Knowledge predication

A semantic typology

Anna Sjöberg
Knowledge predication
A semantic typology
Anna Sjöberg

Academic dissertation for the Degree of Doctor of Philosophy in Linguistics at Stockholm University to be publicly defended on Friday 10 November 2023 at 13.00 in Hörsal 8, hus D, Universitetsvägen 10D.

Abstract
The present thesis is an investigation of the cross-linguistic expression of knowledge predication (‘He knows that it’s raining’, ‘she knows the boy’ etc.) Knowledge predication is investigated in parallel texts—specifically New Testament translations—in a genealogically and areally stratified variety sample of 83 languages. A selection of 100 verses, representing the variety of knowledge as well as near-lying concepts, are manually analysed and annotated for lexical and constructional information and analysed using both qualitative and quantitative methods.

In the first part of the thesis (chapters 2-4), the theoretical and methodological foundations of the investigation are presented. The definition of knowledge is discussed and a survey of previous research on knowledge and its aspects is presented, drawing on work in philosophy and psychology as well as linguistics. The results are the basis for the selection of verses for investigation. The novel sampling procedure employed—a development of the Diversity Values sampling method with added areal stratification—is introduced. This part of the thesis also situates the investigation within the field of semantic typology, describes the theoretical frameworks used – constructional typology and cognitive semantics—and discusses the methodological challenges of parallel corpora drawing on results from translation studies.

In the second part of the thesis (chapters 5-6), the question of whether knowledge is generally distinguished cross-linguistically is addressed. A similarity measure based on lexical and constructional annotations and incorporating automatically inferred paradigmatic relationships is presented and applied. This is then the basis for agglomerative hierarchical clustering, the results of which are interpreted as evidence of an extended knowledge domain (‘know’, ‘understand’ and ‘find out’) and a narrow knowledge domain (‘know’). Also addressed is the question of the universality of knowledge predication, and the sample’s two possible exceptions to this (Kalam [Nuclear Trans New Guinea, PNG] and Fasu [Isolate, PNG]) are discussed in some depth.

In the third part of the thesis (chapters 7-11), the cross-linguistic expression of the extended knowledge domain is investigated in more detail. Using Classical Multi-Dimensional Scaling (also known as Torgerson Scaling or Principal Co-Ordinate Analysis), the dimensions along which the expression of knowledge varies cross-linguistically are explored. Their semantics is investigated, and the typology of their expression is discussed. Additionally, the connections of the different parts of the knowledge domain to other domains such as perception—are explored and analysed from a cognitive semantic perspective. Finally, quantitative evidence for knowledge having a core (‘know that’) and a periphery (‘know person’, ‘understand’ etc.) is presented, including a measure of directionality of formal motivational relationships.

The fourth part (chapter 12) concludes and summarises the thesis. Six generalisations regarding the linguistic expression of the knowledge domain are made, describing for example restrictions on co-expression patterns. Finally, a semantic map of the domain of knowledge predication is presented, synthesising the results of the thesis, and establishing a basis for future research.

Keywords: semantic typology, constructional typology, semantic maps, knowledge, epistemology, parallel texts, language sampling, token-based typology, cognitive semantics.
KNOWLEDGE PREDICATION

Anna Sjöberg
Contents

Abbreviations and conventions vi
List of Figures vii
List of Tables ix
Acknowledgements xi
Sammanfattning på svenska xii

1 Introduction 1
  1.1 Setting the scene ................................. 1
  1.2 Research questions ............................... 3
  1.3 Overview of the thesis ............................ 4

2 Aspects of knowledge 10
  2.1 What is knowledge? ............................... 10
  2.2 The internal structure of knowledge ............... 14
    2.2.1 Knowledge-how and ability .................. 14
    2.2.2 Personal and interpersonal knowledge .......... 17
    2.2.3 Objectual and experiential knowledge ......... 22
    2.2.4 Vertical and horizontal knowledge .......... 26
    2.2.5 Further aspects of understanding .......... 31
    2.2.6 Territory of Information and epistemic status 33
  2.2.7 Event structure ............................... 36
  2.2.8 Summary and conclusions ...................... 38

3 Investigating knowledge predication—theory and method 40
  3.1 Frames, prototypes and exemplars ................ 41
  3.2 Conceptual, comparative and generalised domains 47
  3.3 Connecting meaning and form .................... 52
  3.4 Using the New Testament parallel corpus ........ 59
3.4.1 Different styles of translation ........................................ 60
3.4.2 Translational interference and unnaturalness of register ........ 63
3.4.3 Multiplicity of sources .................................................. 66
3.4.4 Noisiness, low efficiency and difficulty of analysis ............... 68
3.5 Choosing verses ......................................................... 70
3.6 Constructions ............................................................ 75
  3.6.1 Constructional typology ............................................. 75
  3.6.2 Coding for constructions ........................................... 82
3.7 Summary ................................................................. 93

4 Language sampling ........................................................ 95
  4.1 Introduction ............................................................. 95
  4.2 Sampling in typology and the Diversity Value method ................. 95
  4.3 Geographically clustered Diversity Value sampling .................. 97
  4.4 Summary ................................................................. 105

5 Demarcating knowledge .................................................... 107
  5.1 Introduction ............................................................. 107
  5.2 Measuring similarity .................................................. 108
  5.3 Grouping similar verses ............................................. 114
  5.4 Demarcating the knowledge domain ................................ 115
    5.4.1 Methodological issues of the clustering ......................... 116
  5.5 Summary and conclusions ........................................... 123

6 The universality of ‘know’ ................................................ 124
  6.1 Introduction ............................................................. 124
  6.2 Languages without a form for ‘know’ ................................ 125
  6.3 Blurred lines between ‘know’ and other concepts .................... 133
  6.4 Summary and conclusions ........................................... 137

7 The dimensions of knowledge .............................................. 141
  7.1 Introduction ............................................................. 141
  7.2 From distances to semantic maps .................................... 142
  7.3 The dimensions of knowledge: Preliminaries ......................... 148
  7.4 Dimensions 1 and 2: Understanding and finding out ................. 150
  7.5 Dimension 3: Knowing people ....................................... 156
  7.6 Dimension 4: Realisation ............................................. 158
  7.7 Dimension 5: Not knowing ........................................... 160
  7.8 Dimension 6: Knowing how .......................................... 163
  7.9 Summary of the dimensions considered .............................. 164
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 The core-periphery structure of the knowledge domain</td>
<td>166</td>
</tr>
<tr>
<td>8.1 Introduction</td>
<td>166</td>
</tr>
<tr>
<td>8.2 Asymmetries in lexical motivation</td>
<td>167</td>
</tr>
<tr>
<td>8.2.1 Directionality of formal motivation</td>
<td>168</td>
</tr>
<tr>
<td>8.2.2 Domain-external motivation</td>
<td>171</td>
</tr>
<tr>
<td>8.3 The core-periphery structure of the knowledge domain</td>
<td>173</td>
</tr>
<tr>
<td>8.4 Summary and conclusion</td>
<td>176</td>
</tr>
<tr>
<td>9 The dimensions of knowledge: Form and meaning</td>
<td>178</td>
</tr>
<tr>
<td>9.1 Introduction</td>
<td>178</td>
</tr>
<tr>
<td>9.2 From forms to meaning</td>
<td>178</td>
</tr>
<tr>
<td>9.3 Knowing, finding out and understanding</td>
<td>181</td>
</tr>
<tr>
<td>9.3.1 A methodological remark</td>
<td>181</td>
</tr>
<tr>
<td>9.3.2 Achievement and state</td>
<td>182</td>
</tr>
<tr>
<td>9.3.3 Understanding</td>
<td>194</td>
</tr>
<tr>
<td>9.3.4 Finding out</td>
<td>204</td>
</tr>
<tr>
<td>9.3.5 Interim summary</td>
<td>206</td>
</tr>
<tr>
<td>9.4 Recognising and knowing people</td>
<td>207</td>
</tr>
<tr>
<td>9.5 Realising</td>
<td>215</td>
</tr>
<tr>
<td>9.6 Not knowing</td>
<td>216</td>
</tr>
<tr>
<td>9.7 Knowing how</td>
<td>219</td>
</tr>
<tr>
<td>9.8 Summary</td>
<td>223</td>
</tr>
<tr>
<td>10 Knowledge and other concepts</td>
<td>225</td>
</tr>
<tr>
<td>10.1 Introduction</td>
<td>225</td>
</tr>
<tr>
<td>10.2 Previous literature on the inter-domain connections of ‘know’</td>
<td>226</td>
</tr>
<tr>
<td>10.3 Connections to the extended knowledge domain</td>
<td>228</td>
</tr>
<tr>
<td>10.3.1 Connections with HEAR</td>
<td>228</td>
</tr>
<tr>
<td>10.3.2 Connections with SEE</td>
<td>232</td>
</tr>
<tr>
<td>10.3.3 Other connections</td>
<td>234</td>
</tr>
<tr>
<td>10.3.4 Interim summary</td>
<td>236</td>
</tr>
<tr>
<td>10.4 Explaining the connections</td>
<td>236</td>
</tr>
<tr>
<td>10.4.1 Frames, profiles and metonymical change</td>
<td>237</td>
</tr>
<tr>
<td>10.4.2 Bridging frames of the knowledge domain</td>
<td>240</td>
</tr>
<tr>
<td>10.5 Summary and conclusions</td>
<td>243</td>
</tr>
<tr>
<td>11 The dimensions of knowledge: Typology</td>
<td>245</td>
</tr>
<tr>
<td>11.1 Introduction</td>
<td>245</td>
</tr>
<tr>
<td>11.2 Achievement and state</td>
<td>246</td>
</tr>
<tr>
<td>11.3 Understanding</td>
<td>251</td>
</tr>
</tbody>
</table>
11.4 Knowing people .............................................. 255
11.5 Realising ......................................................... 272
11.6 Not knowing ..................................................... 275
11.7 Knowing how .................................................... 278
11.8 Summary and conclusion ................................. 282

12 Conclusions .......................................................... 284
A Appendix: The languages in the study ................. 291
B Appendix: The verse sample of the study ............. 297
C Appendix: Link to Online Supplementary Materials ... 307
References ................................................................. 308
Abbreviations and conventions

1  First person
2  Second person
3  Third person
?  Unknown or uncertain

ABIL  Ability
ABL  Ablative
ABS  Absolutive
ACC  Accusative
AFF.O  Affected object
ANTIP  Antipassive
AN  Animate
ART  Article
AUX  Auxiliary
A  Agent-like argument of canonical transitive verb

CLF  Classifier
COLL  Collective
COMPL  Completive
COMP  Complementizer
CONJ  Conjunction
CONN  Connective
CONT  Continuative
COP  Copula
DAT  Dative

DA  Different actor
DECL  Declarative
DEF.OBJ  Definite object
DEF  Definite
DEL  Delative
DEM  Demonstrative
DS  Different subject
ELA  Elative
EMPH  Emphasis/emphatic
EP  Epenthetic
FAM  Familiar
FUT  Future
GEN  Genitive
HON  Honorific
IAM  Iamitive
IDEOPH  Ideophone
IMPF  Imperfect
INAN  Inanimate
INCH  Inchoative
INDF  Indefinite
IND  Indicative
INS  Instrumental
IPFV  Imperfective
IRR  Irrealis
JUSS  Jussive
LNK  Linker
LOC  Locative
M  Masculine
NARR  Narrative
NEG  Negation
NMLZ  Nominalizer
NOM  Nominative
NPST  Non-past
O  Object
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFV</td>
<td>Perfective</td>
</tr>
<tr>
<td>PF</td>
<td>Perfect</td>
</tr>
<tr>
<td>PL</td>
<td>Plural</td>
</tr>
<tr>
<td>POSS</td>
<td>Possessive</td>
</tr>
<tr>
<td>PR.NAME</td>
<td>Proper name</td>
</tr>
<tr>
<td>PREP</td>
<td>Preposition</td>
</tr>
<tr>
<td>PRS</td>
<td>Present</td>
</tr>
<tr>
<td>PRT</td>
<td>Particle</td>
</tr>
<tr>
<td>PST</td>
<td>Past</td>
</tr>
<tr>
<td>PTCP</td>
<td>Participle</td>
</tr>
<tr>
<td>Q</td>
<td>Question marker</td>
</tr>
<tr>
<td>REDUP</td>
<td>Reduplication</td>
</tr>
<tr>
<td>REFL</td>
<td>Reflexive</td>
</tr>
<tr>
<td>REL</td>
<td>Relative</td>
</tr>
<tr>
<td>REM</td>
<td>Remote</td>
</tr>
<tr>
<td>SA</td>
<td>Same actor</td>
</tr>
<tr>
<td>SBJ</td>
<td>Subject</td>
</tr>
<tr>
<td>SENT.SUFF</td>
<td>Sentence suffix</td>
</tr>
<tr>
<td>SEQ</td>
<td>Sequential</td>
</tr>
<tr>
<td>SG</td>
<td>Singular</td>
</tr>
<tr>
<td>STAT</td>
<td>Stative</td>
</tr>
<tr>
<td>S</td>
<td>Single argument of canonical intransitive verb</td>
</tr>
<tr>
<td>TOP</td>
<td>Topic</td>
</tr>
<tr>
<td>TR</td>
<td>Transitive</td>
</tr>
<tr>
<td>TEMP</td>
<td>Temporal adverbialiser</td>
</tr>
<tr>
<td>VENT</td>
<td>Ventive</td>
</tr>
<tr>
<td>VOC</td>
<td>Vocative</td>
</tr>
<tr>
<td>VRBZ</td>
<td>Verbaliser</td>
</tr>
</tbody>
</table>

Examples are glossed according to the Leipzig Glossing Rules (Comrie et al. 2008), with some abbreviations added. The list of abbreviations used in glosses can be found above. Unless otherwise indicated, all glosses are my own. When glossed examples are cited from elsewhere, the glossing and translation is that of the original source unless otherwise indicated. When the example is from the New Testament, the free translation is from the Lexham version unless otherwise indicated. Free translations supplied by the author are marked with (AS). Language names as well as genealogical affiliation and location is according to the Glottolog (Hammarström et al. 2019). For simplicity’s sake, a single country is given for each language.

Italics are used to refer to words in individual languages, e.g. know for the English word. Single quotes, as in ‘know’, are used for meanings. Capital letters, as in KNOW, are used to refer clusters resulting from automatic clustering methods. Capitalised category labels, as in Know, are used to refer to the a priori categorisation of verses (see section 3.5) Small capitals, finally, as in WORD, are used for more precisely meant concepts, such as those of Natural Semantic Metalanguage.
## List of Figures

1.1 Map of the languages in the sample . . . . . . . . . . . . . . . 6
1.2 A semantic map of the knowledge domain . . . . . . . . . . 9

2.1 Aspectual structures of ‘know’ . . . . . . . . . . . . . . . . 37

3.1 Depiction of the construal process profiling . . . . . . . . . . 46
3.2 Illustration of automatic marker extraction . . . . . . . . . . 72

4.1 Illustration of the sampling procedure . . . . . . . . . . . . . 104
4.2 Map of the languages in the sample II . . . . . . . . . . . . . 106

5.1 Silhouette scores against number of clusters . . . . . . . . . . 116
5.2 Cluster inclusion in three different clustering algorithms . . . 119
5.3 Average similarity across and within clusters . . . . . . . . . . 120
5.4 Agglomerative clustering with different number of clusters . . 122

7.1 CMDS analysis of driving distances between European cities . 143
7.2 Eigenvalues of the dimensions of the CMDS solution . . . . . 149
7.3 Dimension 1 and 2 of the CMDS analysis . . . . . . . . . . . 150
7.4 Dimension 1 and 2 of the CMDS analysis, Swedish . . . . . 152
7.5 Dimension 1 and 2 of the CMDS analysis, Bengali . . . . . . 153
7.6 Dimension 1 and 2, full and lexical, Bengali . . . . . . . . . . 155
7.7 Dimension 1 and 2, constructional, Bengali . . . . . . . . . . 155
7.8 Dimension 1 and 2 of the CMDS analysis, North Tanna . . . 156
7.9 Dimension 1 and 3 of the CMDS analysis, Swedish . . . . . . 157
7.10 Dimension 4 with Dimensions 1 and 2, Swedish . . . . . . . 159
7.11 Dimension 1 and 4 of the CMDS analysis, Mapudungun . . . 160
7.12 Dimension 1 and 5 of the CMDS analysis, Moro . . . . . . . . 162
7.13 Swedish *kunna* and Car Nicobarese *huveueu* . . . . . . . 164
7.14 Summary of the dimensions of the CMDS analysis . . . . . . . 165

8.1 Graph of directionality, knowledge and perception . . . . . . . 170
8.2 Graph of directionality, knowledge and perception . . . . . . . 170
8.3 Graph of formal directionality, narrow knowledge domain . . . 172
8.4 Proportion of colexification by cluster . . . . . . . . . . . . . 173
8.5 Average number of colexifications, narrow knowledge domain 174
8.6 Core-periphery structure and iconic representation . . . . . . 176
9.1 Aspectual construals in Trinitario-Javeriano-Loretano . . . . . 187
9.2 Dimensions 1 and 3 of the CMDS analysis, Hopi . . . . . . . . 213
10.1 Similarity between KNOW and other clusters . . . . . . . . . 229
10.2 Similarity of verses in the knowledge domain to HEAR . . . . 230
10.3 Similarity of verses to HEAR, ‘find out’ excluded . . . . . . . 231
10.4 Similarity of verses in the knowledge domain to SEE . . . . . 232
10.5 Similarity of verses to THINK and REMEMBER . . . . . . . 234
10.6 Similarity of verses in the knowledge domain to BE ABLE . . 236
10.7 Integration of semantic change in language . . . . . . . . . . . 238
10.8 Know-as-achievement event with different profilings . . . . . 240
10.9 Know-as-achievement event with different profilings II . . . 241
10.10 Know-as-achievement event with different profilings III . . . 242
11.1 State-achievement directionality . . . . . . . . . . . . . . . . 250
11.2 State-achievement directionality, zoomed in . . . . . . . . . . 251
11.3 Map of State-achievement directionality . . . . . . . . . . . . 252
11.4 Similarity of KNOW and UNDERSTAND . . . . . . . . . . . 253
11.5 Distribution of forms correlated with Lexical Dimension 1 . . 255
11.6 Distribution of forms correlated with Dimension 3 . . . . . . . 256
11.7 Forms correlated with Dimension 3, discussed languages . . . 265
11.8 Distribution of forms correlated with Dimension 3 by area . . 266
11.9 Map of ‘know person’ forms . . . . . . . . . . . . . . . . . . 266
11.10 Non-person nominal objects along Dimension 3 . . . . . . . . 268
11.11 The meaning of two Tepehua words . . . . . . . . . . . . . . 269
11.12 Distribution of forms correlated with Dimension 4 . . . . . . 272
11.13 Dimasa *mithi* and *tharmai*, Dimensions 1 and 4 . . . . . 274
11.14 Map of markers correlated with Dimension 4 . . . . . . . . . 274
11.15 Dimensions 1 and 5, North Tanna . . . . . . . . . . . . . . . 276
11.16 Map of the distribution of lexical negation . . . . . . . . . . 278
11.17 Map of types of ‘know how’-expression . . . . . . . . . . . 281
12.1 A semantic map of the knowledge domain . . . . . . . . . . . 290
List of Tables

1.1 Summary of the interpretation of the first six dimensions . . . 7
3.1 Colour terms in English and Navajo . . . . . . . . . . . . . . 59
3.2 Translation equivalents in Hungarian . . . . . . . . . . . . . . 86
3.3 Construction in three Hungarian verses I . . . . . . . . . . . . 88
3.4 Construction in three Hungarian verses II . . . . . . . . . . . . 88
3.5 Construction in three Hungarian verses III . . . . . . . . . . . . 89
3.6 Examples of codings of verses from Hungarian . . . . . . . . 91
3.7 Number of occurrences of different coding values . . . . . . . 92
4.1 Overview of sampling procedure . . . . . . . . . . . . . . . . 98
4.2 Number of sample groups in each area without NT translations 103
5.1 Methodological summary of chapter 5 . . . . . . . . . . . . . . 108
5.2 Examples of codings of verses from Hungarian . . . . . . . . 109
5.3 Example of comparison of markers in Swedish . . . . . . . . . 112
5.4 Examples of codings of verses from Hungarian . . . . . . . . 113
5.5 Clustering of the verses at 8 and 10 clusters . . . . . . . . . . 117
6.1 Methodological summary of chapter 6 . . . . . . . . . . . . . . 125
6.2 Number of languages absolutely colexifying concepts in CLICS 126
7.1 Methodological summary of chapter 7 . . . . . . . . . . . . . . 142
7.2 Driving distances between European cities . . . . . . . . . . . . 143
7.3 Summary of the interpretation of the first six dimensions . . . . 149
7.4 The three verses with the lowest values in Dimension 1 . . . . 151
7.5 The three verses with the highest values in Dimension 2 . . . . 151
7.6 The three verses with the highest values in Dimension 1 . . . . 151
7.7 Selected verses with high values in Dimension 3 . . . . . . . . 157
7.8 The three verses with the highest values in Dimension 4 . . . . 159
7.9 The three verses with the lowest values in Dimension 5 . . . . 161
7.10 The three verses with the highest values in Dimension 6 . . . . 163
Acknowledgements

Thank you, to my supervisors, professor Bernhard Wälchli and professor Maria Koptjevskaja-Tamm, for your unceasing support, your excellent advice and your willingness to give so freely of your time and knowledge. This thesis would not have been possible without you.

Thank you, to my parents, for their support throughout my time as a PhD student (as well as before it, and, I hope, after).

Thank you, to the Department of Linguistics, Stockholm University for giving me the opportunity to pursue my PhD during these almost five years. Thank you also to everyone in the department: my fellow PhD students (past and present), who have not only generously contributed with their time and expertise to improve my thesis but have also made it a fun five years; the senior researchers and teachers, for creating an academic environment to thrive in; and of course to the technical and administrative staff who have been rocks in the stormy seas of bureaucracy, administration and technology. I especially thank Marcin Włodarczak for (indirectly) suggesting the paradigm-adjusted similarity measure and Nastia Panova for her help in constructing the Online Supplementary Materials.

Thank you, to Calle Börstell, whose comments as my mock opponent have much improved this thesis.

Thank you, to professor Anju Saxena, for giving me my first job in linguistics and making me infinitely more well-prepared for my PhD studies than I would have been without her.

Last, but certainly not least, thank you, to all those speakers, translators and linguists who have given of their time and knowledge to make so many languages available through New Testament translations; this thesis would not have been possible without it.

Anna Sjöberg
Uppsala, September 2023


I den tredje delen (kapitel 7–11) undersöks mer i detalj det tvärspråkliga uttrycket av den utökade kunskapsdomänen. Genom att använda klassisk multi-deimensionell skalning (även känt som Torgerson-skalning eller Principal Co-Ordinate Analysis) utforskas de dimensioner längs vilka uttrycket av kunskap varierar tvärspråkligt. Dimensionernas semantik undersöks och deras uttrycks typologi diskuteras. Dessutom analyseras kopplingarna mellan olika delar av

1. Introduction

1.1 Setting the scene

Knowledge has been a subject of enquiry for great thinkers across cultures for millennia. In Plato’s Theaetetus, Socrates debates what separates knowledge from true belief; the Ming dynasty Chinese philosopher Wang Yangming argued for the unity of zhi ‘knowledge’ and xing ‘action’; and the various Vedic—concerned with the Vedas, literally ‘knowledge’—schools of classical Indian philosophy discussed the reliability of various knowledge-sources. Also in modern days, epistemology—the theory of knowledge—is a core part of philosophy and introductory courses on philosophy will include it (for introductions to epistemology see e.g. Williams (2001) and Nagel (2014)).

However, it is not only philosophers and academics who are preoccupied with knowledge or its absence. Know is the sixth most common verb in spoken English, sharing the position with say, modal verbs excluded (Leech, Rayson & Wilson 2001). In a corpus of Mandarin Chinese movie and TV series subtitles, 知道 zhīdào ‘know’ is the seventh most frequent word (Qing & Byrsbert 2010). And in a corpus of Māori, consisting of texts for children and materials from broadcasting, mōhio ‘know’ is the 51st most frequent word, including function words (New Zealand Ministry of Education 2010).

In addition to a high frequency within languages, there is good evidence that forms for ‘know’ are highly frequent across languages. For example, ‘know’ is included in the Leipzig-Jakarta list—a list of 100 meanings which in a survey of 41 genealogically, geographically and culturally diverse languages were found to be both highly widespread and unlikely to be borrowed (Haspelmath & Tadmor 2009). And within Natural Semantic Metalanguage, an approach which posits the existence of universally lexicalised atoms of meaning called primes ‘know’ has been found to be one of these (as of the most recent publication, Goddard, Wierzbicka & Farese 2022) 65 primes (Wierzbicka 2018b).

Knowledge thus appears to be a central concern to humans across times and cultures. This alone makes its expression in the world’s languages a worthwhile object of investigation, and this is what this thesis is about. Knowledge, however, can be expressed in many ways in languages. This thesis will focus
on knowledge predication, as exemplified by the Swedish example in (1)

(1) Swedish (Indo-European, Germanic)

a. Gothenburg Dialogue Corpus (Borin, Forsberg & Roxendal 2012)

\[ \text{hon vet var du kommer ifrån och hur det varit} \]
\[ 3SG.F \text{ know.PRS where 2SG come.PRS from and how it be.SUP} \]

‘she knows where you are from and how it has been’

By knowledge predication I understand a construction that assigns participant roles in a knowledge event to particular entities (see further section 3.5).

Arguably, any linguistic expression somehow relates to the knowledge of speakers and hearers. Some languages also have dedicated grammatical categories to expressing knowledge stances, known as evidentiality or epistemicity (c.f. Evans 2005, Aikhenvald 2004). These phenomena do not fall under the definition of knowledge predication, and are thus outside the scope of this thesis.

Whereas knowledge predication has been investigated in a few individual languages (e.g. Vatrican 2004, Van Hulle 2015), the cross-linguistic picture remains largely limited to the identification of rough translation equivalents (e.g. Vanhove 2008, Georgakopoulos et al. 2022). However, one does not need to look very far (from the perspective of English, which currently dominates the discussion) to find mismatches in the meanings of ‘know’ forms across languages. In German, English know corresponds to two verbs—wissen ‘know that’ and kennen ‘know a person, object’. Staying in the Indo-European family, we find Russian, where know corresponds to znat’ ‘know that’ and umet’ ‘know how (to do something)’. As a final example—geographically close to the former and also Indo-European—there is Swedish, where know corresponds to three verbs—veta ‘know that’, känna ‘know person’ and kunna ‘know how’ (also ‘be able’).

Already these few examples illustrate the complexity of identifying of translational equivalents. Does a translation equivalent of English know need to have all these meanings? Then what should be done with languages like German, Russian or Swedish, where some of the senses are expressed by distinct lexemes? And, perhaps more importantly, why should we take what might be only the polysemy—the situation of a form having multiple, related meanings—of a single language as defining a particular concept? Is English an oddity in having the same word for the uses identified above? Or is it the other languages mentioned above that are unusual in making distinctions here?
What has been said so far concerns forms having a narrower meaning than the English *know*. The possibility exists also, of course, of broader meanings. Do we find languages where no distinction is made between, for instance, knowing and remembering or knowing and believing?

### 1.2 Research questions

In short, it is well known that in many languages, there is a form meaning something like ‘know’, but little is known about the range of meaning for these forms—do they roughly correspond to the English *know*, do they have a narrower use or do they have a broader use? This issue is what the first research question and its sub-questions to be investigated in this thesis revolve around:

A. What is the cross-linguistic status of knowledge predication as a distinct construction?

   A1. Is knowledge predication generally distinguished in languages, and if so, what are the boundaries of it?

   A2. How universal is a distinct construction for knowledge predication?

I address these questions in chapters 5 and 6 and argue that cluster analysis of cross-linguistic data shows that a *generalised knowledge domain* (see section 3.2) can be distinguished.

However, as the examples given above of the linguistic expression of knowledge in four languages—English, German, Russian and Swedish—show, there is considerable variation of expression within the generalised knowledge domain. Given this diversity found in four languages with geographically close homelands and which are genealogically related, what kind of distinctions are found if the scope is widened to include languages from all over the world? What possible meanings are discovered? Is meaning carved up arbitrarily or is the variation constrained? How do the different possible senses of ‘know’ forms relate to each other—are they all equal or are some more central than others?

These are the kind of issues that the second set of research questions of this thesis is concerned with:

B. What linguistic distinctions are made in the knowledge domain—what meanings are distinguished?

   B1. How do these meanings relate to one another—are some more basic or central?

   B2. How can the distinctions be characterised semantically?
B3. How do the forms for these meanings relate to forms for domain-external meanings?

B4. How are these distinctions expressed in the languages of the world?

The cross-linguistic orientation of both sets of research questions situates this thesis within the tradition of semantic typology, defined by Evans (2010: 504–505) as “the systematic cross-linguistic study of how languages express meaning by way of signs”. A recurring challenge of semantic typology is what data to use (see e.g. Koptjevskaja-Tamm 2012). The traditional grammars of grammatical typology rarely give the necessary information and for many research questions—including both sets posed here—more detailed data than that found in word-lists or dictionaries is required.

In this thesis, parallel texts are used as the main data sources (cf. Wälchli & Cysouw 2012). Among the various texts available in more than one language, the Christian New Testament is unequalled in the number and variety of translations available and it is therefore the text used. This allows for a fine-grained investigation of the expression of knowledge in what is, for semantic typology, a rather large stratified variety sample of 83 languages. The expression of knowledge predication as well as near-lying domains is investigated in these languages using both quantitative and qualitative methods, the results of which are reported on in this thesis.

In the following section, an overview of the thesis is given.

1.3 Overview of the thesis

This thesis consists of 12 chapters, this introduction included.

In chapter 2, the object of investigation—that is, knowledge—is explored. Section 2.1 considers philosophical definitions of knowledge and how they relate to the question of the universality of knowledge. Note that in accordance with the method chosen—the use of parallel texts—no definition discussed is taken as defining the object of investigation, nor serves as a comparative concept or comparative domain (see also section 3.2). Rather, the domain under investigation is defined as a set of contexts (see section 3.5).

Section 2.2 gives a survey of what previous work in philosophy, psychology as well linguistics and related disciplines can tell us about the aspects of knowledge. This section is an important methodological tool in giving a preliminary sketch of the knowledge domain which serves as a basis for further investigation. The areas considered are know-how and ability (section 2.2.1), personal and interpersonal knowledge (section 2.2.2), objectual and experiential knowledge (section 2.2.3), depth of knowledge as typically expressed by forms like understand (section 2.2.4), further aspects of understanding (section
2.2.5 territories of information and epistemic status (section 2.2.6) and lastly the event structure of knowing (section 2.2.7). The chapter concludes with a summary of the main points which are hypothesised to be of cross-linguistic relevance (2.2.8).

Chapter 3 deals with the theory and method of the investigation. First, section 3.1 introduces relevant concepts from one of the main theoretical perspectives of the thesis, namely cognitive semantics. A view of meaning as conceptualisation, categories as structured around prototypical frames and the importance of construal are discussed. This is followed by a section where a typology of domains is introduced (3.2). I argue that to properly understand and discuss the notion of domain in semantic typology (and, indeed, typology more generally) three uses of it need to be distinguished—conceptual domains, comparative domains and generalised domains. In section 3.3, I discuss the general methodological approach of the thesis, starting from the three central problems of semantic typology as identified by Koptjevskaja-Tamm (2012: 383). The approach used here—parallel text typology—is compared with various other approaches and the problems of data, cross-linguistic meaning identification and metalanguage are addressed.

Methodological issues concerning the use of parallel texts in general and New Testament translations in particular are further discussed in section 3.4. Drawing on work from the field of translation studies, I address some of the difficulties that have been raised concerning the use of parallel corpora for typological investigation and give a theoretical grounding for the notion of translational equivalence.

Section 3.5 describes the choice of which specific verses from the New Testament to include in the investigation, building on the results of section 2.2. The section also describes the process of using automatically extracted correspondences of know and their distribution in the New Testament to construct a cross-linguistically valid characterisation of ‘know’ as a set of verses. Put simply, we are not interested in a typology of how English know is expressed across languages, but rather how a more general cross-linguistic notion of ‘know’ is expressed.

In section 3.6 constructional typology with its roots in construction grammar, the other of the two main theoretical approaches underlying the investigation, is described. I argue that considering only lexical items is not enough to fully capture the variation in the expression of knowledge and that the constructions involved also need to be considered, drawing on Koch (2012a). Section 3.6.2 describes how the construction grammar approach is turned into the concrete coding schema for verses used in the investigation. I also present the automatic process of assigning coded markers to paradigm slots, allowing the similarity measure described later to take paradigmatic relations into
In chapter 4, the novel language sampling procedure used in the thesis is described. The method is an adaptation of the Diversity Value (Baker 1993) which in addition to considering the genealogical relationships of languages also takes into account geography. The details of the method are described and the application to the Glottolog language catalogue (Hammarström et al. 2019) and the New Testament corpus is then presented. The procedure results in an 83 language sample, shown in map 1.1 and Appendix I.

Chapter 5 begins the presentation of the investigation’s results and is concerned with the first research question. The chapter starts with a discussion of how the codings described in section 3.6.2 can be represented as distances between verses. Building on previous works, a distance measure that weighs lexical and constructional markers and also takes into account the paradigmatic relationships of markers is introduced (section 5.2). This distance measure is then used to apply hierarchical agglomerative bottom-up clustering to the verses, thereby giving an objective and quantified answer to how language categorise the investigated contexts (section 5.3). The results of the clustering are presented in section 5.4. I show that distinguished by the clustering are both a narrow knowledge domain (roughly corresponding to English know) and an extended knowledge domain (roughly know, understand and find out)—that is, there does appear to exist a knowledge domain cross-linguistically.

In chapter 6, I discuss research question A2 that is, the universality of ‘know’ as a category. The two languages in the sample which are potential exceptions (Fasu [Isolate] and Kalam [Nuclear Trans New Guinea], both spoken in Papua New Guinea) are discussed, as well as a number of other languages
where the separation of knowledge and other meanings is not clear-cut. I conclude that the near-universality of ‘know’ is a striking fact, even if no definitive answer regarding these languages can be reached based on the current data.

Chapter 7 begins the presentation of the results pertaining to the second research question. To identify the relevant parameters of variation within the domain, Classical Multi-Dimensional Scaling—a statistical method for inferring locations in a multi-dimensional space from distances—is applied to the verse sample. After a general description of this method, including a discussion of other similar methods (section 7.2) the dimensions of the knowledge domain—as identified in chapter 5—are interpreted. The first six dimensions—the choice of cut-off point is determined based on the data and is discussed in section 7.3—are given a first characterisation, summarised in table 1.1. Of particular note is the first dimension, where a distinction between ‘know’ and ‘understand’ coalesces with one between know-as-state and know-as-achievement. I show that the former distinction is made mainly at the lexical level while the latter is made constructionally. This shows how important it is to consider both lexical and constructional information. Of interest is also the very clear connection between knowing people and recognising them.

In chapter 8, question B1 about the internal relationships of the meanings identified in chapter 7 is addressed. Quantitative evidence is presented of the asymmetry of lexical motivation patterns within the knowledge domain (section 8.2)—in addition to considering the likelihood of external motivation or colexification, I introduce a quantitative measure of the directionality of formal motivation (8.2.1). Building on ideas from frame semantics and prototype theory, I argue that these difference can be explained by positing a core–periphery structure of the knowledge domain where ‘know that’ is the core (section 8.3).

In chapter 9, the semantics of the dimensions identified are further explored, addressing research question B2. The chapter begins with an account of how the view of meaning sketched in section 3.1 allows for using the properties of cross-linguistic forms associated with the dimensions to make infer-

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Low pole</th>
<th>High pole</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>States</td>
<td>Achievements</td>
</tr>
<tr>
<td>or</td>
<td>Know</td>
<td>Understand</td>
</tr>
<tr>
<td>2</td>
<td>Internal</td>
<td>External</td>
</tr>
<tr>
<td>3</td>
<td>Know that</td>
<td>Know person, Recognise</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Realise</td>
</tr>
<tr>
<td>5</td>
<td>Not know</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Know how</td>
</tr>
</tbody>
</table>

**Table 1.1:** Summary of the interpretation of the first six dimensions.
ences about the semantics of the dimensions. Based on this, a more detailed semantic characterisation is given of each of the six dimensions identified in 7. A noteworthy result is that the characterisation of understanding given in sections 2.2.4 and 2.2.5 is more or less in its entirety echoed by the forms found expressing these contexts cross-linguistically.

Chapter 10 addresses research question B3—what are the connections of knowledge to other concepts? The chapter begins with a brief survey of previous research on the topic of the connection of ‘know’ with other domains (section 10.2). Results concerning the connections of different parts of the knowledge domains are presented (section 10.3) and I show that different concepts are connected with different parts of the knowledge domain. ‘Hearing’ is strongly connected to ‘finding out’ and ‘understanding’, ‘seeing’ to ‘knowing person’ and ‘recognising’, ‘thinking’ and ‘remembering’ to ‘understanding’ and ‘realising’ and, finally, ‘be able’ to ‘knowing how’. The theoretical framework presented in section 3.1 is applied to suggest a model explaining these connections in terms of metonymy within bridging frames.

Chapter 11 deals with question B4: the expression of the dimensions of knowledge in the languages of the sample is explored through the lens of token-based typology—rather than attempting to fit languages into abstract, discrete types, their behaviour at the token level is considered. The measure of formal directionality mentioned above is applied to the distinction between know-as-state and know-as-achievement. Another noteworthy result is The personal knowledge scale, a tendency scale describing the extension of forms meaning ‘know person’, where I argue that if a form distinct from the general word for ‘know’ is used for some meaning on the scale, it will also be used for meanings to the left of it, barring the use of some other form distinct from general ‘know’:

The personal knowledge scale

Recognise (a person) < Know a person (directly) < Know a person (indirectly)

Finally, chapter 12 outlines the main conclusions reached and identifies possible avenues for future investigation. It does so in part by proposing a network semantic map of the knowledge domain, based on the results presented in previous chapters—see figure 1.2.

There are also three appendices in the thesis. Appendix A gives the language sample of the study, including details of the New Testament translation used and references to secondary sources. Appendix B contains the verse sample of the study. Appendix C contains a link to the Online Supplementary Materials.
Figure 1.2: A semantic map of the knowledge domain
2. Aspects of knowledge

A distinction is commonly made between two approaches in semantic typology: onomasiology and semasiology (see e.g. Koptjevskaja-Tamm 2012). In the semasiological approach, the forms in a given language are the basis of the investigation—what else can form X mean, how does it connect to other forms and so on. In the onomasiological approach, meanings are the basis of investigation—for a given domain of meaning, it is asked how languages carve it up linguistically. The approach taken in this investigation is mainly onomasiological. As is discussed in more detail in section 3.3 contextually embedded situations in the form of translations of Bible verses serve as an etic grid. An etic characterisation aims to include “all logically distinguishable possibilities, regardless of whether or not individual languages group them together” (Evans 2010: 509), as opposed to an emic characterisation which focuses on the actual characterisations of individual languages. However, in the case of knowledge predication, it is far from clear what the “logically distinguishable possibilities” are, or indeed, how to go about establishing this.

The purpose of this chapter is to sketch an answer to this question. In order to do this, I draw on the considerable body of research on the topic of knowledge, including works in philosophy and psychology as well as linguistics and related sciences (section 2.2). Before addressing the variation within the knowledge domain, the question of what knowledge actually is will be briefly discussed, along with related views on the universality of knowledge (section 2.1). Neither of these surveys has any ambition toward completeness and rather draws on what is useful in establishing a preliminary sketch of relevant aspects of knowledge to be used in the subsequent investigation.

2.1 What is knowledge?

As mentioned in the Introduction and discussed in more detail in section 3.3, the method of the investigation into knowledge in this thesis consists of taking contextually embedded situations in the form of translation of Bible verses as the basis of cross-linguistic comparison. It is these contextually embedded situations which are the analytical primitives of the investigation, not some concept defined in the abstract. Nonetheless, it is of interest to consider the ques-
tion of what characterises the situations to be investigated—in other words, what is knowledge?

This is one of the traditional questions of the philosophical field of epistemology or theory of knowledge, and philosophers and others have been investigating it in that context since antiquity. In this section, some of the major views are introduced.

We begin with what is often considered the classical analysis of knowledge, also known as the tripartite analysis. This decompositional analysis was famously expressed by Gettier (1963: 121) as:

(2)  S knows that P [if and only if]

   (i) P is true,
   (ii) S believes that P, and
   (iii) S is justified in believing P.

Gettier attributes “some such definition” (1963: 121) already to Plato (put forth in the dialogue Theaetetus). The view of the tripartite analysis as dominant for most of the history of Western philosophy has been challenged, but there is no doubt that it is the standard definition, found in textbooks and overviews. The first two criteria of the definition are fairly straightforward. First of all, we can only know true things. Examples such as

(3)  English (Indo-European>Germanic, UK)

   a. Ichikawa & Steup (2018)

   I knew that Clinton was going to win [the 2016 presidential election, but she didn’t]

are, in this view, best explained as extended, non-literal uses. Generally, asserting that S knows P presupposes the truth of P. The second criterion — the belief criterion — specifies that we are dealing with a relation that holds between a subject and that which it known. The third criterion — for the true belief to be justified — is the most discussed one. Much of this discussion has been due to Gettier’s famous 1963 paper where he goes on to demonstrate that there are several cases where all three criteria apply, but we are nonetheless hesitant to say that something is knowledge—these have become known as Gettier problems or cases (see Gettier’s original paper Gettier (1963) or e.g. Ichikawa & Steup (2018)).

---

1For reasons of time and space, the following will be limited to approaches from the Western tradition of philosophy
As an example of a Gettier problem (though not one due to Gettier, but Chisholm 1966: 23), imagine a person looking at a field. They see what looks like a sheep, and form the belief that there is a sheep in the field. This belief is clearly justified. However, what the person sees is not a sheep, but a dog disguised as a sheep. But behind the dog, there is in fact a real sheep. The person thus has a true, justified belief that there is a sheep in the field—but do they really know it?

Gettier’s challenge to the classical analysis has given rise to many attempts to refine the definition to exclude these problematic cases from the definition of knowledge. Most have focused on the justification criterion—for instance by suggesting that there must be causal links between P and S’s belief that P for the belief to be knowledge (Goldman 1976) or by positing a “no false lemmas”—condition, i.e. no supposition which figures in S’s justification (i.e. lemma) may be false (Clark 1963). So far, no modified analysis has won general acceptance.

An important question to ask here is of course what it is we are aiming to define by the criteria in 2 or other definitions. When asking the question “What is it to know something?”, what are we inquiring about? Carter & Sosa (2022) distinguish three distinct ways in which the question can be understood—semantic, conceptual and metaphysical. With the semantic understanding, the question is simply about what know means (or the corresponding word in whatever language one happens to be doing epistemology in). With the conceptual understanding, the goal is rather to characterise the abstract concept which is expected to underlie the word. With the metaphysical understanding, finally, the question concerns knowledge per se. It is far from always clear which understanding an author adopts.

Relevant to the view taken on how the object of epistemological analysis is properly understood is the view of the question of knowledge’s existence. When we talk about knowing things, does this correspond to some actual facts of the matter, or does it come down to social conventions? Depending on the stance taken on this issue, meta-epistemological realism and anti-realism can be distinguished (Carter & Sosa 2022). The realist position can basically be characterised as the belief that knowledge ascriptions are true or false by virtue of some property of reality, independent of the people involved. As Carter & Sosa (2022) put it, for the realist “when you say that you know that Paris is the capital of France, there is a fact of the matter—either you do or you do not know this—and further, the obtaining of this fact is not simply down to, e.g., cultural consensus, approval, or what you think about what you know.”

Given the realist view, we would expect knowledge predicates to be similar across the world’s languages independent of culture, since they express a shared, universal property or state—Knowledge, if you will.
For meta-epistemological anti-realists, statements about what someone knows and does not know are not objectively true or false. Among the more familiar theories in this group we find relativism, which claims that a statement such as “S knows that P” is only true or false in relation to a certain epistemic system. Thus, it can be true that Galileo and the churchman arguing against him both knew contradictory things, since they did their knowing in different epistemic systems where the criteria for justification differed (Rorty 1980: 330). On this view, the similarity or dissimilarity of knowledge predicates cross-linguistically is fundamentally an empirical question—there is no shared, universal Knowledge. This view opens up discussions about knowledge and power—if there is no objective truth about what you know, the knowledge of a group is entirely defined by those who have the power to set the terms of the epistemic system.

Note that though the relativist view rejects a metaphysical basis for knowledge, it at least in some variants retains the classical tripartite analysis, or some variant of it. As mentioned, however, the tripartite analysis has been difficult to rescue from the attacks of Gettier and those coming in his wake. In part because of the difficulty in settling on a set of necessary and sufficient conditions for what it means to know something, some scholars have abandoned the decomposition attempts—knowledge is knowledge, and that is all that can be said about it. This is sometimes known as knowledge first epistemology (the term is particularly associated with the philosopher Timothy Williamson, e.g. Williamson 2000). Within linguistics, a famous proponent of knowledge first epistemology is Wierzbicka (e.g. 2018b) who includes KNOW as a prime in her Natural Semantic Metalanguage (NSM). NSM is discussed more in section 3.3 but for now it can be noted that being a prime in NSM means being immediately graspable and undefinable without circularity. In other words, Wierzbicka rejects the tripartite analysis of knowledge outlined above. She does this by arguing that it is the conversational sense found in ‘I know (it)’ which is the basic meaning of KNOW and not, as more traditionally held, the ‘know that’ sense. This solves the problem of the apparent factivity of know—if to KNOW something implies its truth, it would seem that this truth is part of the concept KNOW, which is thus decomposable. The solution is to relegate the truth implication to a so called frame of KNOW—KNOW that—and not the prime itself.

In the NSM view, then, knowledge is universal and knowledge predicates cross-linguistically all express the same basic concept, KNOW. NSM is most readily interpreted as a psychological theory (cf. Dessaix 2016) and it is thus in some sense interested in the (non-)analysis of knowledge at the conceptual level. The claim that there are primes that constitute a universal “alphabet of human thought” (cf. section 3.3), however, means that the analysis at
least approaches the metaphysical. Whether it is a realist view or not depends somewhat on one’s point of view; on the one hand, the truth of knowledge ascriptions depends on mental concepts, i.e. primes—making it an anti-realist view—but on the other hand, these primes are supposedly universal and given, which gives the approach a realist element.

To summarise, knowledge has been traditionally analysed as having three components—S know that P if, and only if, i) P is true, ii) S believes that P and iii) S is justified in believing P. This is known as the tripartite analysis of knowledge, and remains the standard analysis in (Western analytical) philosophy [Ichikawa & Steup 2018]. even though it has been challenged by so called Gettier cases since the 1960s. In part as a response to these challenges, some authors have suggested a knowledge first approach where knowledge is taken as basic and unanalysable. Within linguistics, the most famous such suggestion is that of the NSM school. Views also differ, however, at a deeper level of analysis; how is the question of what knowledge is to be understood—semantically, conceptually or metaphysically? The answer depends on the meta-epistemological view taken—are there facts of the matter about what someone knows (the realist view) or not (the anti-realist view)? This last question also has implications for how the universality of knowledge is viewed—in the realist view, it is necessarily universal while in the anti-realist view, similarity between knowledge predicates is not a priori given but a question to be empirically determined.

2.2 The internal structure of knowledge

The definitions discussed above centred mainly on what is usually called propositional knowledge, that is, knowledge of the kind ‘I know that X’. But as we saw already in the Introduction, know has a broader meaning than propositional knowledge—and many philosophers have taken knowledge to also include more. Others have argued that there is a difference in kind between the relations that know can express. The following sections (2.2.1–2.2.7) consider aspects of knowledge that have been discussed in these arguments and other other aspect that may be of relevance to the linguistic expression of knowledge. The result is a preliminary sketch of the parameters of variation of knowledge in section 2.2.8.

2.2.1 Knowledge-how and ability

Consider the following sentences:
Does know mean the same thing in these contexts? This has been a debate in philosophy at least since [Ryle (1945, 1949): whether or not knowledge-how (4b) is of a different kind or not as knowledge-that (4b). The positions may be summarised as follows, following [Fantl (2017)]:

(i) Intellectualism: knowledge-how is reducible to knowledge-that, i.e. knowing how to do something consists of knowing a set of propositions

(ii) Moderate anti-intellectualism: knowledge-how and knowledge-that are of different kinds

(iii) Radical anti-intellectualism: knowledge-that is reducible to knowledge-how, i.e. knowing a proposition consists of knowing how to do something

Within the modern field of philosophy, proponents of all three positions may be found. The philosophical discussion philosophy has mainly been concerned with, as well as conducted in, English. [Rumfitt (2003)] is one of the exceptions. In his criticism of [Stanley & Williamson (2001)], who argue that knowledge-how is analysable analogously to knowledge utterances containing other question words, Rumfitt brings up French, which does not use a question word, and Russian, with a completely different word for knowledge-how. He argues that any claims of how ‘know how’ or ‘know that’ function on a metaphysical level, unlike claims about language-specific particularities of vocabulary and grammar, cannot be grounded only in English and writes that “this is an area where a trawl through cross-linguistic data might be of real philosophical interest” (Rumfitt 2003: 165).

The distinction in philosophy between ‘know that’ and ‘know how’ is closely related to the distinction between declarative and procedural memory found in psychology and cognitive science (or declarative and non-declarative memory, of which procedural memory is one kind). Evidence for this divide comes in part from amnesiac patients: it has been demonstrated that patients who cannot form new declarative memories due to brain damage, may still be capable of learning new skills, including more cognitive skills such as reading
mirrored text (Squire 2004, Adams 2009). Some evidence also seems to suggest that it is possible to be able to acquire new declarative knowledge but not new procedural knowledge (Adams 2009).

Further complicating the issue is the relationship between knowledge-how and ability or skill. Some, like Ryle 1949, consider them to be the same, while others point out that there are many examples where we (at least in English) would ascribe knowledge-how but not ability. Consider for example a pianist who loses an arm: intuitively, many would still say that they know how to play the Moonlight Sonata, but are not able to do so (at least not for a while, until they learn how to play with one arm).

Abilities may be further subdivided into specific and general abilities, quoting Maier (2018): “Consider a well-trained tennis player equipped with ball and racquet, standing at the service line. [—] every prerequisite for her serving has been met. [—] Let us say that such an agent has the specific ability to serve. In contrast, consider an otherwise similar tennis player who lacks a racquet and ball, and is miles away from a tennis court. There is clearly a good sense in which such an agent has the ability to hit a serve: she has been trained to do so, and has done so many times in the past. Yet such an agent lacks the specific ability to serve, as that term was just defined. Let us say that such an agent has the general ability to serve.” That is, there may be a difference between what an agent is currently able to do, due to external circumstances, and what they would be able to do, if the circumstances were appropriate.

The distinction is reminiscent of the distinction sometimes made in the discussion of modality between participant-internal and participant-external possibility, using the term of van der Auwera & Plungian (1998). The latter often develops out of the former, which in turn has as a common source a word for ‘know’. As van der Auwera & Plungian (1998: 88) suggest, writing “it [=representation of meaning in semantic maps] gives the various meanings an air of categorical discreteness, while they in fact often function on a cline”, it is a reasonable view that a situation may not be characterised strictly as one type of modality, but rather as being positioned somewhere on a continuum between the two, which may further vary depending on the perspective taken by the speaker.

Another distinction sometimes found, pointed out by van der Auwera & Plungian (1998), is the distinction between learnt and inherent knowledge. In French, for instance, savoir is used for learnt abilities (such as reading or swimming) while pouvoir is used for inherent abilities. However, they also note that the distinction may be better put as between intellectual and physical.

To summarise, there is experimental evidence for a neurological division

---

1 An example frequently found in the literature with unclear origins.
between declarative and procedural memory, a distinction that appears to be echoed in some languages. There is also linguistic evidence of a distinction between what might be termed theoretical knowledge-how and practical knowledge-how or ability. We might posit a scale from purely declarative, propositional knowledge (e.g. “X knows that WWII began in 1939”) to non-cognitive ability (“X can lift 50 kg”) with cognitive skills coming somewhere in between (“X knows how to solve a second degree equation”). Note, however, that even what has here been termed non-cognitive ability contains a cognitive component—even lifting weights requires motor control, which is in some sense cognitive. Perhaps it is therefore better to postulate the scale from entirely cognitive to minimally cognitive.

The hypothesis for cross-linguistic encoding would be that this scale is divided into contiguous segments, but the exact cut-off points (and the number of categories) are variable. Related to this continuum, but at least theoretically separable, is the distinction between learnt and inherent skills. In addition to this, a distinction can be made between specific and general abilities, a scale which would seem to coincide partly with that of participant-external and participant-internal modality.

It is clear that the three dimensions of ‘know how’ and ‘know that’ which have been identified (practical-theoretical, inherent-learnt and internal-external) here are closely intertwined. An ability that is learnt is more likely to be more theoretical and in some sense also less sensitive to outside circumstances. It seems a reasonable prediction that natural languages will generally conflate these closely related dimensions. In French, then, it might be said that the distinction between savoir and pouvoir is influenced by all three dimensions (theoretical/intellectual–physical, learnt–inherent, specific–general). The exact semantics of the constructions found in languages is an empirical question, and one which may perhaps not be answered in the present study, since it seems likely that felicity judgments and elicitations of minimal pairs might be necessary as in most cases the dimension will at least partially coincide.

2.2.2 Personal and interpersonal knowledge

Similarly to examples 4a and 4b above, for the following sentences it can be asked if know means the same thing:

(5) English (Indo-European>Germanic, UK)


Hanna knows that penguins waddle
The question is perhaps even more pressing than that concerning ‘know how’ and ‘know that’ since English is something of an outlier in the European context in using the same word to express these two. In most European languages, including Swedish, a lexical distinction is made between propositional knowledge or knowing that (5a) and knowing persons (5b). In Swedish, *veta* is used for the first, whereas *känna* is used in the second, as seen in 6a-6b.

(6) Swedish (Indo-European>Germanic, Sweden)

a. Constructed example (based on 5a)

```
Hanna vet att pingvin-er rulta-r
PR.NAME know.PRS COMP penguin-PL waddle-PRES
```

‘Hanna knows that penguins waddle’

b. Constructed example (based on 5b)

```
Ted känner John
PR.NAME know.PRS PR.NAME
```

‘Ted knows John’

Similar distinctions are found, for example, in French (*savoir* vs. *connaître*), German (*wissen* vs. *kennen*) and Hungarian (*tudni* vs. *ismerni*). In the history of philosophy, it is mainly the ‘know that’ sense that has been discussed while the ‘know person’ sense has been largely ignored.

In principle, viewpoints analogous to those discussed in section 2.2.1 regarding ‘know that’ and ‘know how’ can be taken on the relationship between these two senses: one is reducible to the other or they are of different kinds. In the latter case, they can be seen as sub-species of the same super-ordinate category (that is, ‘knowledge’) (e.g. Wierzbicka 2018b) or they may be seen as separate (e.g. Farkas 2020).

Benton (2017) distinguishes between three grades of personal involvement in knowledge of people. The first two are cases of propositional knowledge, and it is only the third that is distinct. The first grade is second-hand known propositions of people, such as knowing about a famous person from reading about them. The second grade is knowledge of people from “perceptual access” to them, that is, knowledge by direct experience. In Benton’s view, this knowledge may be both propositional and objectual, that is, consisting only of stored experience and not propositions.
The third grade is what Benton calls interpersonal knowledge proper and consists of “reciprocal causal contact” and the knower’s treatment of the known as a subject, and vice versa. Benton gives the example of Julia and Juan who belong to the same committee and often go to the same large meetings, but have never spoken directly to one another. They know each other’s roles, names, appearances and so on: that is, they have much knowledge at the first and second grades of involvement. But for there to be knowledge at the third grade of involvement, something more is required. This component of interpersonal involvement is what makes personal knowledge non-reducible to and of a different kind from propositional knowledge in Farkas’ view; there is a non-cognitive aspect to it (Farkas 2020: 270).

It is interesting to note that the third grade of involvement appears to fairly accurately correspond to the meaning of the Swedish word käänna ‘know (someone)’, which always—when used without verb particles—means personal acquaintance. It is normally not acceptable in contexts that would be classified as the first or second grade of personal involvement in Benton’s account. In this, the Swedish verb differs from its counterpart in German. The Swedish variant in [8a] of the German [7] is not acceptable in the context: it is grammatical, but would mean that all readers are personally acquainted with Margrit Fuchs. A more accurate Swedish translation is given in (8b), with the particle verb käänna till (which is not limited to persons as objects).

(7) German (Indo-European>Germanic, Switzerland)

a. Oltner Tagblatt 2018 (Universität Leipzig, Sächsische Akademie der Wissenschaften & InfAI 2023)

Leser dieser Zeitung kennen Margrit Fuchs

‘Readers of this paper know (of) Margrit Fuchs’

(8) Swedish (Indo-European>Germanic, Sweden)

a. Constructed example (based on [7])

?Tidningen-s läsare käänner Margrit Fuchs

‘Readers of this paper know Margrit Fuchs’

---

1This example is from a Swiss newspaper
b. Constructed example (based on 7)

Tidningen-s läsare känner till Margrit Fuchs
magazine-GEN reader.PL know.PRS to PR.NAME

‘Readers of this paper know of Margrit Fuchs’

Interpersonal knowledge differs from other personal knowledge in another way, hinted at in the description above: it is normally symmetrical. If A knows B (in the third grade of involvement sense), B also, in most contexts, knows A. Indeed, Benton considers this a necessary condition for interpersonal knowledge and it follows from his view that interpersonal knowledge comes from reciprocal causal contact (2017: 18). The same is clearly not true of knowledge of propositions or of knowing of people.

Within the linguistic literature on reciprocals, a class of verbs for what is known as naturally reciprocal events—that is, events that are “either necessarily (e.g. ‘meet’) […] or else very frequently (e.g. ‘fight’, ‘kiss’) semantically reciprocal”—are often distinguished (Kemmer 1993: 102). Predicates belonging to this class, whose exact make-up varies somewhat from language to language, are often treated differently in expressions of reciprocity, either by receiving no overt marking at all or by receiving lighter marking. Generally, words like kennen are not mentioned among these (see e.g. the surveys in Evans et al. 2011). This is the case in Swedish; compare the examples in (9a) with (9b).

(9) Swedish (Indo-European>Germanic, Sweden)

a. Constructed example

De slog-s/krama-de-s/träffa-de-s
3PL hit.PST-S/hug-PST-S/meet-PST-S

‘They hit/hugged/met each other’

b. Constructed example

?De kände-s
3PL know.PST-S

*‘They knew each other’
¿‘They were known’

The -s morpheme is productive as a passive morpheme, but with some verbs — which ones is lexically determined — the reading is reciprocal instead, as shown in the examples in 9a (there are also other uses of the morpheme, see
This is not the case with kännan, as in example 9b, where the only possible interpretation is as a passive, even if this too is rather odd. Is Swedish typical in this respect, or do languages tend to treat forms for ‘know person’ as semantically reciprocal?

Another interesting feature of interpersonal knowledge is that it is itself gradable, as noted by Benton (2017: 17–18). We (usually) know our family better than we know our colleagues. Discussing the uses of English know with a person object, Benton reflects that there are contextual differences in when these are appropriate. In some cases, it might be appropriate to say that we know someone we have only met once, in others a longer acquaintance is necessary.

Here should perhaps be noted another sense of knowing, often referred to as “knowing someone Biblically” (or “carnal knowledge”), that is, sexual intercourse. Farkas (2020: 267) briefly discusses this sense and concludes that this cannot be regarded as a kind of knowledge in the normal sense, which is an intuitively appealing conclusion. As hinted at by the name, this colexification is a feature of the language of the Bible. Some translations keep it and some do not; it may vary between translations in the same language. Take for instance Luke 1:34, given here in two English versions.

(10) English (Indo-European>Germanic, UK)

a. Luke 1:34 (King James)

Then said Mary unto the angel, How shall this be, seeing I know not a man?

b. Luke 1:34 (Lexham)

And Mary said to the angel, “How will this be, since I have not had sexual relations with a man?”

In the King James translations we find know, but in the Lexham version (which, it should be noted “is a relatively literal translation which closely follows the original”) we find have sexual relations—both as translations of Greek ginosko ‘know’. On the understanding of interpersonal knowledge as being the result of “causal reciprocal contact”, the meaning extension is not very strange: sexual intercourse is clearly reciprocal contact, and we would be dealing with a semantic shift from the effect (knowing each other) to the cause.

---

1 Colexification refers to the expression of two functionally distinct senses by the same lexeme (such as mouse for both the animal and the computer accessory. The term was coined in François (2008) as an option neutral to the distinction between polysemy, vagueness and homonymy.
(reciprocal contact—in this case, sex). It seems likely that the use is motivated at least partly by taboo.

To summarise, a number of recent philosophical approaches have argued for the distinctiveness of interpersonal knowledge, a distinction for which there is also cross-linguistic evidence. In Benton’s account, knowledge of persons comes in grades: only the third constitutes interpersonal knowledge proper, whereas the first two do not. As seen for Swedish and German, languages may vary in which grades a lexical item encodes. Furthermore, interpersonal, that is third grade, knowledge is itself gradable, and what depth of knowledge is required for ascriptions to be acceptable is context-dependent. It was also noted that interpersonal knowledge is a mostly symmetrical and reciprocal relation: it is an interesting question whether the symmetry is ever encoded in the linguistic form.

Finally, the use of know to mean having sexual intercourse was noted. While this is undoubtedly an interesting case of meaning extension and taboo, it is not included in this investigation due to its rarity in the NT (there are two examples, both negative).

2.2.3 Objectual and experiential knowledge

As pointed out in the section above (2.2.2), German kennen ‘know’ is not limited to interpersonal knowledge, unlike the Swedish counterpart känna ‘know (someone)’. As is demonstrated by the following examples, it is also not limited to taking persons as its object:

(11) German (Indo-European>Germanic, German/Switzerland)

a. Wilhelm Meisters Lehrjahre (von Goethe 1795(1990))

   Kennst du das Land, wo die Zitron-en blüh’n?

   know.2SG.PRS 2SG DEF.N.ACC land where DEF.PL lemon-PL bloom.3PL.PRS

   ‘Do you know the land where the lemons bloom?’

b. Beobachter (Universität Leipzig, Sächsische Akademie der Wissenschaften & InfAI 2023)

   Viele Zahnärzte kennen solche Fälle.

   many dentist.PL know.3PL.PRS such case.PL

   ‘Many dentists know such cases’
The wider extension is quite common, at least in European languages, although it is not universal. The Swedish *känna*, without particles, is, as mentioned, largely confined to the interpersonal knowledge, including not being possible with non-person arguments. Swedish variants of [11a](#) and [11b](#) are thus doubtful:

(12) Swedish (Indo-European>Germanic, Sweden)

a. Constructed example (based on [11a](#))

```
??Känner du land-et, där citron-er-na blomma-r?
know.PRS 2SG land-DEF there lemon-PL-DEF bloom-PRS
```

‘Do you know the land, where the lemons bloom?’

b. Constructed example (based on [11b](#))

```
*Många tandläkare känner sådana fall
many dentist.PL know.PRS such case.PL
```

‘Many dentists know such cases’

In the examples presented so far, German has differed from Swedish in the use of a single form—*kennen* being used. What is of interest to us, however, and as was noted above, is that a lexical distinction similar to the Swedish one is made in German, as shown in the following example.

(13) [Fukuda](#) (1970: 91)

```
Ich *kenne* ihn, aber ich *weiß* nicht, wo
1SG know.1SG.PRS 3SG.ACC but 1SG know.1SG.PRS NEG where
er wohnt
3SG live.3SG.PRS
```

‘I know him, but I don’t know where he lives’

The most obvious difference here is the form of the known: in the case of *kennen* nominals and in the case of *wissen* clauses. This is not, however, an exhaustive description, as is clear from the following examples, taken from the main German-language lexicon ([Duden](#) 2023) where we see *wissen* used with nominal objects.
German (Indo-European>Germanic, Germany)

a. “wissen” (Duden 2023)

\[
\begin{align*}
\text{den} & \quad \text{Weg}, \quad \text{die} \quad \text{Lösung}, \quad \text{ein} \quad \text{Mittel} \\
\text{DEF.M.ACC} & \quad \text{road} \quad \text{DEF.F.ACC} \quad \text{solution} \quad \text{INDEF.N.ACC} \quad \text{means} \\
\text{gegen} & \quad \text{etwas} \quad \text{wissen} \\
\text{against} & \quad \text{something} \quad \text{know.INF}
\end{align*}
\]

‘know the road, the solution, a means against something’

b. “wissen” (Duden 2023)

\[
\begin{align*}
\text{jemandes} & \quad \text{Adresse}, \quad \text{Namen} \quad \text{wissen} \\
\text{somebody.GEN} & \quad \text{address} \quad \text{name} \quad \text{know.INF}
\end{align*}
\]

‘know someone’s address, name’

In general, it has been claimed that for both German (Fukuda 1970: 101) and French (Van Hulle 2015: 101) abstract nouns are preferred as objects by wissen/savoir and concrete nouns by kennen/connaître.

It should also be noted that although the two are of course closely correlated, there is in fact a potential disconnect between linguistic form and content. Take for example nouns such as answer, idea or indeed proposition itself. The linguistic form is that of a noun, but these clearly have what might be called inherent propositional content — that is, they do not denote physical things, as objects prototypically do, but rather propositions. Languages may, it appears, single out these nouns. Swedish veta ‘know’ is overwhelmingly more common with clausal objects and is usually not acceptable with nominal objects. In the Swedish language corpus Korp (Borin, Forsberg & Roxendal 2012), around 1% of the occurrences of veta are with a noun as an object and the top three objects are svaret ‘the answer’, sanningen ‘the truth’ and saker ‘things’. These can all be seen as denoting propositions despite their linguistic form.

However, neither this gives an exhaustive account of the distinction in German. In many cases, both German verbs may occur with the same object with more or less subtle variations in meaning, a recurring topic in blogs for (English-speaking) learners of German. The following examples are given in Schuchart (2023).
The two sentences in (15) are roughly equivalent, but (15b) has the additional notion of being familiar with the way through personal experience. Thus, *kennen* is knowledge through personal acquaintance, experience or familiarity, whereas *wissen* lacks this component (see also Fukuda 1970). Similar connections to personal experience exist for French *connaître* and Spanish *conocer* (Vatrican 2004).

It is interesting to note the parallels to evidential systems here. In the evidential systems of the world’s languages, first-hand evidence plays an important role; one of the most common distinctions made world-wide is between what has been experienced directly and what has not (Aikhenvald 2004). To be clear, I am not suggesting that the lexical distinction in, for example, German is an evidential system in disguise; I am only noting that some of the same parameters seem to play a role. This source-based meaning seems to be even clearer in some languages. Fortescue (2001: 19) writes that Hopi has a separate word for ‘knowing by observation’, as opposed to ‘knowing a fact’ and ‘knowing someone or how to’ and according to Dixon 1993, there is no clear counterpart to ‘know’ in Dyirbal, instead one must say for example “I saw that X” or “I heard that X”, as reported by Evans & Wilkins 2000: 570.

In psychology and cognitive science, a distinction similar to the personal experience dimension of *kennen* vs. *wissen* is made between memories stored as personal experience and other memories: this is the distinction between episodic and semantic memory (which are part of declarative memory, as opposed to procedural memory, discussed above, 2.2.1). The distinction was first suggested in Tulving (1972). Episodic memory refers to the memory of lived-through experience and is characterised by the experience of mental time travel in retrieval. That is, in recalling something from episodic memory, we can re-experience the original event. Semantic memory is knowledge about the world that is not tied to a specific experience. The separation of the two
is, as with procedural memory, supported by patients with brain injuries that have impacted one of the systems but not the other, as well as by neuroimaging experiments (Gardiner 2008, Nyberg 2008).

Note that within cognitive science and psychology, the difference between episodic and semantic memory has generally been considered to correspond to the distinction between ‘remember’ and ‘know’, rather than for example *kennen* and *wissen*. The semantic difference between *know* and *remember* in focus here is the subjective experience of the cogniser: the experience of mental time travel. It should be noted, however, that this does not exhaust the differences in meaning between ‘remember’ and ‘know’: there is, for example, ‘remember’ as ‘manage to retrieve from stored knowledge’ or ‘manage to attend to’. Whether the same meanings are commonly colexified in other languages is an open question.

To summarise, two parameters in the lexical organisation of the knowledge domain have been identified: the form of the object and personal experience. For the first, it was noted that there is a certain mismatch between linguistic form and content: in some cases, nouns are most appropriately regarded as having propositional content. The second parameter identified is that of personal experience. It is easy to see a functional motivation as to why this would be intertwined with object-form in lexical differentiation: often, we have personal experience of objects, but not of propositions. However, these are two distinct parameters. Personal experience has also been connected with the distinction between ‘know’ and ‘remember’. In this case, there seems to be no accompanying difference in preferred object-form, although there are other differences between the two.

2.2.4 Vertical and horizontal knowledge

In many languages, there is a lexical distinction similar to that between English *know* and *understand*; in this section I will argue that one of the differences expressed by the two terms is between what Mizumoto (2018) has (metaphorically) termed vertical and horizontal knowledge. The two terms are, of course, closely related, and can in many contexts be interchanged without any noticeable change in meaning. However, as the following examples show, they can also be contrasted:

(16) English (Indo-European>Germanic, UK)

a. People 2003 (Davies 2008: COCA)

   *Diego “knows his brothers [sic] gone,” says Janelle Portillos, “but he doesn’t understand.”*
b. Music Educators Journal 2003 (Davies 2008-: COCA)

Students may **know** concepts, but they do not **understand** them if their knowledge is not contextualized.

c. Technology Teacher 2003 (Davies 2008-: COCA)

*I knew* what assets I had, but I didn’t **understand** how to fully utilize them.

(17) Swedish (Indo-European>Germanic, Sweden)

a. 8 sidor (Borin, Forsberg & Roxendal 2012)


‘The pupils learn by rote, but they do not understand, say the experts of the School Inspection Authority’

In examples 16a-16b, a fact is described as known but not understood. This suggests that although the two words are often interchangeable, there are cases where one applies but not the other. Except for in 16b there is no indication of what is lacking for understanding: there, it is a question of the knowledge not being “contextualised”.

In 16c and 17, the relationship between knowing and (not) understanding is further elaborated. In 17 learning something by rote is contrasted with understanding it. The same is not as naturally done with *veta* ‘know’ being used instead of *förstå* ‘understand’ (18a). Neither does *kunna* ‘be able, know how’ which may normally be used with certain nouns (such as *kunna saker* ‘know things’) work (18b).

(18) Swedish (Indo-European>Germanic, Sweden)

a. Constructed examples (based on 17)

??Eleverna **lär sig sak-er utantill**, men **vet** pupil-PL-DEF learn REFL thing-PL by.rote but know-PRS dem inte 3PL.O NEG

‘The pupils learn things by rote, but they do not know them’
b. Constructed examples (based on 17)

??Eleverna lär sig sak-er utantill, men kan dem pupil-PL-DEF learn REFL thing-PL by.rote but can.PRS 3PL.O inte NEG

‘The pupils learn things by rote, but they do not know them’

In [16c] knowing a fact is contrasted with understanding the practical applicability of it. The difference here is also expressed in the object clause.

In [19a-20a] below, sensory perception\(^1\) is contrasted with understanding in a way that is parallel to knowledge and understanding in [16a-17] 20a is a somewhat special case, in that it concerns the distinction between perceiving speech as sound and perceiving and understanding it as meaningful language.

(19) Swedish (Indo-European>Germanic, Sweden)

a. GP 2004 (Borin, Forsberg & Roxendal 2012)

Jag **kände** men **förstod** inte hur **skadad** jag 1SG feel.PST but understand.PST NEG how injured 1SG var be.PST

‘I felt but did not understand how hurt I was’

b. Webbnyheter 2013 (Borin, Forsberg & Roxendal 2012)

Sauron **se-r** allt men **förstå-r** inte alltid PR.NAME see-PRS everything but understand-PRS NEG always vad det är han se-r what it BE.PRS 3SG.M see-PRS

‘Sauron sees everything but doesn’t always understand what it is he sees’

(20) English (Indo-European>Germanic, UK)

a. Dateline 1998 (Davies 2008-: COCA)

I can **hear** the voice, but I don’t **understand** the words.

\(^1\)Note that **känna** colexifies ‘know (person)’ and ‘feel’
From the examples presented, I argue that there are cases where speakers ascribe knowledge (Swedish vetande) but not understanding (Swedish förståelse). One of the differences between the two is aspectual: ‘understand’ appears to be more commonly conceived of as an achievement (see section 2.2.7 below for further discussion on the aspectual distinctions in the domain). This does not seem to be the difference in the examples above. Rather, the words are used to differentiate in what can be described as depth of knowledge or, adopting the terms of Mizumoto (2018), horizontal (‘know’) versus vertical (‘understand’) knowledge. Note the parallel with everyday use, seen in expressions such as “deep/shallow knowledge/understanding”.

Mizumoto (2018: 89) writes the following on the characterisation of this dimension, regarding the Japanese forms shitte-iru ‘know’ and wakatte-iru ‘understand’: “horizontal knowledge, is usually a matter of how much and how widely the subject has obtained information” whereas “vertical knowledge, is a matter of deep appreciation, so that we can go deeper under what is already within one’s ken or territory”. That is, horizontal knowledge is just a matter of having a piece of information within one’s mental storage, so to speak, whereas vertical knowledge requires more than this: some kind of metaphorical depth.

Note that Mizumoto explicitly argues against an exact equivalence between understand and wakatte-iru (e.g. Mizumoto 2018: 95) but from his discussion it is clear that he sees the vertical-horizontal dimension as something shared by the terms in both languages. Citing Kvanvig (2003) he also argues that understand and wakatte-iru share a sense of implying usefulness for actions (e.g. Mizumoto 2018: 95). Compare example 16c above.

The main thesis of the philosopher Kvanvig (2003) is that understanding is not a kind or species of knowledge and that understanding is valuable in itself (whereas knowledge, he argues, has no added value beyond the value of its parts). The difference between knowledge and understanding, in Kvanvig’s account, is that the former consists mainly of the relationship between a knower and single true propositions whereas the latter generally cannot be characterised by picking out a single true proposition that the subject believes, but is characterised by the discovery and creation of systematic relations between propositions that are within the subject’s knowledge (Kvanvig 2003: 185–203).

Within psychology, a similar distinction is sometimes made between deep and shallow processing in the context of memory encoding. The idea, proposed by Craik & Lockhart (1972) is that we remember things better if they are encoded via deep processing than via shallow processing. Experiments by Craik & Tulving (1975) appear to confirm this. Subjects were presented with words and then asked one of the following types of questions: 1) questions about the physical characteristics of the printed word, 2) questions about whether the word rhymed with another word and 3) questions about the mean-
ing of the word. These represent, from 1 to 3, successively deeper levels of processing. It was found that subjects remembered more words when they were asked questions of type 3. Several further experiments supported the thesis, but there is also counter-evidence, as discussed by Dinsmore & Alexander (2012) who attribute the conflicting evidence partly to the lack of clear definitions of what is meant by depth of processing. The same sentiment is shared by Goldstein (2014: 181) who writes that “[t]he idea of levels of processing motivated a great deal of research but became less popular when it became apparent that it was difficult to define exactly what depth of processing is” and that “[e]ven though the term levels of processing is rarely used by present-day memory researchers, the basic idea behind levels of processing theory - that memory retrieval is affected by how items are encoded - is still widely accepted, and a great deal of research has demonstrated this relationship”.

The above discussion has concerned the encoding of memories, as opposed to the storage of memories, the latter which is what the philosophical discussion of understanding and knowledge is about. However, evidence, as well as common sense, seems to suggest that there is a connection between the two, for example the research of Bower et al. (1969) which showed that when subjects were presented with a word-list structured in a meaningful taxonomic tree - that is, a form representative of deep processing - they not only recalled the words better than those who were presented with random trees, but also tended to recall the words using the same meaningful structure.

To summarise, in several languages, including English and Swedish as shown, understanding and knowing can be contrasted in ordinary language use, and I have argued that in some cases this corresponds to differences in what can be called the depth dimension of knowledge. Some philosophers have claimed that there is a difference in kind between knowledge and understanding, and that they are not reducible to each other. Knowing cannot become understanding just by increasing the number of pieces of information one possesses, that is, by widening the scope of knowledge. In Kvanvig’s account, what is needed is a grasp of the relationships between the pieces of knowledge and their implications. It could of course be argued that there is actually no reason to posit a difference in kind to explain this: are not relationships between pieces of information themselves information? On that account, the difference between understanding and knowing is not in the states themselves, but in what objects (propositions, and perhaps relations between them) are known.

Psychological research seems to support that the level of processing used when encoding memories affects memorisation. As discussed, there are unclarities in what exactly is meant in psychology by depth of processing, but it is clear that the core idea is similar to Kvanvig’s characterisation of understanding: the meaningful integration of possessed knowledge into the cognitive life.
of the subject.

2.2.5 Further aspects of understanding

In the above section, we saw that some factors that have been suggested to separate ‘know’ and ‘understand’ are depth, practical usefulness, and connections among the known facts. In this section, some more linguistically oriented analyses will shed further light on what characterises understanding. [Apresjan (2000: 171) gives the following description of Russian ponimat’ ‘to understand’ (enumeration added by me):

[...] at the basis of understanding lies [i] knowledge or imagination comprising [ii] fairly complex facts or situations. For this reason understanding [iii] requires a certain deductive effort which [iv] relies on the subject’s prior experience. The knowledge or notion obtained in this way [v] makes it possible to predict how the situation will develop further.

Note that this description covers only the construction ponimat’ chto, ‘understand that’. Apresian also gives an alternative explication from Boguslavsky (1984: 623) which is meant to cover also the other two constructions in which ponimat’ is found, namely ponimat’ + Interrogative clause and ponimat’ + Noun phrase (both of which are also found in English and Swedish). The explication is (boldface and enumeration added):

X ponimaet Y [X understands Y] = ‘[i, iii] the fact that X has processed or usually processes with a component of his psyche W, usually his mind, certain facts connected with Y has resulted in X (a) having or (b) [vi] beginning to have in his consciousness reliable information Z about [vii] the substantive features of Y’.

As Apresjan (2000) points out, this explication seems geared particularly toward the ponimat’ + Noun phrase construction. It is otherwise quite similar, containing the notion that understanding consists of having information and that this relation is the result of a mental process where the previous experiences of the individual are involved. Two points present in Boguslavsky’s explication but not Apresian’s are the possibility of ponimat’ as “beginning to have” a certain mental state and ponimat’ as relating to the “substantive features” of something.

These explications explicitly concern the Russian ponimat’. From a more general angle, [Fortescue (2001: 17)] gives the following characterisation of understanding, based on English but aiming toward “typologically general” (my enumeration).
When you think, you do so about what you know or imagine or believe to be the case, or what you remember or intend doing, and you may judge the object of your thoughts to display some quality; you may in the process get to understand it better. You understand something when you recognize it as corresponding to or relating to something you have previously experienced or learnt (and may thus act accordingly). So, when you have understood something, you know it and can remember it later and think about it again.

This characterisation bears many similarities to the explications by Apresian and Bogluslavsky, as well as the philosophical discussions of Mizumoto and Kvanvig. Understanding is characterised by creating connections between facts in the subject’s mind and previous experience. This means that understanding is an internal process, that is, not necessarily caused by external stimuli but by thinking about things. Fortescue also raises the same usefulness for actions as argued for by Kvanvig. Finally, there is also the idea of understanding as achievement.

Another aspect of understanding that has been brought forward is the intersubjectivity of it. Sweetser (1990: 43) writes, based on her etymological investigation, that “hearing is connected with the specifically communicative aspects of understanding” and the action of putting oneself in someone’s “emotional space”. That is, understanding appears to be connected specifically with understanding other individuals.

As we can see, there are several recurring points in these characterisations of understanding, despite their somewhat different viewpoints. The most important characteristics can be summarised as follows.

Understanding...

a. is like knowing (i), but deeper (iii)
b. consists in connecting new information with old information, or of creating new connections (iii, iv)
c. comes about by internal processing (not primarily external stimuli) (iv)
d. is applicable to the future and can guide action (v)
e. concerns the substantive qualities of its objects (vii), which are complex (ii)
f. is often an achievement (vi)
g. is intersubjective
2.2.6 Territory of Information and epistemic status

A core observation which is common both to Kamio’s theory of Territory of Information (e.g. Kamio 1997)) and the Conversation Analytical work on epistemic status (e.g. Heritage 2012) is that people in interaction do not behave as if they are equally entitled to speak about everything. It is, for instance, perfectly natural to utter 21a-21c:

(21) English (Indo-European>Germanic, UK)

a. Kamio (1997: 8)
   \[I \text{ am nauseated}\]

b. Kamio (1997: 5)
   \[That \text{ lady is my mother}\]

c. Kamio (1997: 2)
   \[I \text{ was born on September 29, 1942}\]

but replacing the first person pronoun with the second person in all cases results in a more or less unnatural utterance, 22a-22c.

(22) English (Indo-European>Germanic, UK)

a. Constructed example (based on 21a)
   \[??You \text{ are nauseated}\]

b. Kamio (1997: 5)
   \[??That \text{ lady is your mother}\]

c. Kamio (1997: 2)
   \[??You \text{ were born on September 29, 1942}\]

Circumstances can of course be imagined where the sentences in 22a-22c would sound natural, but in most cases they would be considered strange. Kamio’s explanation of this is that in 21a-21c the information lies in the speaker’s Territory of Information, whereas in 22a-22c it lies in the hearer’s Territory of Information. Direct, declarative utterances are only appropriate concerning one’s own Territory of Information. The Territory of Information is determined by the rather abstract notion of closeness to a certain piece of information: if a piece of information is closer to the speaker than a given limit, it belongs to their Territory of Information. This closeness may consist of direct experience,
(professional) expertise, information received from reliable third parties and so on. Kamio seems to consider the exact make-up of the notion of closeness to be culturally dependent.

The placement of a piece of information within or outside the speaker’s and addressee’s Territory of Information is something that is oriented toward by speakers in several ways. Kamio argues that it has significance in e.g. psychological utterances, demonstratives and a range of other constructions (see [Kamio 1997] especially ch. 4-6). [Heritage (2012)] argues that relative epistemic status (basically equivalent to Territories of Information) is the deciding factor in action formation, i.e. in letting interlocutors understand something as a statement, a request for information etc.

The theory of Territories of Information may be subsumed under Evans’ more general theory of multiple perspectives ([Evans 2005], as in [Bergqvist (2016)]). In many cases, Evans argues, more than one perspective needs to be considered when accounting for the choice between grammatical forms or lexical items. Evans gives the example of the false belief marker molkkûn(h)- in the Australian language Dalabon. This marker is attached to a verb and means that the event expressed is currently known to the speaker but unknown to someone else (who exactly is pragmatically decided; it may be the speaker themselves previously). This kind of marker may not be entirely uncommon: [Spronck & Vuillermet (2019)] and [McGregor (2023)] discuss mistaken-belief constructions in several Australian and South-American languages. Note that these by definition incorporate multiple perspectives: a belief can only be mistaken in view of an external party or a subsequent revision.

It should be stressed that Territory of Information and actual knowledge state are separate things. Something may be known by a speaker but not lie within their Territory of Information. Conversely, something may lie within a speaker’s Territory of Information but be unknown to them. [Bergqvist (2016)] argues that both factors—actual knowledge state and Territory of Information—are relevant in the choice of epistemic marker in Cogui (Chibchan, Colombia).

Highly relevant to the present study is the claim by [Lee (2006)] and, following her, [Mizumoto (2018)] that the difference between the two Japanese words shiru, usually ‘know’, and wakaru, usually ‘understand’, is in part explained by Territory of Information. Lee focuses in particular on examples of negating knowledge, but argues that the same is true in the affirmative. Consider examples 23a, 23b.
(23) Japanese (Japonic, Japan)

a. [Lee(2006: 192)]

Donna-hito-to kekkonshi-tai-ka zenzen wakara-nai
what-kind-person-with marry-want-Q at.all understand-NEG
/*shira-nai*/
/know-NEG

‘(I) don’t know what kind of person I want to marry at all’

b. [Lee(2006: 192)]

Oosaka-kara hakata-made kuruma-de nanjikan kakaru-ka
Osaka-from Hakata-to car-by how.many.hours take-Q
wakara-nai /shira-nai
understand-NEG /know-NEG

‘(I) don’t know how many hours it takes to get from Osaka to Hakata by car’

Shira-nai is not possible in 23a but is fine in 23b. Lee argues that the explanation for this can be found in the Territory of Information theory: if a piece of information falls within the speaker’s Territory of Information, wakaru is preferred to shiru.

Lee (2006) focuses on utterances with a first person subject, as does Mizumoto in his discussion of the Territory of Information part of the difference between the two predicates, writing that “in the case of the distinction of shitte-iru and wakatte-iru, we may think that it depends simply on whether the information falls into the territory of the attributee of the knowledge, rather than on whether it falls into the speaker’s or hearer’s territory” (Mizumoto 2018: 84), clearly suggesting that the distinction may be important also in utterances in the third-person.

Sadler (2010) tests this idea against a database of thirty recorded naturally occurring conversations in Japanese, totalling around three hours and ten minutes of analysed material. The study finds Lee’s claim to be largely backed up by the data: out of 99 uses identified as being within the speaker’s territory, 72% percent use wakaru and out of 30 uses outside the speaker’s territory, only 17% (the rest in both cases use shiru).

To summarise, we find in many languages a general orientation toward the Territory of Information of speakers and addressees that manifests in different areas of the language. In Japanese, the evidence suggests that this dimension is partly responsible for the distinction between the two knowledge-verbs wakaru and shiru. An orientation toward the similar, but distinct, dimension of the
knowledge states of interlocutors is also found in many languages and many
domains of language, such as false belief markers.

To the best of my knowledge, no other language than Japanese has been
claimed to make a lexical distinction in knowledge-verbs based on Territory
of Information. If there are, and if it is something that runs through the entire
language system and is not limited to first person utterances, it is a division
that should be observable in the corpus.

Constructions expressing a difference in knowledge state is unlikely to be
found within the domain of knowing: an utterance containing ‘know’ by defi-
nition aligns the knowledge state of the speaker and the one attributed knowl-
edge. The exception is of course the negation of knowledge: in this case the
knowledge states are by definition different. In languages where false belief is
marked, it is possible that this also extends to lack of knowledge—that is, just
as a marker may express that S believes P which I know to be false, it might
express that S does not believe P, though I know it to be true. In both cases,
the asymmetry in knowledge states is the focus of the expression.

2.2.7 Event structure

To know something, is, in the classic taxonomy of [Vendler (1957)] a state, one
of four ‘time schemata’ distinguished by Vendler which concern “the particular
way in which that verb presupposes and involves the notion of time” [Vendler
1957: 143]. It is this classification which most works on aspect take as their
point of departure. Aspect is, as pointed out by [Croft (2012: 31), “notoriously
vexing […] for semantic analysis” and the amount of literature on the topic
reflects this notorious vexation.

What is relevant for the purposes of this sketch of the parameters is the
observation, also going back to [Vendler (1957)] that knowing—in its typical
sense—is a state, but that the English know can also function as an achieve-
ment. In this section, this observation as made by Vendler will be outlined and
supplemented by the description of Croft.

There are verbs, Vendler writes, where it “is hardly possible to establish the
[aspectual] category to which they “originally” belong” [Vendler 1957: 152].
Among these, we find know. According to Vendler, know is in its dominant
use a state. States are distinguished by not being processes ongoing in time
(and thus not permitting the continuous construction in English—*I am being
Swedish) and by lasting for a period of time. Typical states are to know, to be
married or to love.

However, know also has a sense that does not appear to be a state as evi-
denced in sentences like 24.
Figure 2.1: Two-dimensional representation of the aspectual structure of ‘know’ after Croft’s (2012) schema, profiled part bolded

(24) English (Indo-European>Germanic, UK)

a. Vendler (1957: 151)

   And then suddenly I knew!

Here, know is an achievement. Achievements are momentaneous and bounded—typical examples are reaching (the top) and winning (Vendler 1957). To know something in the achievement sense, then, relates to knowing something in the state sense the same way getting married relates to being married. In some terminologies, this makes knowing what is sometimes known as an inceptive state—an event with both state and achievement senses (Croft 2012).

In the approach of Croft (2012), this duality is explained by the notion of construal (see section 3.1). Croft (2012) argues for a two-dimensional model representing event structure—the time dimension and a quality dimension, in accordance with his definition of aspect as “how events are construed as unfolding over time” (2012 p. 53). Time is represented on the x-axis of a diagram, representing the linear progression of time. The quality dimension is shown on the y-axis and represents the meaning that a predicate expresses. Depending on the properties of the predicate, the dimension may be different, e.g. continuous or binary.

The representation of knowing in this kind of schema is shown in figure 2.1a. For simplicity’s sake, we will assume for the moment that know is non-gradable and that the quality dimension thus has two values: ‘know’ and ‘not know’. When know is used in the state sense, this represents a construal—a way of viewing the event—in which the period of knowing is singled out—see figure 2.1b. If, on the other hand, the achievement sense is used, the speaker focuses on the moment in the event where there is a change from not knowing to knowing (figure 2.1c).
The arrows in the diagrams in 2.1 represent that knowing is a permanent state, in Croft’s 2012: 57–58 terminology. This separates it from a transitory state. It is also an acquired state—this can be seen by the fact that the change from not knowing to knowing is included, in the opposing inherent state a state is thought of as having no beginning. These characterisations, however, are subject to construal—although the possibility that we will cease to know something is usually not present when talking about knowing, there are contexts where it might be relevant and the aspectual structure would have to reflect this.

To summarise, then, there is both a state sense and an achievement sense to ‘know’—a well-known fact at least since Vendler (1957). In English, both may be expressed by the same form—know. Following Croft (2012) we can understand these senses as different construals of the same basic aspectual structure.

2.2.8 Summary and conclusions

In the above sections, a number of parameters of variation of knowledge have been identified and discussed.

Starting with differences in the object of knowledge, we can note that the type of knowledge most often discussed concerns knowledge of propositions. Some languages (such as English) treat knowledge of how to do things or of objects in the same way. The following three cases were discussed: knowledge how (this dimension appears to have the form of a continuum, ranging from propositional knowledge to physical ability, with theoretically oriented knowledge somewhere in between), knowledge of objects and knowledge of persons. For all of these, the difference can be viewed either as being only in the object of the knowledge state — that is, the state itself is essentially the same — or the states themselves can be considered to be different from each other. We saw suggestions that knowing a person is fundamentally different from knowing a fact, as well as pairs like German kennen and wissen being characterised by a presence versus lack of personal acquaintance.

The difference between knowing and understanding also at least in part has to do with the objects of the relation, where understanding concerns complex objects. There is also, however, a difference in the relation itself, where understanding is deeper and consists of making connections between known facts. These differences have led some philosophers to argue that knowing and understanding are fundamentally different, and that the latter cannot be reduced to the former.

The idea of Territories of Information has also been suggested to be relevant for the form of knowledge predication. Territories of Information, too, is
a dimension that could be said to concern the objects of knowledge, though it is oriented toward the social properties of facts—in whose Territory of Information does a fact lie?—rather than the more absolute differences between for instance a proposition and a person.

Finally, it was observed that knowing can be both a state and an achievement, depending on which part of its aspectual structure is in focus.

The preceding sketch of the knowledge domain is in large parts—especially when philosophy and cognitive science are concerned—based mainly on English. This is of course unfortunate, and it shows the need for a systematic investigation of the cross-linguistic expression of the field—which is the topic of this thesis. The aspects sketched above are a guide in choice of materials and analysis, but should not be understood as an a priori limit on possible ranges of variation.

Finally, the expected relation between these dimensions and lexical distinctions in any given language should be addressed. From the examples that have been discussed, it is clear that a linguistic distinction usually cannot be described as encoding variation in just one aspect. To take an example, consider Swedish förstå ‘understand’. As its English counterpart, it appears to encode the more vertical aspect of knowledge (as opposed to horizontal veta ‘know’), but it also differs in preferred aspectual usage—veta on its own is awkward as an achievement, while this is not the case with förstå. There are also several combinatorial differences between förstå and veta. There is every reason to expect that this multidimensionality holds also in other languages.
3. Investigating knowledge predication—theory and method

The previous chapter gave a preliminary sketch of the knowledge domain, based on research in fields such as philosophy, psychology and linguistics. A number of dimensions — such as depth of knowledge and intersubjective concerns — were identified as potentially relevant to the organisation of this domain in the languages of the world—the object of investigation for this thesis. This chapter describes the theoretical and methodological foundations of the investigation.

The chapter begins with a section outlining crucial assumptions about meaning, drawing mainly on cognitive linguistics. Concepts necessary to to explanations offered in later chapters are introduced and some key terms are defined. A section which deals with the senses in which domain—a central notion to semantic typology—can be understood follows (section 3.2), and I outline the three distinct uses of domain which will play a role in this thesis. Then, section 3.3 discusses the question of how the expression of meanings cross-linguistically can be compared and describes the approach taken in this thesis. Section 3.4 addresses some methodological questions particular to the approach taken—using a massive multilingual corpus in the form of New Testament translations. Sections 3.5 and 3.6 further describe the materials of the investigation. Section 3.5 describes how the contexts — i.e. the verses — for the study were chosen or, put differently, how the sketch of the knowledge domain in chapter 2 is operationalised. Section 3.6 specifies the objects of investigation on the formal side; what are the linguistic units that are of interest? The answer is constructions, and an overview of constructional approaches to language and constructional typology is given. The section concludes with a description of how the constructional approach was concretely used for coding the verses of the investigations (section 3.6.2).

The chapter concludes with a summary of the main points and a brief summary of the more specific methodologies used and discussed throughout the thesis.
3.1 Frames, prototypes and exemplars

This thesis aims to investigate meaning. It is therefore appropriate to make explicit here a few assumptions made, following work in cognitive linguistics, in the investigation about what linguistic meaning is and how it works. No attempt will be made to comprehensively cover the questions addressed and the discussion will be limited to what is directly relevant to this thesis.

One of the key insights in cognitive linguistics is that the meaning of a word can only be understood against a network of background assumptions (e.g. Lakoff 1987). As an example, consider a word like *weekend*. Understanding its meaning requires an understanding of the concept ‘week’ and to understand its use fully, an understanding of the industrial work-week, religious notions of holy days and so on is required. To take another, classical, example, *bachelor* can be defined as an unmarried, adult man, but leaving it at that obscures that any understanding of *bachelor* presupposes a society where the institution of marriage exists and it is expected of men to enter into it when they reach a certain age (and, indeed, presupposes the concepts of ‘man’ and ‘adult’). This notion of a concept being inseparable from the background against which it is understood is an extension of a more basic and abstract idea—that some concepts (e.g. the arc of a circle) require something more in order to be at all possible (a circle). The first is the *profile* and the second is the *base* Langacker (1987: 183–184), Croft (2002: 165).

Given this view, there is no way to make a sharp distinction between linguistic knowledge and world knowledge—thus, cognitive linguistics takes an encyclopaedic view of meaning, and is adopted here. Our knowledge of language is intertwined with our knowledge about the world—as Langacker puts it “meaning is equated with conceptualisation” (Langacker 1986: 3). But what, then, is conceptualisation? Conceptualisation, in cognitive linguistics, refers to an individual’s representation of experience. It is not merely an objective snapshot of reality, but is structured by the individual’s previous experiences, attitudes and categories of thought (see Lakoff 1987 for an in-depth discussion). The subjectivity becomes clear if we consider for example commercial transactions. An individual’s conceptualisation of a person handing another person a piece of paper with things printed on it as *payment* depends on pre-established categories and cultural knowledge.

Conceptualisation is based on concrete experience, but the conceptualisations which are the basis for how linguistic forms are applied and understood are abstractions and generalisations. As Langacker (1987: 1–2) puts it:

A lexical item embodies the commonality in form and meaning observable across a substantial number of *usage events* (i.e. actual utterances in their full phonetic detail and contextual under-
standing). Its acquisition comes about through the reinforcement of recurrent features, the progressive entrenchment of whatever aspects of form and meaning are constant across events. It thus involves a process of decontextualization, whereby non-recurrent features are filtered out, as well as schematization, for it is only by abstracting away from specific points of fine detail that commonalities become apparent.

I take this to mean that the meaning of a linguistic form is a generalisation and abstraction across individual conceptualisations. The view taken here is similar to the exemplar semantics described by Croft (2007) see also Wälchli & Cysouw (2012: 674) who characterise exemplar semantics as “a cover term for all approaches to semantics where exemplar meaning is considered more fundamental than the meaning of abstract concepts”.

An advantage of this model of meaning is that it offers an explanation for the often observed prototype effects in natural language categories. Discussion of categories has a long history (see van der Auwera & Gast 2011 for a survey). In what is known as the classical view, categories have been thought of as being defined by necessary and sufficient conditions and having all-or-nothing membership. Returning to the example of a bachelor, a classical definition is that it is an adult, unmarried man (e.g. Fodor, Fodor & Garrett 1975). Anyone who fulfils all these criteria is a bachelor, while anyone who does not is not a bachelor; there are no intermediate cases, and there are no more bachelor-like bachelors than others.

Taking off in the 1970s and pioneered in particular in the work of Eleanor Rosch, another theory of categorisation has gained considerable importance in (particularly cognitive) linguistics, namely prototype theory.

Rosch showed in psychological experiments that participants did not behave as if the assumptions of the classical theory were true; they considered some members of a category better examples of it than others, and the ratings were fairly consistent across subjects and test conditions (among others Rosch 1978, 1975, Rosch & Mervis 1975). A domain where this was observed early on was the colour domain, where the seminal study of Berlin & Kay (1969) showed the existence of cross-linguistic focal hues, that is, hues which are better examples of colour words than others. We return to this study below. This degree of representativeness was subsequently shown in a large range of studies to correlate with several psychological and linguistic variables. To mention some, it was shown that the more representative an object is of a category (as determined by the rating experiments mentioned above) the quicker participants are to judge it as part of that category. Another finding was that if there are two objects, A and B, which are members of a category X, and A is more
representative of X than B, B will be judged as more similar to A than vice versa.

Similarly, it was shown that some categories have cases of intermediate membership, contrary to what the classical model predicts. This was shown both between subjects and within subjects. It is worth noting that these two characteristics—non-equivalence of members and indeterminate membership—do not necessarily co-occur.

In short, much research has shown that many categories which people use do not appear to respect the classical structure of equivalence of members and all-or-nothing membership. Rather, they are characterised by a non-equivalence of members in terms of representativeness and in some cases also cases of indeterminate membership. In other words, they have a prototype structure—one or more paradigmatic cases surrounded by less paradigmatic cases.

It was early on in the research on prototypes pointed out that it is important to distinguish between prototype effects and the possible sources of these in the underlying representation of the category (e.g. Rosch 1978, Lakoff 1987). What was discussed above is prototype effects. These may be the result of different underlying structures.

Perhaps most commonly associated with prototypes is structure by family resemblance. This notion goes back to Wittgenstein (1953) who argued that for some categories (such as game) it is not possible to give a set of defining characteristics shared by all members of the category. What connects the category instead, is that each member shares some characteristics with some category members, but few if any characteristics are shared by all category members (and if they are, they are not enough to be distinctive of the category).

As another kind of structure underlying prototype effects, Lakoff (1987) introduced the notion of radial categories, which are hypothesised to underlie polysemous categories. As an example, Brugman & Lakoff (1988) consider the English preposition over. This word has a large range of connected but distinct meanings, for example, what Brugman & Lakoff (1988) call the above-across sense (this is the central sense in their analysis), the above sense, the reflexive sense and so on. What is characteristic of a radial category is a central case with connected subcategories which are variations on the central case. The variations are neither predictable from nor completely unrelated to the central case; rather, they are motivated by it, through semantic shifts. An interesting feature of radial categories is that each variant subcategory can itself have a prototype structure.

There is a clear connection between the view of meaning described above and prototypicality. Categories will be prototypical because they are centred around an idealised abstraction—a prototype—to which new experiences are
compared. Some experiences will be more similar to the prototype, and others less.

Despite its wide use in linguistics as a whole and cognitive linguistics in particular and the vast amount of literature on prototypes, there is, to quote Koptjevskaja-Tamm & Vejdemo (2020) “on the whole relatively little discussion of prototype semantics and focal meanings in cross-linguistic research on domain categorization”. The main exception is the colour domain, and to some extent the domain of containers. Within the colour domain, there is good evidence that categorisation is centered around a number of focal hues. In the body of research that began with Berlin & Kay (1969) (see also Kay et al. 2011) it has been shown that although languages vary considerably in how many basic colour terms they have, the range of variation is quite constrained—colour terms seem to center around a number of focal hues, regardless of how many colour terms there are in the language.

The experiences expressed by the kind of expressions investigated in this thesis are arguably more complex than colour. Empirical work, however, shows, as Koptjevskaja-Tamm & Vejdemo (2020) point out, that for most domains investigated so far, the variation in actual cross-linguistic expression can be explained by a relatively small number of parameters. Koptjevskaja-Tamm & Vejdemo (2020) further suggest that the organisation of (some of) these domains can, like colour categorisation, be understood via prototypes. By positing the existence of a limited number of prototypes capturing salient variations along the relevant dimensions, the expression of experiences across languages can be understood.

What differentiates this from straightforward componential analysis (i.e. characterising contexts as +/- whichever dimensions turn out to be relevant) is the recognition that the different dimensions are interdependent. To quote two proponents of the frame-based approach of the Moscow lexical typology school, Rakhilina & Reznikova (2016: 104) “they [=the features of situations] fall into gestalts, often so closely knit together that selecting one argument restricts the others.” A given domain, therefore, is structured around a number of prototypes representing the salient and relevant combinations of features. Following Rakhilina & Reznikova (2016), such prototypes will be referred to as frames.

The notion of frames comes from the work of Fillmore (e.g. 1975, 1977, 1982). In his early work, Fillmore distinguished scenes and frames. Scenes, in Fillmore’s account “include not only visual scenes but also familiar kinds of interpersonal transactions, standard scenarios defined by the culture, institutional structures, enactive experiences, body image, and, in general, any kind of coherent segment of human beliefs, actions, experiences or imaginings” (Fillmore 1975: 124). Frames are the linguistic correlates of scenes—for in-
stance collections of words and rules of verbal behaviour. Later applications have generally abandoned this distinction.

In this thesis, I will use *frame* to refer to the idealised and schematised prototype which is assumed to underlie the meaning of linguistic forms, and *scene* to refer to the individual conceptualisations which the *frames* are based on and used to categorise. The relationship is reminiscent of *types* to *tokens*, thus:

**Definition 1.** Scenes are the conceptualisations of individual situations

**Definition 2.** Frames are the idealised and schematised prototypes through which categorisation functions

Cross-linguistic regularities in expression suggest that the frames which make up an individual’s world knowledge are not just haphazardly collected but are organised in a structured way which can be described metaphorically in spatial terms. For example, Haspelmath (2017) shows that there are restrictions in how forms map onto indefinite pronoun functions and that these restrictions can be stated in the form of a semantic map, i.e. a graphical representation of spatial structure. It is often assumed that this structure—more or less directly—reflects an underlying mental structure, often referred to as conceptual space (e.g. Croft 2001) (see section 7.2 for other views of the relationship of semantic maps to cognition).

**Definition 3.** Conceptual space is the metaphorical space defined by the frames making up an individual’s knowledge and the relations between them

Note that this definition gives an individualistic view of conceptual space, contrary to e.g. Croft (2003), according to whom there is “a common conceptual space, which represents a common human cognitive heritage, indeed the geography of the human mind” (p. 139). It seems safe to assume that individuals have highly similar conceptual spaces—which means that speaking of conceptual space in the abstract without reference to a specific individual is still meaningful—but I do not believe that a common conceptual space can be assumed.

A further important point about frames is that not every part of the frame corresponding to a form is equally important when the form is used. A form like *walk* presupposes a frame which contains a person walking, something that is walked on, a direction and so on. What is expressed by the form, however, is the act. This can be explained by the notion of construal. For any given situation, there are a large number of ways it can be represented or construed. To take again the example of commercial transaction, this can be seen
from the point of view of the person handing over the money, the person getting the money, the money itself, the items expected to change hands after the money has been exchanged so and so on. The different construals correspond to different linguistic expressions (e.g. *bought* vs. *sold*).

The notion of construal is central in explaining many phenomena in cognitive linguistics and there are many different kinds of phenomena which fall under the construal-heading, with different ways of categorising them (see e.g. Verhagen 2010, Croft & Cruse 2004: chapter 3). What is relevant to us here is the process Langacker calls *prominence* with the particular sub-type *profiling*. This is the construal process which allows us to pick out a particular part of a scene as the profile—i.e. the focused, referred to part. As an example, consider the situation of someone admiring someone else. We can choose to focus here on the person doing the admiring—this corresponds to the noun *admirer*—or on the relation itself—this corresponds to the verb *admire*. This is shown schematically in figure 3.1, after Langacker (2010: 436).

Properly then, a linguistic form expresses a profiled part of a frame. Importantly, however, the profiled part presupposes the background—in this, the frame is necessary for the linguistic expression. As hinted at above, the understanding in cognitive linguistics is that the frame implied is quite broad—as Fillmore puts it: “nobody can really understand the meanings of the words in that domain who does not understand the social institutions or the structures of experience which they presuppose” (Fillmore 1982: 31). He takes as an example the various concepts connected to commercial transactions—*buy*, *sell*, *sale* etc. These can be understood only in relation to the scene of commercial transactions—in other words, understanding a word like *sell* presupposes the commercial transaction scene and all that it includes in terms of cultural understanding (money, value etc.)
This view of meaning pinpoints another source of prototype effects in categories. The idea, then, is that categorisation proceeds from the comparison of encountered scenes with available frames. In some cases the scene will fit quite well with a frame, but in others, if may be more difficult to fit a scene to a frame—these will be less good members of the category. To return to the example of the bachelor, we generally do not feel that the Pope or a Tarzan-like man raised in the jungle are good examples of bachelors (Lakoff 1987: 70). This is a mystery if we accept that the classical definition—unmarried adult man—is adequate.

The view of categorisation through prototypical frames, however, offers an explanation; the reason that the Pope or Tarzan make poor bachelors is that the scenes in which they are situated are poor matches for the frame against which bachelor is understood. Lakoff (1987) offers a similar explanation, but in terms of Idealised Cognitive Models. These are largely equivalent to what I understand by frames.

To summarise the view of meaning presented here, then, we can say that linguistic forms are associated with frames. Frames are abstractions across individual instances of conceptualisation—what I call scenes—and represent salient and relevant kinds of situation. The frames are also the basis for the categorisation of new scenes. A scene is judged as belonging to a certain frame and thus is expressed by the form associated with that frame. Scenes may be more or less similar—this gives rise to prototype effects in categories.

3.2 Conceptual, comparative and generalised domains

Closely related to categories are domains. The domain is an important concept in cognitive linguistics, but it also often figures in semantic typology. To give a couple of recent examples, Georgakopoulos et al. (2022: 439) “focus on the domains of perception and cognition” and Koptjevskaja-Tamm (2022: 125) considers “the close interaction between lexicon and grammar in the encoding of the temperature domain across languages” (my emphasis). But what does it mean when, in semantic typology, we speak of the X-domain or the domain of X?

To properly understand the use and application of term domain in semantic typology, it is necessary, I will argue in this section, to distinguish between three distinct—albeit interacting—senses of domain. The first I will call the conceptual domain sense, the second the comparative domain sense and the third the generalised domain sense.

By conceptual domain, I mean domain as a coherent and connected area of conceptual space, similarly to how domain is usually understood in cognitive linguistics. For instance, Langacker (1987: 488) defines domain as a
“coherent area of conceptualization relative to which semantic units may be characterized.” Another definition along these lines is Croft’s (2004: 15); “a semantic structure that functions as the base for at least one concept profile”. In this definition, domain is closely related to the notions of base and frame, discussed in 3.1. A domain, in this view, is the body of conceptualisation against which a concept is understood—to repeat an example from 3.1, the domain of weekend is the work-week with all its cultural baggage. Domains, then, represent cohesive areas of conceptual space.

The conceptual domain, in the sense I intend it, however, is not limited to domain as a conceptual base. The essential part of the definition is that a domain identifies an area of conceptual space on the basis of this area’s internal structure. The following definition may be given.

Definition 4. A **conceptual domain** is a coherent area of conceptual space

In addition to the domains of cognitive linguistics, also such uses as those of Nida (1979), Louw & Nida (1988), and Mel’čuk (2008/1988) where domains are established by shared semantic components are included here. Also the lexical fields of authors such as Trier (2008/1934, 1931) would qualify as conceptual domains under this definition. To refer to the cognition domain is to identify a coherent area—by whatever criteria one chooses—of conceptual space.

This characterisation of domain, however, does not fully capture how the concept functions in semantic typology. In the onomasiological approach to semantic typology, the researcher takes as a starting point an area of semantic space and then investigates how this is expressed linguistically. The area is frequently pointed out by reference to a domain—as in the quotes above. The function of domain in such cases is mainly the delimitation of an area for investigation. In this sense, it is, I will argue, similar to Lazard’s arbitrary conceptual frameworks (Lazard 2002) or Haspelmath’s comparative concepts (Haspelmath 2010).

The comparative concepts of Haspelmath are conceived against what he calls categorial universalism, where it is assumed that there is a universal set of linguistic categories for languages to choose from—linguistic categories in particular languages instantiate universal categories of the same kind (Haspelmath 2010). This position is mainly associated with generative linguists (see Newmeyer 2007) but is also implicitly assumed by some functionally oriented linguists (e.g. Dixon 2010). On this view, a cross-linguistic comparison can be made on the basis of these universal categories.

At the opposite pole is what Haspelmath (2010) calls categorial particularism, where categories are seen as language-specific. A category in a language
can only be described and understood in relation to that specific language system. This is a view which goes back to the structuralists of the early 20th century (Haspelmath 2010: 664, see e.g. Boas 1911) and has in some form been adopted by many linguists. In this view, cross-linguistic comparison becomes a problem. If categories are defined entirely language-specifically, how can we compare, say, the Swedish Subject and the English Subject? The answer, in the views of Lazard (2002) and Haspelmath (2010), lies in the fact that what languages do have in common is semantics.

Semantics, however, in the view taken by Lazard, consists of “amorphous meaning content, liable to any kind of arbitrary modelling” (Lazard 2002: 147). It is this arbitrary modelling which linguists use to make cross-linguistic comparisons; they construct arbitrary conceptual frameworks to impose structure on the “amorphous meaning content”. Lazard stresses that the frameworks “are not haphazardly chosen”. They have their basis in previous experience—“common sense, general knowledge of the world, personal opinions, even philosophical considerations” (Lazard 2002: 148) and what the linguist already knows about how languages tend to work. Lazard also stresses the provisionality of these frameworks; they are tools which if they fail to produce results can be refined or discarded.

A similar suggestion is made by Haspelmath (2010) with his comparative concepts. Since categories in different languages cannot be directly compared, cross-linguistic comparison requires defining universally applicable conceptual-semantic concepts—comparative concepts. The linguist then matches language-specific categories to the comparative concept.

To give an example, as a comparative concept—i.e. universally applicable and not making any reference to language-specific criteria—the dative case is defined by Haspelmath (2010: 666) as “a morphological marker that has among its functions the coding of the recipient argument of a physical transfer verb (such as ‘give’, ‘lend’, ‘sell’, ‘hand’), when this is coded differently from the theme argument.” This means that the Russian Dative and Finnish Allative are instances of dative cases whereas, for example, the Nivkh Dative-Accusative does not count as a dative—since it does not fulfil these criteria.

What these two approaches importantly have in common is their view of the linguist defining a structure imposed on conceptual space as a tool for cross-linguistic comparison. Domains in semantic typology often fulfil the same function; they point to an area in conceptual space the linguistic encoding of which we are interested in. In the extreme version of this view, pointing to a set of functions as a domain includes no claims about internal coherence or relatedness.

I suggest that a domain in this sense might be referred to as a comparative domain, analogously to comparative concepts. The main different to a
comparative concept lies in the degree of internal complexity—a comparative concept specifies one function, while a comparative domain a set of functions or meanings:

Definition 5. A **comparative domain** is a set of functions defined by the researcher as a tool for linguistic investigation

This is a similar notion to what is described by de Haan (2010: 105) when he defines *domain* as “a set of two or more functions”. Note that de Haan requires these functions to be *connected*—a domain is, more specifically, “a contiguous area in the semantic space” (de Haan 2010: 104). To know that a set of functions fulfil the criterion of being contiguous clearly requires knowledge of how semantic (or conceptual) space is structured.

In this sense of being analogous to a comparative concept, a domain cannot be right or wrong (cf. Haspelmath 2010: 665)—this is also Lazard’s view of arbitrary conceptual frameworks, see Lazard (2002: 148).

In answer to the ideas of Lazard and Haspelmath that categories are primarily *constructed*, it has been argued that some categories can in fact—without assuming categorial universalism—be *discovered*. Dahl (2016) argues that in doing bottom-up typology—typology which “builds on comparable cross-linguistic datasets based on primary linguistic data” (Dahl 2016: 434)—and comparing the distribution of markers in e.g. a text (or the answer to a questionnaire or a set of stimuli) we can find pattern clusters—markers that pattern in similar ways. If a cluster contains patterns from a sufficient number of varied languages, there is reason to posit what Dahl calls a *cross-linguistic gram type* (Dahl 2016). These “are not primarily entities “constructed” by the researcher to be used for comparisons but objects that are “out there” to be discovered” (Dahl 2016: 435).

A similar approach concerning the bottom-up construction of domains is argued for in de Haan (2010) in whose view domains, in the non-arbitrary sense, are secondary to functions. The construction of a semantic map (which is what they are concerned with) should start with functions rather than with a pre-determined domain; only as a secondary step should domains be mapped, “if we find that a given meaning range on a semantic map occurs over and over again in the world’s languages, then we can assign domain-status to that segment of the semantic map because it is obviously linguistically salient” (de Haan 2010: 116).

On this view, then, a particular type of cross-linguistic domain can be discovered through the comparison of many language-specific patterns. Importantly, this differs both from conceptual domains and the domain as a tool for cross-linguistic research—comparative domains. I propose that such a domain
can be referred to as a generalised domain, adapting the term generalising concept from Dahl (2016):

**Definition 6.** A **generalised domain** is an abstraction from regularities in the categorisations of individual language varieties.

By introducing this terminology we can distinguish between *domain* in three senses: **conceptual domains**, which are defined directly by reference to the structure of conceptual space, **comparative domains**, which are defined by linguists as tools of research and finally **generalised domains**, which are generalisation and abstraction based on the categorisation of individual languages.

It might be thought that the difference between these kinds of domains is not as radical as the use of distinct terms suggests. For instance, as both Haspelmath (2010) and Lazard (2002) emphasise, for comparative concepts or arbitrary conceptual frameworks to be successful, they must be based on consideration of empirical evidence, be it “common sense, general knowledge of the world, personal opinions, even philosophical considerations” (Lazard 2002: 148). This is clearly also the case for comparative domains, and in many cases the use of *domain* can be read as either comparative or conceptual. Returning to the quotes from Georgakopoulos et al. (2022) and Koptjevskaja-Tamm (2022) which introduced this section, cognition, perception and temperature are comparative domains in being the delimitation of the object of investigation. But it is also easy to see that these domains have internal coherence, and in this sense they are conceptual domains.

The reason for this correlation is that we investigate something as a domain because we expect there to be connections within it, that is, we view it as a conceptual domain.

Nonetheless, the distinct functions and roles in the investigation merit keeping the senses apart. A comparative domain is mainly a tool in the delimitation of the area of investigation. As such, it is not a result and it is also not a claim to anything beyond being a tool of research. Generalised domains, however, are concrete results of empirical cross-linguistic investigations; they are generalisations over language-specific categorisations. Conceptual domains can also be results, in contrast to comparative domains, but unlike generalised domains they are not primarily claims about cross-linguistic facts but about conceptual space(s). However, we expect—as pointed out by de Haan—that the cross-linguistic facts expressed by generalised domains reflect facts about conceptual space. Thus, claims about both generalised and conceptual domains may result from the same kind of investigation, but, importantly, they are claims about different things.
3.3 Connecting meaning and form

In the definition of Evans, semantic typology is “the systematic cross-linguistic study of how languages express meanings by way of signs” (Evans 2010). The previous sections have dealt with meaning; below forms, in the shape of constructions, will be discussed. This section deals with connecting the two, or, in other words, how can we investigate what Evans defines as the object of investigation for semantic typology—“how languages express meanings by way of signs”? Three central problems in any such investigation within semantic typology are identified by Koptjevskaja-Tamm (2012: 383), 2008 p. 43-44: the problem of data collection, the problem of identifying the same meanings across languages and the problem of metalanguage. In this section, these three problems are introduced, largely following the account of Koptjevskaja-Tamm (2008) and the approach to them used in this thesis is described, thereby situating this investigation within semantic typology as a field.

The first problem, then, concerns the data used in the investigation; what kind of materials should we study and how do we get hold of them? Unlike in traditional grammatical typology, secondary sources such as grammars and dictionaries are relatively little used in semantic typology. This is because it is in general difficult to find the kind of information needed for most non-grammatical domains in these works, regardless of how extensive the works are (Koptjevskaja-Tamm 2012). Therefore, studies in semantic typology generally need to collect data in other ways, more suited to the research questions posed in semantic typology. Methods that have been used include questionnaires (e.g. Rakhilina, Ryzhova & Badryzlova 2022), elicitation using non-linguistic stimuli (e.g. Majid et al. 2007) or texts (e.g. de Swart, Tellings & Wälchli 2022). In the case of texts, the use of parallel texts—i.e. the same content translated into two or more different languages—has a special position, due to parallel texts’ immediate comparability and the fact that they are generally more accessible to a researcher not fluent in the language. These ways of doing typology have come to be known as primary data typology (Wälchli & Cysouw 2012, Koptjevskaja-Tamm 2012) and can be contrasted with the more traditional grammatical typology using reference works — secondary data typology. An exception to the trend of using primary data in semantic typology which may be mentioned is the development of large-scale lexical typological studies applying quantitative analyses to large databases (e.g. Georgakopoulos et al. 2022).

The inadequacy of grammars and dictionaries as sole information sources is very clear in the specific domain under study in this thesis, in that it is not possible to get the necessary information about the use of words for knowing from grammars or dictionaries. The most common word for ‘know’ will gen-
erally appear also in a fairly short grammar, since it is usually a frequent word, but there will not be information on how this word compares to other possible words within the domain in the language. This is true for semantics generally. The same is true of word lists and dictionaries; even if all the relevant forms in a language can be found, there is insufficient information about their respective uses.

This leaves us with primary data typology of some kind as a methodological option. Non-linguistic stimuli are very difficult (if not impossible) to design for the knowledge domain. This leaves the use of questionnaires or texts. Both methods have their advantages and disadvantages. Questionnaires can be designed to eliminate as far as possible source language interference (for example by, as done in the study of Dahl (1985) on the expression of tense-aspect, by giving the terms of interest neutralised for TAM category). Using parallel texts, this is not possible and there may therefore be thought to be a greater risk of interference from the source language. For questionnaires, the researchers themselves can also select those native speakers or language experts whom they believe to be suited to the task. This selectiveness is more difficult when using parallel corpora, as the researcher often has only limited access to information about the person or persons responsible for the translation.

The main disadvantages of the questionnaire method are that it is very time-consuming (in terms of designing, distributing, collecting and analysing questionnaires) and that the language sample is limited by who the researcher has or can establish contact with. It is here that the advantages of using parallel texts become clear. Since the work of translation is already done, as is in some cases the work of collecting these translations (e.g. in the movie subtitle corpus ParTy by Levshina 2016), using parallel texts can give immediate access to a large collection of data. For example, there are approximately 1700 translations of the New Testament or portions thereof, representing around 1300 languages in the corpus available to me (Mayer & Cysouw 2014). This represents a coverage that cannot in any reasonable time be achieved by using questionnaires. For an exploratory project aimed at world-wide variety, this makes the use of New Testament translations an attractive solution. There are of course challenges with using parallel texts in general as typological data and New Testament translations in particular; these are discussed in section 3.4.

The second methodological problem that is identified by Koptjevskaja-Tamm (2008, 2012) is central to all typology but perhaps especially so in semantic typology; how do we know that the formal units picked out in different languages mean the same thing? Put differently, how do we know that we are comparing apples with apples, and not apples with oranges (Koptjevskaja-Tamm 2008)?
A popular solution in semantic typology has been to use non-linguistic stimuli. This stimuli is presented to speakers of different languages who are asked to name or describe them. The most famous example is perhaps Berlin and Kay’s (1969) study on colour vocabulary using wooden chips of particular colours. This is also the approach of the so called Nijmegen group, where semantic domains are represented by pictures (e.g. spatial location) or by video clips (e.g. putting and taking) which are used as elicitation materials (Levinson et al. 2003, Majid, Boster & Bowerman 2008). The central idea of these approaches is that by comparing the extension of forms over non-linguistic stimuli, we can compare meanings. The stimuli work as a denotational or etic grid which serves as a neutral ground for comparison (Levinson et al. 2003 487, see also Evans 2010).

This approach has been challenged on several grounds. Firstly, the neutrality of this approach depends on the neutrality of the etic grid chosen and even in a case where this seems entirely straightforward—the colour domain—it has been argued that the etic grid of coloured chips in fact represents an English or broadly Western view of the phenomenon (Lucy 1997). Thus, the view of the colour domain through this etic grid is not a view through a neutral tertium comparationes ‘the third [part] of the comparison’ but a view through a particular cultural conception of the domain. It is easy to see how this problem is even greater in cases where there is no obvious physical parallel to the domain, but the approach has challenges even for very concrete domains. How can we know which different situations or contexts are expressed linguistically? As an example, how do we know that it is relevant to elicit descriptions of something in a container with contact (e.g. a fruit in a bowl) vs. without contact (e.g. a fly in a bowl) but not of something in a circular vs. an oval bowl? The answer has generally been to treat the construction of an etic grid as an empirical question and gradually develop and refine the representation of a domain as new data come to light (see e.g. Levinson et al. 2003, Rakhilina, Ryzhova & Badryzlova 2022).

The second challenge is that context-free stimuli are not always appropriate or sufficient for getting at what is truly the meaning of words. For the domain of colour, for example, Wierzbicka (2018, 2015, also Goddard & Wierzbicka 2013 ch. 4) argues that colour-terms in many languages cannot be divorced from the environments in which they occur. For example, in Russian there are two terms which roughly correspond to the English brown—koričnevýj and buryj—which differ partly in what referents they can be used for. The former cannot be used of eyes or hair, for example (Wierzbicka 2018a). Sometimes there are deep-rooted cultural associations of words used to describe colour—something which is obscured by an exclusive focus on extension across an etic grid (Lucy 1997, Wierzbicka 2018a).
Thirdly, and perhaps most importantly, reference is not necessarily the same thing as meaning, in most people’s, including many linguists’, understanding of the word. This observation goes back to Frege (1892, for an overview see Zalta 2022: section 3) who distinguished between sense and reference (Sinn and Bedeutung). For example, the Morning Star, the Evening Star and Venus all have the same reference (a certain planet) but they do not all have the same sense. There is a mismatch between reference or extension and sense or intention.

Evans (2010) discusses this in terms of etic and emic representations — the extension of a form over the set of all possible referents vs. the language-specific characterisation of the underlying category. As an example, in Kayardild (Tangkic, Australia) the word kularrind can refer to an older brother of a female, a younger brother of a female, an older sister of a male or a younger sister of a male. This is the extensional meaning. But the intentional meaning of the word when used by speakers of Kayardild is hardly appropriately represented as this kind of disjunction; rather, it would be something like “opposite-sex-sibling” (Evans 2010: 510). In this case, the move from etic to emic or from extension to intention is fairly easy and uncontroversial, but there are cases where the move is more difficult. One such example is discussed by Koptjevskaja-Tamm (2008: 12): the words teplyj (Russian) and ljummen (Swedish), both approximately ‘tepid’. The denotational range — i.e. what temperatures the terms can cover — are about the same, but the descriptive meaning is very different. Teplyj is oriented toward warming (i.e. more teplyj equals warmer) and has positive connotations in metaphors. Furthermore, it is defined by reference to the human body. Ljummen, on the other hand, lacks a clear orientation (i.e. more ljummen equals warmer or colder, depending on context) and is negative or neutral in metaphorical contexts. It also lacks the connection to the human body. Here, the move from reference to sense is more difficult and cannot be made solely on the basis of the temperature terms’ extension across an etic grid of physical temperatures.

This problem in semantic typology is sometimes discussed in relation to the philosophical problem known as the “gavagai” problem after the thought experiment of the philosopher Quine (1964, see e.g. Koptjevskaja-Tamm 2008, also Hylton & Kemp 2022: section 6 for a discussion of the problem from the philosophical point of view). Imagine that you find yourself with a group of people whose language you do not know at all. One of them points to a rabbit running past and says “gavagai”. How can you know that what the man is referring to is the rabbit, and not, say, the notion of rabbitness being instantiated, or a collection of un-detached rabbit parts? Quine’s problem is closely connected to the possibility of the indeterminacy of translation (Hylton & Kemp 2022) which is discussed further in section 3.4.
To counter this gap between reference and “true” meaning, some researchers have preferred other ways of approaching the problem of cross-linguistic comparison of meaning than looking at denotation. The perhaps leading approach of this kind to meaning cross-linguistically is the Natural Semantic Metalanguage (commonly abbreviated NSM), first proposed by Wierzbicka (1972) and since then adopted by a large number of scholars. The fundamental assumption of NSM is that all languages are decomposable into the same set of a finite number of primitive and universal concepts, so-called primes—this idea goes back to the philosophy of Leibniz and his “alphabet of human thought”. These primes are immediately understandable and cannot be explained in other words without circularity. Examples are KNOW, BODY and GOOD. The meaning of more complex words can be expressed using primes and a set of connectives, through reductive paraphrase. For explanations of the NSM methodology see e.g. Goddard & Wierzbicka (2013), Wierzbicka (2016), and Goddard (2021).

NSM analysis has been applied to a large range of languages from different families and areas with several highly interesting results, for instance in the domain of emotions (see e.g. Harkins & Wierzbicka 2001) and pain (Goddard & Ye 2014 and papers therein). An obvious drawback of the NSM approach from a methodological point of view is that it more or less requires near-native speaker knowledge or close co-operation with native speakers for explication of more complex meanings. This strongly limits the possible language sample, both in terms of size and what languages can be included. There have also been theoretical objections to NSM, questioning the validity of some of its tenets. For example, Barker (2004) argues that names and natural kinds are not paraphrasable (based on the argument of Kripke and Putnam) which would make NSM an incomplete theory of meaning, and Koptjevskaja-Tamm & Ahlgren (2004) bring up the issue that the primitiveness of the primes identified by NSM should be demonstrable also by empirical evidence and not just theory-internal reasoning.

Another important approach within semantic typology is the frame-based methodology associated with the Moscow Lexical Typological Group, which builds on the Moscow school of semantics (e.g. Apresjan 2000 in English) and the frame semantics of Fillmore (e.g. 1982). The key insight of the Moscow semanticists is that even words with very closely related meanings differ in their combinatorial possibilities, e.g. for verbs, what kind of subjects they can have. Systematically describing these collocations for individual lexemes and semantic fields gives insight into the organisation of the lexicon (Apresjan 2000). The Moscow Lexical Typological Group applies this cross-linguistically through the use of semantic frames (cf. section 3.1), prototypical situations which characterise a domain and are distinguished, among other things, by the prototypical arguments involved in them (for an overview,
see [Rakhilina & Reznikova 2016]). For example, [Lander, Maisak & Rakhilina (2012)] give the following four relevant frames for the domain of “aquamotion” (swimming/ floating): 1) active swimming, 2) passively drifting with the current, 3) floating on the surface, 4) travelling (of vessels and people on vessels). The identification of what frames make up a semantic field is considered an empirical question, and the analysis of this typically starts with the researcher’s native language(s) and is then expanded to other languages and gradually refined. The frames are then used as points of comparison cross-linguistically. In terms of data, the Moscow Lexical Typology Group relies on dictionaries, corpora, questionnaires and verification by native speakers.

The approach to comparing meaning across languages taken in this thesis is intimately connected to the method of data collection. Recall that the data take the form of parallel texts, i.e. translations into several languages of the same passage of text. Meaning is viewed as the extension of forms across these contexts, following [Wälchli & Cysouw (2012)]. Instead of a denotational grid consisting of non-linguistic stimuli, we have a denotational grid consisting of what [Wälchli & Cysouw (2012)] refer to as contextually embedded situations. That is, we compare the meaning of forms across languages by comparing what contexts they are applied to. In this, the method takes an denotational approach to meaning, in contrast with, for example, NSM. Using parallel texts has clear parallels with both stimuli methods and the frame approach, with the main difference being what the ground of comparison is, or, in other words, what the analytical primitives are—physical stimuli, frames or contextually embedded situations in the form of parallel texts. The two latter approaches share the characteristic of looking at how forms apply to situations or events, but differ in that the situations of the frame method are prototypical abstract entities whereas the situations of parallel texts are actual instantiated situations—the use of parallel texts is an example of applied exemplar semantics (cf. [Croft 2007], [Wälchli & Cysouw 2012] and section 3.1). The parallel text approach has of course the disadvantage that nothing guarantees that the situations expressed in for instance the New Testament are representative of or comprehensive in relation to the domain under study. On the other hand, it has the advantage of not necessitating a prior model of the domain and opening up the possibility of discovering that distinctions the researcher is unaware of are relevant to linguistic expression.

The third methodological question concerns how to describe our findings. Supposing we have found proper materials for our investigation and have also found a way to identify the forms encoding certain meanings; how do we describe these meanings? In other words, what can we use as a metalanguage?

As implied by its name, Natural Semantic Metalanguage gives a clear and explicit answer to this. The same units that allow us to compare meanings
cross-linguistically — universal semantic primes — also function as metalinguistic descriptions of meaning. The terms making up the metalanguage are therefore terms from a natural language (in practice usually English) but since they represent semantic primes they are, according to the theory, immediately translatable into any other language without loss or change in meaning, and are therefore suitable as a metalanguage.

Most approaches are not as explicit about the use of metalanguage. In approaches which are explicitly denotational (such as the Nijmegen method) the real or underlying metalanguage is, in a sense, the etic grid established. However, since descriptive reference to the etic grid is often required for readability the language of the publication in practice takes the role of metalanguage. This is addressed explicitly in Wälchli and Cysouw (2012: 682) who write that “[i]nterpretation is not possible without labeling of some sort [—]. However, these labels — or descriptive meanings — are no semantic primes but rather have a similar status to the lexical part of glosses in grammatical analyses in reference grammars. The labels are not precise semantic definitions, but rather indicate that there is a certain range of meaning that is recurrently expressed by a number of categories in different languages and is therefore likely to represent a cross-linguistic semantic category type.” A similar approach of labelling part of semantic space is taken in Haspelmath’s use of comparative concepts—this approach is discussed further in section 3.2.

Even though their use has been limited in typological investigations, diagrams and schemas should be mentioned here. The use of these visual abstract representations of meaning is mainly associated with cognitive linguistics. Not everyone agrees, however, that depictions of this kind are explanatory in the sense that they are more basic or more immediately understandable than the linguistic meanings themselves; Goddard (2018: 311–312) argues that without verbal descriptions accompanying them, the image schemas of cognitive linguistics are uninterpretable.

Another visual approach to representing meaning that has been widely used is semantic maps. Note that the level of visualisation is quite different between the diagrams of cognitive linguistics and semantic maps; in the former, the goal is to visually portray a single meaning, whereas in the latter, it is the relationships between discrete meanings that are illustrated. The meanings included in a traditional semantic map are generally expressed with forms from the language of the publications (generally English). Semantic maps are discussed in more detail in chapter 7.

In this thesis, I follow Wälchli & Cysouw (2012) in using labels from natural languages (generally English for convenience) in lieu of any specific metalanguage. Importantly, however, these forms should be understood specifically as labels for a region in the semantic space in which the Bible verses are lo-
To summarise, this investigation uses translations of the New Testament to investigate the linguistic expression of the knowledge domain. The contextually embedded situations expressed in the texts are taken as analytical primitives, and the meaning of forms is compared by comparing the distribution across these situations.

### 3.4 Using the New Testament parallel corpus

The materials used for this investigation, then, are translations of the New Testament. This brings with it some methodological issues, some that apply more generally to parallel text typology and some that are more specific to New Testament translations. The discussion in this section draws heavily on de Vries’ discussion of the problems of using Bible translations for typology in a special issue of *Sprach und Universalienforschung* devoted to parallel corpora (de Vries 2007) and an article by Levshina dealing with the opportunities and problems of using corpora in typology more generally (Levshina 2022).

The use of a parallel corpus depends on the notion of *translational equivalence*. This is close to the common-sense view of translation being the operation of putting the same meaning into a different language, but was introduced in early translation studies by Catford as a result of his view that meaning is language-specific. Therefore, we cannot speak of meaning being the same across languages. What we can speak of is “the ability to function in the same situations” (Malmkjær 2011a: 110). What gives this ability is, in the terms of Catford, *relatability to the same features of substance* (Catford 1965).

As an illustration of what Catford means by this, consider colour. In English, there are (among others) the colour words *red*, *orange*, *yellow*, *green*, *blue* and *purple*. In Navajo, on the other hand, in Catford’s account, we have *ìicii?*, *ìico* and *dootl’iż*. The terms’ relationship to the colour spectrum is shown in table 3.1 after Catford (1965: 51).

<table>
<thead>
<tr>
<th>English</th>
<th>red</th>
<th>orange</th>
<th>yellow</th>
<th>green</th>
<th>blue</th>
<th>purple</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colour spectrum</td>
<td>x</td>
<td>y</td>
<td>z</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Navajo</td>
<td><em>ìicii?</em></td>
<td><em>ìico</em></td>
<td><em>dootl’iż</em></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 3.1:** Colour terms in English and Navajo

The reason that *ìicii?* and *dootl’iż* can function as translation equivalents of *red* and both *green* and *blue* respectively is that “they are relatable to the same substantial features x, y and z. It is solely this relationship to the same substantial features that justifies their use as translation equivalents since they clearly...
have quite different meanings from the English items” (Catford 1965: 51).

The colour spectrum is an example of what Catford calls “situational substance” (Catford 1965: 51). For many cases of translation, this is the most important feature, but depending on the type, scope and purpose of translations, other kinds of substance may also be relevant, e.g. phonetics in poetic or lyrical texts. Later works in translation studies have both challenged and developed the notion of equivalence, but it remains, in the words of Leal (2012: 44) “very much present, predominantly as a blanket, useful concept”.

In short, then, we consider the New Testament verses across different languages to be translation equivalents because they are usable in the same situation (the situation here being their place in the Bible text) by virtue of sharing the most relevant features. Again, depending on the purposes of the translation what is considered ‘most relevant’ may differ. This brings us to the first issue of using translations as data – the problem of what might be called different styles of translation.

3.4.1 Different styles of translation

In his overview of potential problems with using Bible translations as data for typological studies, de Vries (2007) argues that Bible translations are particularly sensitive to considerations of *skopos*. The theory of *skopos*, from the Greek for ‘purpose’, was introduced by the translation scholar Vermeer (1989/2021) in the late 1970s and takes as its starting point *translation as action*. Actions, in Vermeer’s view, always have a purpose, and consequently, translations are always done with a purpose. Applying this to equivalence, what the relevant features to be considered are (that is, what counts as being usable in the same situation) depends on the *skopos* of a particular translation.

Returning to the colour example above, for a translation with a *skopos* centred on conveying information and factual content, the relations to the physical colour spectrum are relevant, and we end up with the translation equivalents discussed above. If, on the other hand, the *skopos* is centred on conveying emotion — as perhaps in the translation of poetry — the relations of words to feeling and values are more important than their relation to the physical colour spectrum and the translator may end up choosing different forms in the target language. The crucial point of *skopos* theory is that both can be valid translation equivalents; it depends on the translator’s purpose.

In Bible translations, there can be a wide variety of *skopoi*, some grounded in theological differences. For some communities the translation is expected to function in ritual and sacred contexts and thus the ‘otherness’ of the text may be emphasised (cf. the unwillingness of some Swedish congregations to adopt the 2000 translation instead of the 1917 in liturgy, Rydbeck 2011). For
others, translations are a missionary tool and the main purpose is for the text to function independently in the cultures of prospective converts (de Vries 2007).

All of these different skopoi mean that different features may be relevant in determining an appropriate translational equivalent. In the case of someone who believes that the Bible is literally the word of God, word-for-word correspondence is the most relevant, placed above considerations like idiomaticity in the target language. An example is the revision of a translation of the Old Testament from Hebrew into Greek by Aquilas in around 125 CE where each occurrence of the particle marking the accusative case in Hebrew is translated into συν ‘with’, despite this not being more linguistically appropriate than doing the corresponding thing in English (de Vries 2007: 154). On the other end of the scale, we find very free translations. Here the purpose is to bring the message of the Bible to a group of people and the relevant features are therefore the content and, perhaps most of all, the appropriateness to the target culture. Word-for-word correspondence or even sentence-for-sentence correspondence is less relevant.

Free translations are often influenced by Nida’s theory of dynamic or functional equivalence (highly influential within Bible translation) as opposed to formal equivalence (Nida 1964). Translations focusing on formal equivalence, such as the extreme example of Aquilas above, were criticised by Nida who suggested instead a method of dynamic (now often called functional) translation. In brief, Nida’s method consists of extracting the so-called kernel sentences of the source text, basically a kind of component analysis, consisting of expressing the content in basic, more fundamental sentences. This underlying message is then put into a form suitable for the target language (Porter 2008, Barnes 2011).

A similar distinction is that between domesticating and foreignising translations (Venuti 2008, Paloposki 2011). Domesticating translations adapt the content of the source text from the source culture to the target culture, for instance by adapting settings and names, whereas foreignising translations retain the cultural features of the originals.

To summarise then, the nature of translational equivalence — or what is considered able to function in the same situation, in other words: what counts as a translation of a verse — depends on the skopos or purpose of the translation. If the purpose is to bring the Bible to a new group of speakers and the understandability of the message is at the centre, the translator may strive for dynamic equivalence and domestication — a free translation. If, on the other hand, the purpose is to preserve the perceived holiness of the source text for use in liturgy, the translation may strive for formal equivalence and foreignisation — a literal translation.

What, then, does this variety of translation styles mean for the use of Bible
translations in this investigation? The translations in the corpus used range from quite literal to very free in their styles. Most fall somewhere in between the two extremes and mostly correspond with the source text at sentence level; the translators appear to have considered the information content to be the most relevant feature. These translations present no particular problems. In the fairly rare cases where there is more than one translation available, I have, when such information can be found, chosen those where the goal is idiomaticity in the target language rather than faithfulness to the original. For example, the Thai New Contemporary Version was used rather than the Thai Standard Version since it takes a less literal approach and “flows a bit better” (Dahlfred 2012).

Regarding the extremes of translation style, a number of quite free translations are present in the sample. These are problematic because it is not always clear what, if any, stretch of text corresponds to the original knowledge ascription of interest. In most cases, however, it is possible to find a correspondence. As an example, consider the passage from the Fasu translation in 25 in comparison with how the verse reads in English (here the Lexham version, but most are very similar).

(25) Fasu (Isolate, PNG)

a. Mark 15:45

Ai-rakano satini-mo some-raka, “Ao, kakaro kusua-po.”
speak-and soldier-ERG say-NEUTR yes truly dead-DECL

Ai-rakano kai-kia Pairate-mo Yosefe-aki some-raka,
speak-and hear-SEQ PR.NAME-ERG PR.NAME-to say-NEUTR

“Yasu kusu-ane nómo maka pu-ane siahoasirea-po,”
PR.NAME dead-DS 2PL.ERG take-SEQ go-DS be.good-DECL

ai-sa-po.
speak-PST-DECL

Lit.: ‘And (it being spoken) the soldier said, “Yes, he is truly dead.”
And (it being spoken) hearing [this] Pilate said to Joseph “Jesus being dead, it is good if you take him,” he spoke’ (AS)

‘And when he learned of it from the centurion, he granted the corpse to Joseph’

Two things happen in this verse. Pilate hears from the soldier that Jesus is indeed dead (he has previously asked him this) and then gives Joseph the body of Jesus (which Joseph has requested earlier). The Fasu version reports these two events, but in a form which is quite different from the English or the Greek.
Pilate’s learning that Jesus is dead is represented both as direct speech from the soldier and as Pilate ‘hearing’. Also the granting of Jesus’ body is expressed as direct speech. In this, the Fasu version takes some freedoms with the source text but nonetheless conveys roughly the same events—the kernels, in Nida’s view.

Despite the relative freedom of form in relation to the source, it is in this case possible to identify a counterpart to the knowledge predication—Pilate’s hearing of the soldier’s words (bolded in 25). This construction fulfils the requirements on being predication—assigning participant roles in an event—and it is knowledge predication by virtue of appearing in a relevant context (as mentioned, knowledge in this investigation is defined extensionally by a set of verses).

The other extreme – very literal translations – are harder to spot in unfamiliar languages, especially if there are no original texts of the same type available to compare with. If there are other sources saying that a translation is very literal, I have tried to avoid that translation. An example is the translation into Northern Saami, which is very closely based on Finnish (Richard Kowalik p.c.). Another candidate is Car Nicobarese which very closely follows the English—so closely that one may suspect that some idiomaticity has been lost, though this is hard to establish without non-translated texts to compare with. An example is give in 26. Note that the Car Nicobarese translation is one of the oldest in the sample—the original version is from the 1920s.

(26)  Car Nicobarese (Austroasiatic, India)

a. Mark 12:14

\[ \text{Mahaköphöti; akahalön in tö Meh tö-köen} \]
\[ \text{teacher know 1PL COMP 2SG ADJ-be.truthful} \]

‘Teacher, we know that you are truthful’

In this case, the problem is not identifying the appropriate counterpart, but rather the risk of interference from the source language; that the translation does not well represent the target language, but rather the source language. This, and related problems, is discussed in the following section.

3.4.2 Translational interference and unnaturalness of register

The problem of influence from the source language — leading to varieties sometimes called translationese — is brought up by Levshina (2022) as a gen-

---

1Which we, as is usually the case, do not know what exactly it is, though it is likely some English or Tok Pisin version
eral problem of using translations in typology. Scholars in translation studies also emphasise that the process of translation itself gives rise to particular features of translationese; these are argued to be universal and not dependent on the features of either the source or target language. These universal features of translation, summarised by Baker (1993), include a tendency toward greater explication, a tendency toward disambiguation and simplification, a preference for “conventional grammaticality” and a tendency to avoid repetition. Baker (1993: 243) writes that based on “small-scale studies and casual observation, a number of scholars have noted features which seem, intuitively, to be linked to the nature of the translation process itself rather than to the confrontation of specific linguistic systems”, and calls for the systematic investigation of these using large corpora. This call has been taken up, and much evidence has been brought forward for these features of translation—for instance Kenny (2001) on normalisation or Laviosa (1996) on simplification—though their universality and the exact list remain debated (Malmkjær 2011b).

Another problem of using translation corpora (and indeed, almost any corpora) pointed out by Levshina (2022: 20) is the underrepresentation of some registers and text types, and she questions “whether the conclusions made for the available text types will hold if we use data representing other registers”. As a hint that they may not, she cites an example from Östling (2015: 209), who finds that in New Testament translations, the word order in Romance languages is predominantly adjective-noun, whereas these languages are usually classified as noun-adjective in sources such as the World Atlas of Language Structure (Dryer 2013) (in part due to the dominance of basic high-frequency adjectives such as ‘good’, which often have adjective-noun order in Romance).

For the New Testament corpus in particular, we can make a further subdivision of this underrepresentation into texts representing only written language and furthermore representing only some very particular registers (for instance narratives or persuasive discourse).

As Levshina (2022: 20) points out, there is in general a great lack of spoken corpora and even more so parallel spoken corpora—indeed, it could be argued that parallel truly spoken corpora are impossible. The New Testament is clearly an example of written language, even if large parts are dialogue and thus to some extent imitations of speech. Biber (2014) argues that the distinction between what he calls oral (i.e. spoken) and literate (i.e. written) registers is fundamental and universal. Using multi-dimensional analysis of a range of linguistic features, he shows that the oral-literate dimension is highly important in explaining the distribution of features in several languages from different language families (see references in Biber 2014). This divide, however, is not binary, but consists of a continuum from the most oral-like registers (conversation) to the most literate-like (written information prose). Written fic-
tion and letters fall somewhere in between, probably in part because, as Biber (2014: 3) notes, “writers can choose to employ linguistic features associated with stereotypical speech”. The text of the Gospels (from where the majority of the verses in the present study are taken) is clearly closer to the narrative registers of fiction than written information prose. Thus, extrapolating from the results discussed in Biber, we should not discard the New Testament corpus simply for being written; being narratives, they are likely to share many features of spoken language. This, of course, is not to deny that NT translations are written language, with the limits on representativity that come with that.

The other part of the problem pointed out by Levshina (2022) is that corpora often represent only a particular register. In the case of the New Testament, “[m]issionaries are notorious for creating new varieties of the languages into which they translate the Bible or parts thereof” according to Stolz (2007: 103). Wälchli (2007: 131) (citing Briggs 1993) discusses the case of so called aymara misionero, a variant of Aymara (Aymaran, Bolivian) used by missionaries in which the use of evidentials departs from the standard variety. Like the issue of too literal translations, this danger of an “artificial register” is difficult to safeguard against. Without original texts as a ground for comparison or the judgements of others familiar with the language in question, we must take the texts at face value.

As hinted at by the quote from Levshina (2022: 20) given above — “[will] the conclusions made for the available text types [...] hold if we use data representing other registers” — the key issue here is representativity. In typology, we are generally not interested in a text in itself but as a representative of a language variety. We do not really want to say that the linguistic feature so-and-so is expressed so-and-so in the translation of the New Testament made by so-and-so in this or that year; rather we want to say that since we can observe the linguistic feature is expressed this way in this text, we can conclude that it is expressed in this way in language X. I have brought up three related challenges to allowing us to perform this jump of reasoning when working with New Testament translations—translationese, spoken vs. oral registers and the potential particularity of New Testament registers.

As an explicit acknowledgement of these issues, some authors prefer to speak of doculects rather than language and varieties (e.g. Wälchli & Cysouw 2012). A doculect refers to a language variety as instantiated in a particular text, bypassing the difficult question of what constitutes one language or variety. I will not use doculect here, but the results presented should of course be understood as applying empirically only to the specific texts investigated—extrapolation to languages as a whole is tentative and likely incomplete.

In discussing the two problems raised so far — different translation styles
and translationese — we have been concerned with the process of translation and regarded the source text as unproblematic and unchanging. In the case of Bible translation, as de Vries (2007) points out, this is not always the case. This issue is discussed in the following section.

3.4.3 Multiplicity of sources

Unlike, say, the Declaration of Human Rights or Harry Potter, “the Bible” cannot be identified with a unique source text. What we refer to as the Bible is a collection of texts which are available to us through a collection of manuscripts of varying age, geographical origin, degree of completeness and, in places, content. Their merging into a coherent book is the result of long traditions of text-critical work. This is the source of what de Vries (2007) points to as one of the challenges particular to using specifically Bible translations in linguistic work; it cannot uncritically be assumed that the same passage in two different translations are translations of the same source text. As an example, de Vries (2007: 152–153) gives the ending of the Lord’s Prayer in Matthew 6:13 in the King James Version: “for thine is the power and the glory for ever. Amen.” This passage is missing in most modern versions.

To control for this danger, the verses in the sample are checked against text-critically annotated versions of the Bible. Relevant points are given in Appendix B.

Another issue here is that translations are often made from languages other than the original Greek (for example, English or Spanish). This problem is exacerbated by the fact that in general it is difficult to know from what source text a translation was made, since this is usually not indicated. As an example, consider the two versions of Mark 14:68 below, in English (King James Version) and in Spanish (Reina Valera Contemporánea 2011, the latest revision of the most authoritative translation into Spanish).

(27) English (Indo-European>Germanic, England)

a. Mark 14:68 (King James)

But he denied, saying, I know not, neither understand I what thou sayest. And he went out into the porch; and the cock crew.
a. Mark 14:68 (Reina Valera Contemporánea)

Pero Pedro lo negó, y dijo: «No lo conozco, ni sé de qué hablas.»

Lit.: ‘But Peter denied it and said: “I do not know him, nor do I know of what you are speaking”. And he turned to the entrance. In that moment the rooster crowed’ (AS) ‘But he denied it, saying, “I neither know nor understand what you mean!” And he went out into the gateway, and a rooster crowed.’

The rendering of Peter’s word is clearly different in the two translations: in English, both know and understand have what thou sayest as the object. In the Spanish, the object of the first verb — conozco ‘know(person).1SG’ — is lo, most readily interpreted as ‘him’ (i.e. Jesus) and only the following verb (sé ‘know(fact).1SG’) has the relative clause as its object. This structure is followed in a number of South American language translations, indicating that they are translated from the Spanish. Where it seems that differing source texts may play a role in the results, this is addressed throughout the thesis.

However, the problem of source multiplicity goes further than possible multiplicity of source texts, and in this extended sense it is not unique to Bible translations. Even supposing that we know that the text passage which is the basis for two translations is the same in both cases, how can we be certain that the translators understood the source text in the same way? de Vries (2007: 149) discusses the problem as underdetermination: for any given source text, multiple interpretations and translations are possible. We cannot be certain that different translators have the same situation in mind when translating a stretch of text. Fortunately, such certainty is not necessary: what is necessary is that we can safely assume that the situations are similar enough for the purposes of comparison.

This is possible by taking advantage of the situations translated being contextually embedded. All the verses used are part of a larger textual context, be it a narrative or a letter. This means that the surrounding text constrains the possible interpretations of the verses. Consider for example the following two
From the discussion in section 2.2.2 we expect that different degrees of personal acquaintance might be differently encoded; these two verses might seem a good starting point for investigating this. However, in the King James English version, as in the Greek, we find the same lexical item expressing the relation. How can we be sure that translators into different languages are imagining, and thus putting into words, the same situations? Through the constraints of the surrounding textual context, I argue. Consider the first example. A causal explanation is given for the knowledge already within the same verse, making it clear that it is knowledge through direct contact. In the second case, we need to look outside the immediate verses for constraining context. The verse is from a passage in which Stephen recounts the history of the Jewish people and is the segue between the story of Jacob and Joseph and the time of Moses, years in which the Jewish people “grew and multiplied” (Acts 7:17). Clearly, then, the knowledge in this verse cannot be direct personal acquaintance, but rather knowing of or about.

Thus, in most cases, the possible interpretations are constrained by the surrounding context. We cannot expect the underdetermination to be solved entirely, but sufficiently so for practical purposes.

Finally, we come to a bundle of difficulties that holds for all typological corpus work, not just parallel or Bible corpora, namely the fact we are dealing with generally un-glossed and otherwise unannotated pieces of text in unfamiliar languages. This is discussed in the following section.

3.4.4 Noisiness, low efficiency and difficulty of analysis

Corpus data are generally “noisy”, in the words of Levshina (2022: 22). We may find spelling mistakes, punctuation errors or in the case of computationally processed texts errors resulting from that. But even apart from that, texts are “noisy” compared to the kind of information given in grammars or dictionaries; what we have is simply a string of characters—no paradigms or explanations. This means that using corpora is “extremely time-consuming” (Levshina 2022: 21). But, perhaps more problematically, using and analysing
un-glossed corpora is also quite difficult and therefore susceptible to error, as discussed in [Wälchli (2007), 128].

A question that presents itself here is, if there are so many problems with using (New Testament) parallel corpora, why should we do it? The answer is that the advantages outweigh any problems. [Levshina (2022)] gives a detailed overview of the possibilities of using corpora in typology and the new avenues of investigation it opens up, and [Wälchli (2007)] discusses similar points specifically with regard to parallel corpora. I will not repeat all of their points here; instead, I will simply present the argument for why using New Testament translations is a suitable method for the research questions I ask.

In short, the main questions this thesis seeks to answer are: how do the languages of the world express knowledge predication and what distinctions are made within the domain? As mentioned in section 3.3, grammars and dictionaries rarely contain the kind of information needed to answer these kinds of questions. We are left with elicitation or texts, original or translations of some kind. A number of considerations lead to the use of New Testament translations, generally relating to the exploratory nature of the investigation. Very little is known about the organisation of the knowledge domain cross-linguistically. It is therefore desirable to include a large variety of examples as well as a large number of geographically and genetically diverse languages. The first consideration points toward using parallel corpora; it would be difficult to find comparable contexts across languages in original texts, especially when not knowing what the relevant dimensions are. The latter would also make it difficult to design elicitation material (together with the mental and non-physical nature of the domain). Having decided on parallel corpora, the second consideration points us toward the New Testament. It is, by far, the most widely translated text and allows for worldwide coverage, albeit with gaps.

Some of these gaps may be systematic, and it is important to consider the results in light of. Which languages the New Testament is translated into depends on the choices of individuals, something which may be the result of biases. As a very clear systematic bias reflected in the materials, translations of the NT are almost entirely limited to languages which are easily put into writing; this means that sign languages are excluded.

Nonetheless, using translations of the New Testament allows us to get a general picture of the knowledge domain in a worldwide variety of languages. This is something we do not yet have. The possibility that the picture is flawed or incomplete due to the shortcomings of the materials discussed above must be weighed against this advantage.

And with this basic understanding in place, of course, better placed to continue the investigation in other ways and refine the picture.
3.5 Choosing verses

It has thus been established that the materials used for the investigation of the linguistic expression of knowledge in this thesis are verses from translations of the New Testament. This immediately brings with it two related methodological issues, namely:

a) which translations to include

b) which verses to include

Problem [a] can in general be restated as which translations to include—it is only very rarely that there is more than one translation in a language—and is as such central to most typological investigations and has been discussed extensively under the rubric language sampling. The approach to language sampling taken in this thesis is described in chapter [4]. Issue [b], however, is more specific to the parallel texts method. To be sure, all typological investigations need to specify the domain under investigation, but the problem when using parallel texts is to concretise this domain as passages of text (in the case of the NT verses).

First of all, recall from the Introduction that the domain under investigation is knowledge predication, which I defined as a construction in a language which assigns participant roles in a knowledge event to particular entities. This follows Stassen’s (1997: 12) definition of predication as “the application of a general concept to a particular entity” or, put differently, “a particular entity [—] is assigned the role of participant in a certain state of affairs”. Stassen deals with intransitive predication, that is, one-place predicates. In the domain of knowing, we are generally dealing with two-place predicates, which explains the slight modification of the definition. I take ‘event’ to be a broader term than state of affairs, including also dynamic situations. Thus

Definition 7. **Knowledge predication** is the assignment of participant roles in a knowledge event to particular entities, or a construction expressing such an assignment.

In practice, this means that we are interested in verses which contain a predicate similar to *know*. We cannot, however, simply take all the verses containing *know* as the domain of investigation, for several reasons. First of all, there would be too many—English *know* occurs around 500-600 times in most translations. Second, while *know* roughly captures the concept under investigation, it does not define it. Some other way to choose verses is therefore needed. In this section, I will describe the process of choosing the verses to include in
the current investigation. This process can be divided into two phases. In the first phase, the extension of automatically extracted ‘know’-like forms in all languages across all verses was compared. In the second phase, the results of this and the sketch of the knowledge domain given in chapter 2 were used to construct what might be called a variety verse sample (analogously to a variety language sample, see chapter 4).

The purpose of the first phase was to avoid the influence of English in particular and European languages in general on how the domain of investigation is defined. The literature on knowledge has been very focused on English and sometimes some other European languages but, of course, neither English nor any other language deserves a privileged place in the definition of the domain cross-linguistically. This is not overly problematic in the case of including meanings based on English; the worst that happens is that we find that a particular co-expression is a peculiarity of English and does not occur in other languages. What is problematic, however, is the risk of excluding categories based on a small non-representative sample of languages. In other words: if, based on the linguistic behavior of English and a few other languages, we entirely exclude meaning a from our domain of investigation B, we will never discover if languages include a in B.

To minimise this risk, I used two versions of automatically extracted correspondences (Östling p.c. and Wälchli p.c.)—see Östling & Kurfal (forthcoming) and Wälchli & Sölling (2013) for details. What these automatic ways of extracting forms do is start from a defined set of verses (in this case those containing forms of English know) and then find what form(s) in each language best corresponds to this set in its distribution. Once the forms in each language which most likely mean something like ‘know’ are identified, the next step is to consider the distribution of these forms within each language. Thus, while the method picks the “closest match” to a pre-determined set, so to speak, the match will rarely (if ever) be one hundred percent accurate. What is of particular interest to us here is if we find that some of the forms identified recurrently have a wider distribution than English; in other words, if there are meanings that should be included in the knowledge domain which are not found in English or other well-known European languages. Note that this method is automatic and it is therefore possible to apply it to all the available Bibles in the corpus, ca 1700.

The procedures generally are illustrated in figure 3.2.

---

1 I use co-expression as a broader variant of colexification, encompassing all kinds of forms expressing two or more distinct functions and not just lexical items.

2 The 8-digit numbers refer to verses: the first 2 the number of the book (01 is Genesis and so on), the following 3 the chapter (001 is chapter 1) and the last 3 the verse (001 is the first verse of relevant chapter).
As a matter of fact, the adjustments made to the initial characterisation of the domain based on this comparison were relatively minor. It became clear that meanings similar to understand are often covered by the same word as ‘know’, something which strengthened the decision to include these meanings in the investigation. It was also clear that the English way of co-expressing knowing facts and knowing people, as opposed to how it is done in most other European languages, is not a peculiarity of English, and that the decision to include both the meanings in the sample was reasonable.

Otherwise, the comparison revealed interesting but for this investigation not relevant recurrent patterns of expression including words for ‘know’. An interesting example is that ‘heathen’ rendered in some languages as ‘someone who does not know God’. This is an interesting use of the word for ‘know’, but it does not represent a different sense to be included in the investigation. These apparently formulaic uses of know-words are a promising area for further research.

The possibility that some senses which are frequently recurrent for ‘know’ forms are excluded from the verse sample is thus minimised. As an important caveat, it can be noted that this of course only applies to senses found in the NT which, while quite varied, is far from a comprehensive reflection of meanings in a language.

The choice of verses could thus proceed to the next phase. There are several possible approaches to a verse sample. In some cases, it may be possible to include all relevant verses, if the phenomenon under investigation is relatively uncommon in the NT. This, however, is not the case for knowledge predication, as was seen above. It is therefore necessary to select a subset of

**Figure 3.2:** General illustration of the procedures to extracts forms automatically
all potentially relevant verses. One possibility is to randomly select a subset of all the potentially relevant verses. Another possibility is to consider all the relevant verses from a specific section of the next, as done e.g. in Wälchli & Cysouw (2012).

In the current investigation, however, we already have ideas about what parameters might be important in the expression of knowledge; chapter 2 considered this based on previous research in linguistics, philosophy and psychology. To effectivise the investigation, it makes sense to tailor the verse sample to these parameters. In other words, verses are chosen to represent the (different values of) the parameters identified in chapter 2. It should be noted that this does bring with it a risk that variation which is not predicted by previous work may go unnoticed. To counter this, I attempted to make the verse sample as varied as possible in different ways; varying subjects, dialogue vs. narration, tense etc.

This phase of the verse sampling was carried out mainly based on English translations (Lexham and King James), but also on Swedish (Bibel 2000) and on Greek.

To begin with, a number of verses representing typical knowledge (see section 2.1 for a discussion of definitions of knowledge) are included. These take the form of know with complement clauses, including wh-question complement clauses. The verses were chosen to vary in subject person and tense. This gives the verse categories Know that (7 verses) and Know wh- (3 verses).

Early piloting showed that a not entirely uncommon phenomenon is what can be referred to as lexical negation of knowledge (see section 7.7). That is, distinct lexical items are used to express ‘know’ and ‘not know’. Early work with the language North Tanna showed that the distribution of these forms may also be sensitive to the depth dimension discussed in section 2.2.4, which motivated including variation in this parameter as well. This gives the category Not know (5 verses).

As discussed in section 2.2.2, English is an exception in the European context in subsuming knowing facts and knowing people under the same category. Additionally, there appear to be cross-linguistic differences in how widely a form for knowing people is used—Benton’s theory of three degrees of personal knowledge was applied. To investigate this, verses representing different degrees of personal knowledge are included—this yields Know people (7 verses). In section 2.2.3, I discuss the fact that words for knowing people are often also extendable to other nominal objects, and some underlying semantic parameters in regard to this were discussed. Unfortunately, the New Testament is not overflowing with contexts which are relevant to the investigation here. Nonetheless, the category Know other nominal (6 verses) is included. The category is somewhat biased toward what might be called proposition-like
Sections 2.2.4 and 2.2.5 discuss the parameter of depth of knowledge and the semantics of associated lexical items, i.e. forms like *understand*. As stated, the first phase of the verse sampling showed that these are meanings often subsumed under ‘know’. The preliminary sketch showed that there are several components associated with forms like *understand* and it would not be feasible in an investigation of this exploratory and large-scale nature to try to disentangle them all. Since the discussion of the depth dimension has been especially prominent, the choice of verses is focused on this, in contrast to a group of verses chosen to represent primarily the achievement component of ‘understand’. This yields the two categories *Understand—depth* (5 verses) and *Understand—achievement* (4).

The first automatic phase of considering the extension of ‘know’ forms also made it clear that English is by no means alone in including the achievement sense of ‘know’ in the meaning of the form. A number of distinct categories were identified here, in part based on the forms in Greek. This resulted in the addition of four categories—*Recognise* (5 verses), *Realise* (3), *Hear—clear* (3) and *Hear—unclear* (4). The distinction between the two last categories lies in whether the event is immediately tied to a specific utterance or not—i.e. how clearly the source of knowledge is specified as being through hearing.

Finally, as was seen in section 2.2.1, some languages distinguish a ‘know how’-meaning. I suggested that the relationship between this and ‘know that’ and ‘be able’ is best seen as a continuum. To investigate this, a group of verses representing the ‘know how’-meaning was selected—*Know how* (5 verses). Also selected were verses representing various stages along the continuum from cognitive to non-cognitive abilities—*Ability* (8). Additionally, verses expressing external ability, i.e. situations where the condition allowing or not allowing the subject to perform an action are external to the subject (‘He cannot play, because he does not have his violin with him’) (cf. van der Auwera & Plungian 1998) were added—*External ability* (3). The reason for the larger number of verses from this domain was that it made it possible to investigate whether the extension of ‘know’ forms into ability is variable.

In addition to investigating the internal structure of knowledge, this thesis aims to delimit knowledge against other concepts. This requires contexts representing concept close to knowledge. The concepts represented were chosen based on intuitions about conceptual similarity as well as previous research on what concepts knowledge is connected with (see also section 10.2). Additionally, the concept, of course, had to be represented in the New Testament. This last requirement unfortunately disqualified ‘feel’ from consideration—there are too few occurrences.

The meanings added here were *See* (5 verses), *Hear* (4), *Obey* (4), *Believe*
Think (3) and Remember (8). The relatively high number of ‘remember’ verses was to represent the internal variability of this concept as well as due to the very close conceptual similarity of knowing and remembering—a possible extension of ‘know’ forms seems particularly likely here. In addition, three verses representing the middle ground between ‘know’ and ‘see’ were added—verses where sight is probably not to be understood as literally referring to the sense organs, but involves at least some intellect—Know/see (3 verses). This was motivated by the observed strong and supposedly universal connections between knowing and seeing (see section 10.2). In total, this resulted in 100 verses to be used in the investigation. The majority are from the Gospels, with some from the Acts and a smaller number from later letter books. The full sample of verses is given in Appendix B.

3.6 Constructions

We have now defined a domain for the investigation, consisting of exemplars in the form of translations of New Testament verses. The goal of this thesis is to investigate the linguistic expression of this domain. A central issue here is clearly what is meant by ‘linguistic expression’ — what is it we are interested in investigating on the linguistic side? This is the topic of this section; 3.6.1 introduces constructional typology and section 3.6.2 describes the coding schema used in this thesis.

3.6.1 Constructional typology

Traditionally, semantic typology has been concerned with how a given domain is carved up lexically, i.e. what lexemes are used for what subparts of the domain. A typical example is much of the work on colour terms. The further linguistic behaviour of the lexemes is not considered.

Recent works have argued that this approach is inadequate for fully describing the linguistic expression of at least some domains. Koch (2012a) argues that in describing the expression of existence-possession-location, looking simply at the verbal lexeme is not enough. He raises three problems with limiting oneself to a purely lexical typology:

Problem A: Verbal lexical items do not always correspond to simple morphemes. We also have to cope with internally complex lexical items, such as idiomatic expressions.

Problem B: Although the choice of the verbal lexical item is central for the expression of these concepts, the valency of any particular verb involved is important as well.
Problem C: The “expression” of these concepts can vary between full lexical verbal items, explicit copulae, and so-called “zero copulae”.

(Koch 2012a: 546)

Similarly, Koptjevskaja-Tamm (2022) shows that only looking at lexemes and discounting grammatical properties cannot fully capture the cross-linguistic variety in the expression of the temperature domain. Similar to Koch’s problem A, there are complex patterns of full and partial colexification within the temperature domain. As an example, consider the Indo-Aryan language Palula in which tactile temperature (e.g. ‘the stone is hot’) is expressed with an adjective and personal feeling (‘I am hot’) is expressed with a noun. In the case of ‘hot’ these are unrelated (táatu/téeti vs. húluk) but in the case of ‘cold’ they are related (šidáalu vs. šid) (Liljegren & Haider 2015, Koptjevskaja-Tamm 2022: 6).

Similar to problem B, but more generally, there are cases where functions are not lexically distinguished, but are distinguished by the construction used. As an example, consider again Palula, in which both personal feeling and ambient temperature (‘it’s hot’) are expressed with the same noun, but in the first case a construction with the experiencer as the goal is used, and in the latter the construction is existence or coming-into-being:

(30) Palula (Indo-European>Indo-Aryan, Pakistan)

Koptjevskaja-Tamm (2022: 4) from Liljegren & Haider (2015)

a. Personal feeling

\[
\begin{align*}
\text{asaám} & \quad \text{húluk} & \quad \text{dit-u} & \quad \text{de} \\
\text{1PL.ACC} & \quad \text{heat[M]} & \quad \text{fall.PFV-M.SG} & \quad \text{be:PST}
\end{align*}
\]

‘We were feeling hot.’ (lit. ‘Heat had fallen on us.’)

b. Ambient temperature

\[
\begin{align*}
\text{páar} & \quad \text{biid-u} & \quad \text{húluk} & \quad \text{de} \\
\text{last.year} & \quad \text{much-NOM.M.SG} & \quad \text{heat[M]} & \quad \text{be:PST}
\end{align*}
\]

‘It was very hot last year.’ (lit. ‘Last year there was much heat.’)

Finally, corresponding to Koch’s problem C, there is the fact that different lexical items may have different properties or class memberships. As already mentioned, Palula also illustrates this; in some functions the temperature word is a noun, in others an adjective (Koptjevskaja-Tamm 2022).
Koptjevskaja-Tamm (2022: 26) summarises this great variety as analysable from three perspectives, corresponding to the problems mentioned above: formal relatedness of lexical items, the constructions in which the lexical items occur and finally the morphosyntactic properties of the items. A narrow focus on lexical items does not include these perspectives in the description, and therefore cannot fully capture the expression of a domain.

The limits of a narrow lexically focused typology described by Koch (2012a) and Koptjevskaja-Tamm (2022) are not specific to the domains they investigate; the same kinds of problems are found in the knowledge domain.

Beginning with relations between lexemes, the knowledge domain is rich with lexical items with some degree of formal relatedness to each other. We have complex lexical items, such as derivations or particle verbs (example 31). There are also cases like serial verb expression, light verb expression or metaphorical idioms like that in example 32.

(31) Swedish (Indo-European, Sweden)

\textit{känna} ‘know personally’ – \textit{känna till} ‘know of, about’


\textit{niki-POSS} t-\textit{oht} \\
inside-POSS 3SG-hit

‘remember, recall’

In describing these linguistic forms, we want to be able to capture both the similarities and differences between them, the possible similarities with forms in other languages as well as regularities in for example the direction of relationships.

As for the grammatical constructions in which the lexical items occur, we find a number of recurring ways in which these need to be considered to fully account for the encoding of the domain. We find the well-known issue of experiencers being grammatically object-like rather than subject-like (often known as “dative subjects”, see e.g. Bayer 2004)—example 33. The verb \textit{ile} ‘know’ in Yuracaré is one of a small set of verbs in which the experiencer is realised as the object. In the example below this can be seen in the \textit{ma}-prefix, third person plural object, that is, the ones knowing (see van Gijn 2006). There are examples where where the case of an argument expresses meaning distinction (example 34) and in the case of clausal complements, we may find a difference in complement marker or verbal morphology (example 35); in Basque,
the complementiser -ena is factive, presupposing the truth of the complement, whereas -ela has no such interpretation (Artiagoitia et al. 2003: 646). Finally, we sometimes need to consider TAM-constructions to account for differences in meaning which are elsewhere lexicalised (example 36).

(33) Yuracaré (Isolate, Bolivia)
   a. Luk 8:53
      
      mammui ca-l-aimimë-u Jesús
totality 3SG.O-AFF.O-laugh-3PL.SBJ PR.NAME
ma-ile-ti-la naa tuhui-ti-la
3PL.O-know-DS-INS DEM dead-DS-INS

‘They all laughed at Jesus because they knew that she was dead’ (AS)

(34) Hungarian (Uralic, Hungary)
   a. tud- NOUN-ACC ‘know (e.g. the way, answer)’
      John 14:4

      oda tud-játok az ut-at
there know-2PL.DEF.OBJ DEF road-ACC

‘[and] you know the way there’ (AS)
   b. tud- NOUN-DEL ‘know about’
      John 4:32

      ami-röl ti nem tud-tok
REL-DEL 2PL NEG know-2PL.INDF.OBJ

‘[I have food to eat] that you do not know about’

(35) Basque (Isolate, Spain)
   a. Artiagoitia et al. (2003: 646)

      Entzun dut Amaiaren neba hil dena (*baina ez
      hear AUX PR.NAME.GEN brother die AUX(ena) (*but not
da egia)
is true)

      ‘I heard that Amaia’s brother died (*but it’s not true)’
b. Artiagoitia et al. (2003: 646)

Entzun *dut* Amaiaren *neba* hil *dela* (baina ez hear AUX PR.NAME.GEN brother die AUX(ela) (but not *da* egia)
is true)

‘I heard that Amaia’s brother died (but it’s not true)’

*sol* ‘know’ – *da-sol* PFV-know ‘come to know, learn’

Also in these cases we want to be able to capture both differences and similarities, both inter- and intra-linguistically.

Finally, we find variation in the morphosyntactic properties of the lexical items expressing ‘knowing’ and related concepts. An example where we find two knowledge-words belonging to different word classes is given in 37. In [37a] knowledge is expressed using a construction in which the noun *asigi* ‘knowledge’ is possessed by the knower, whereas in [37b] it is expressed as a verb, taking for instance the past tense inflection *-ana*.

(37) Daga (Dagan, Papua New Guinea)

a. Mark 12:14

*nu* *asigi-nu-p*

1PL knowledge-1PL.POSS-LOC

‘we know’

b. Matt 25:26

*Ge anu-ana ne nawan uon*

2SG know-PST 1SG my.work NEG

‘you know that I do not work’ (AS)

Clearly, then, just looking at lexeme choice is not enough to get a full picture of the linguistic expression of the knowledge domain. This is the main point of *constructional typology*: typology through the lens of construction grammar (Koch 2012a; Koptjevskaja-Tamm 2022). Construction grammar is a collection of theories of language which have in common that they argue for doing away with the traditional divide between lexicon and grammar. Just as lexemes are traditionally thought of as sign-signified pairings, construction grammar
takes grammatical patterns, such as ditransitive or passive constructions, to be form-meaning pairs. As an illustration, consider below.

(38) **English (Indo-European>Germanic, UK)**


*She smiled herself an upgrade.*

This sentence can clearly be interpreted as depicting a transfer, analogously to e.g. “She bought herself an upgrade”. How is it that this is so? It is, as *Goldberg (2006: 6)* writes, not particularly plausible to ascribe this sense to the verb itself. More reasonably “[i]t is the argument structure constructions that provide the direct link between surface form and general aspects of the interpretation” (*Goldberg 2006: 6*). Just as *smile* means something (“[t]o give to the features or face a look expressive of pleasure or amusement” according to the Oxford English Dictionary “*smile, v.*” (2023)) the construction itself has meaning. Of course, it is a much more general meaning than that of most lexical items — perhaps something like “A transfers X to B”. The crucial point is that this interpretation does not come about due to the meaning of an individual part of the sentence but of their combination in the specific construction.

Construction grammar originated with an attempt to make sense of idioms (*Fillmore, Kay & O’Connor 1988*). It is clear that a phrase such as the English “pull NP’s leg” is not interpretable only from its parts: in this way, it is idiosyncratic and more like a lexicon entry than explainable through grammatical rules. But as Fillmore et al. point out, idioms like this also interact with the normal rules of the grammar (1988: 504). We can, for example, say that someone pulled someone’s leg, that you shouldn’t pull my leg, (somewhat playfully) that legs were pulled and so on. How can this interaction be accounted for? The construction grammar answer is by recognising that the language system at all levels can be described as part of the same network. Idioms and the ditransitive example considered above are not exceptions; language in its entirety can be described as a network of form-meaning pairings of varying complexity and generality. In other words, even such a basic grammatical rule as (in phrase structure terms) S->NP VP — a sentence consists of a noun phrase and verb phrase — is a construction with a meaning of its own (see e.g. *Fillmore, Kay & O’Connor (1988)* or *Croft (2001)* chapter 1).

In short, then, construction grammar takes a language to be a collection of form–meaning pairs. This goes for the units classically considered as such — morphemes and lexemes — but also for grammatical patterns, such as the ditransitive or passive constructions. This means that there is no sharp distinction
between the lexicon and a grammar of a language — all parts of a language are analysable in the same way, as form-meaning pairings. Of course, this does not mean that no difference is recognised between, for example, lexical items and argument structure constructions. However, the differences are not absolute, but rather form a continuum.

The division of labour between regions in the continuum may vary from language to language and between different domains. Koptjevskaja-Tamm, Divjak & Rakhilina (2010) argue that the domain of aquamotion (i.e. motion through water, e.g. swimming) is more finely distinguished at the lexical level in Swedish, Dutch and German than in Polish and Russian, but that some of these distinctions are instead found at levels of more complex constructions in the latter languages. Since form-meaning pairings are taken to be the fundamentals at all levels in the lexicon-syntax continuum in construction grammar, constructional typology can accommodate these variations and describe both systems.

In short, in investigating the linguistic expression of a domain we cannot assume that all distinctions will be made in the lexicon (rather, as has been shown, there is good reason to assume that this is not the case). Languages will likely vary in the division of labour between lexicon and other means. Construction grammar, which takes not only words to be form-meaning pairings but also grammatical constructions, gives us a way to describe linguistic encoding of domains taking division of labour into account.

Therefore, the linguistic units under investigation in this thesis are constructions.

There is a problem of demarcation here. One of the characteristics of constructionist approaches is that it’s “constructions all the way down” (Goldberg 2006: 26). All expressions are combinations of numerous constructions working together. Goldberg (2006: 10) gives the example expression in (39) and the constructions that it is made up of (40):

(39) What did Liza buy Zach?

(40) a. Liza, buy, Zach, what, do constructions
    b. Ditransitive construction
    c. Question construction
    d. Subject-Auxiliary inversion construction
    e. VP construction
    f. NP construction
It is generally not possible to say that an expression is an example of only one particular construction; each expression is made up of several constructions at various levels of complexity and schematicity. The constructions involved, however, are not simply an unordered mass. Rather, the constructions making up a speaker’s knowledge of a grammar exist in a taxonomic network. Some constructions are instances of other constructions — that is, they are a further specified variant of a more general construction. The construction “S buy IO DO” is an instance of the more schematic ditransitive construction and inherits any properties of it that are not specifically overridden in the inheriting construction (Goldberg 1995: 72–81, Croft 2001: 25).

Another important kind of relation among constructions is the subpart link, defined by Goldberg (1995: 78) as “when one construction is a proper subpart of another construction and exists independently”. “Liza” is a subpart of “What did Liza buy Zach?”, and an instance of the proper name construction (which may be an instance of the noun construction).

Something can be said here about what is meant by “a construction”. The term may be used for a node in the network making up the linguistic knowledge of a speaker. It may also, however, be used for a specific utterance (such as “What did Liza buy Zach?”). The term is thus ambiguous between these two senses, but generally context determines which sense is intended.

3.6.2 Coding for constructions

The task of the investigation is then set. We will identify the construction corresponding to the knowledge predication in each context and will then detail which more general constructions this construction is related to, through instantiation or subpart. Already the list for the relatively uncomplicated expression in example 40 (“What did Liza buy Zach?”) is quite long, and not exhaustive (we lack, for example, word class constructions, tense constructions and so on). Clearly, we would have a problem if we were to attempt to specify each construction involved in each context in the investigation in this way. Crucially, however, it is intuitively clear that apart from the practical infeasibility, not all constructions are equally involved in the expression of all aspects of meaning.

There need to be some criteria for delimiting which constructions are of interest.

It should be reiterated here that the main method of this thesis is the investigation of exemplar data, i.e. individual verses in individual languages. In practice, this means that each verse in each language is coded (or annotated) for what constructions it is an instance of. It is this coding which I am primarily concerned with in the discussion to follow. Apart from this data at the
exemplar level, there is of course also a more general level on which it is noted, for instance, that the word for ‘know’ in a given language is a transitive stative verb or that a nominal object of ‘know’ always takes the accusative. But on an exemplar level, this specification is superfluous.

In deciding what to code for at an exemplar level, I have used two criteria. First of all, the investigation is focused on predicating constructions, by which I understand a construction which assigns participant roles in a state of affairs to particular entities, following Stassen (1997: 12). These predicating constructions potentially involve a large number of constructions, of which the following types are considered directly relevant to the investigation:

i) The lexical morphemes functioning as predicates

ii) Predicate constructions — that is, constructions which are involved in the realisation of the lexical morphemes as predicates. This may include:
   a) Auxiliary and serial verb construction
   b) Derivational morphology and particles
   c) Inflectional morphology

iii) Argument realisation constructions — that is, the way in which participants in the event are realised. This may include:
   a) Case marking
   b) Complementisers
   c) Verbal morphology of the complement

This is of course a very broad list, but excludes some things that are presumably not of immediate interest for the expression of knowledge ascriptions, for example time or place adverbials.

The second criterion is considerably more restrictive, and consists of a focus on differentiation within the domain. That is, the main interest is whether in a language two constructions contrast with each other. Therefore, features that are constant across the domain (such as nominative subjects) are not coded for, although they are noted on a more general plane. There are two reasons for this decision, in addition to the impracticality of coding such things.

The first is that coding for this risks giving rise to dimensions in the data that are irrelevant. Assume we decided to code all tense and aspect distinctions on verbs. This would probably mean that we end up with a strong divide between present and past tense in our analysis, something which is of very limited interest. Furthermore, many common grammatical categories tend to be neutralised in some contexts (for instance, tense and aspect in subordinate
clauses) which leads to the problem of how to analyse these cases. Should a value be inferred or should it be left unspecified? Regardless of the solution, this will have consequences for the analysis of the data. There are thus practical problems beyond time consumption with “over-including”.

The second reason has to do with the view of meaning as paradigmatic. We could note that all objects of the word most closely corresponding to ‘know’ in a language are always marked with a particular case, which a grammar may call Accusative. But without knowing much more about the language (Are the objects of all transitive verbs marked like this? What is included in the class of transitive verbs?) this is of no real interest. Sometimes this kind of information is readily found in grammars, but certainly not always and it lies beyond the scope of this thesis to chart the entire linguistic systems of the languages investigated.

These two criteria give a framework for solving the problem of demarcating which constructions should be coded for, but the decision still has to be made on a case-by-case basis. A problem is that many grammatical categories cut across the entire language, for instance, tense or person. This is in some sense a differentiation within the domain, but, crucially, it is not something which is particular to the domain. That a language can mark knowing as taking place in the past, present or future does not tell us anything specifically about how it divides up the knowledge domain—it tells us something about how it divides up time and events more generally. However, as will be seen, in some cases a more general constructional distinction can be applied to the knowledge domain to express domain-particular distinctions. An example is the Turkish progressive which when combined with bilmek ‘know’ gives a (slight) meaning distinction versus the use of bilmek without progressive (Vafaeian 2018: 133). There is also the case of perfective aspect markers combining with the word for know to mean ‘realise, learn, come to know’ — domain-particular or not (for a more detailed discussion of this particular issue, see section 9.3.1)?

Ultimately, the demarcation largely comes down to the tension between detail and scale. We can devise a very detailed coding schema which takes into account as many as possible of the details of the material. But as a consequence, the investigation will be severely limited in terms of how many languages can be included, which, especially for an exploratory typological investigation is a potential shortcoming. On the other hand, we can use a very broad coding, letting us include many more languages, but at the risk of not including important details. The balance between the two demands of course depends on the specific research question. Here, I have attempted to devise a minimally complex coding schema still allowing the research questions to be addressed, which makes a fairly large (in the semantic typology context)
language sample possible.

The precise decisions on what to include or not are made construction-by-construction and language-by-language. To illustrate the process and to exemplify what the coding looks like in practice, consider the three verses from the NT in Hungarian below, given un-glossed to reflect the state of the un-analysed data

(41) Hungarian (Uralic, Hungary)

a. John 4:32

˝Ő pedig azt mondta nekik: “Nekem van eledelem, amit egyem, amiről ti nem tudtok.”

b. Acts 23:34

Amikor a helytartó elolvasta a levelet, megkérdezte Páltól, hogy melyik tartományból való; és amikor megtudta, hogy Ciliciából,

c. Matt 7:20

Tehát gyümölcsékről ismeritek meg Őket.”

These verses are three of a hundred data points for Hungarian, for which we want to describe how the knowledge predication (or related concept) is expressed linguistically. This involves:

1. identifying which string of language corresponds to the knowledge predication
2. determining which constructions make up this string of language
3. deciding which of these constructions are relevant according to the criteria described above
4. recording this in an appropriate way

Beginning with step 1, there is no difficulty with this in most cases. Recall that the object of investigation is defined as predicated constructions where the predicate ascribes knowledge. In the majority of cases, a one-to-one mapping of translational equivalence can be established between these predicated constructions in different languages. For example, consider an English equivalent of 41a, given in 42. We here have a knowledge ascription (with the main lexical item of the predicate know, the Knower expressed as the subject of the predicate and the Known as a prepositional phrase).
“Nekem van eledelem, amit egyem, amiről ti nem tudtok.”

“I have food to eat that you do not know about.”

Table 3.2: Illustration of identification of translation equivalents in Hungarian

(42) English (Indo-European>Germanic, UK)

a. John 4:32

But he said to them, “I have food to eat that you do not know about.”

With almost exact correspondences, there is no difficulty in deciding on a translation equivalent in the Hungarian example, see table 3.2.

As mentioned in section 3.4, difficulties arise mainly when translations are very free and then in two main ways. In some translations, there is a many-to-one correspondence—in other words, what is one knowledge predication in the original is several in the translation. In these cases, I have chosen the constructions most closely resembling the original in terms of its parts (i.e. subject and object). The other issue is a lack of correspondence in the translation. That is, in the translated text there is no knowledge predication. This may be because the event depicted is construed differently (such as B telling A that X, rather than A finding out that X). Another common situation is knowledge predication in dialogue being omitted. In these cases, the verse is coded as missing a knowledge predication in this translation. The exception is when the role of the predicking constructions is filled by some other grammatical device of the language, typically evidential or epistemic markers. These forms of expressing knowledge ascription do not really fall under the scope of this study (i.e. knowledge predication) but are coded due to their high typological interest.

Having decided what string of languages expresses the knowledge predication, the next step is to determine what constructions make up this string of language. This is by no means simple. As was seen in the discussion of English, detailing the constructions involved in an expression is laborious even for a language well-known to the researcher. Most languages in this investigation were not well-known to the researcher. The texts comprising the New Testament corpus do not come pre-analysed or pre-glossed, and thus a large part of the analysis consisted in manual linguistic analysis of individual verses in individual languages. I have deliberately excluded from the language sample those languages for which no descriptive work exists, which means that some reference materials were available to aid in this analysis. The languages
in the sample range from large, national languages with a wealth of resources (e.g. Swedish or Hungarian) to severely under-described languages with only word lists or grammar sketches. The reference works used for each language are given in Appendix A.

The analyses of the relevant strings of language for 41 are given, glossed, in 43. From this, the move to list all relevant constructions involved is considerably shorter than before. Recall that the first criterion for inclusion in the coding has to do with the type of constructions. We are interested in i) lexical morphemes functioning as the predicates of the knowledge predication, ii) predicate construction, and iii) argument realisation constructions.

(43) Hungarian (Uralic, Hungary)

a. John 4:32

\begin{verbatim}
Nekem van eledel-em, ami-t egye-m, ami-röl ti
1SG.DAT be food-1SG.POSS REL-ACC eat-1SG REL-DEL 2PL

nem tud-tok
NEG know-2PL.INDF.OBJ
\end{verbatim}

‘I have food to eat that you do not know about’

b. Acts 23:34

\begin{verbatim}
és amikor meg-tud-t-a, hogy Ciliciá-ból,
and when PFV-know-PST-DEF.OBJ COMP PR.NAME-ELA

\end{verbatim}

‘and learning that he was from Cilicia’

c. Matt 7:20

\begin{verbatim}
Tehát gyümölcs-ei-k-röl ismer-itek meg őket
so fruit-PL-3PL.POSS-DEL know-2PL PFV 3PL.ACC
\end{verbatim}

‘you will recognise them by their fruits’

Using the components of this criterion as starting points, we can begin specifying for each verse which constructions are involved. Note that the lists are not strictly speaking complete, since they leave out for example very general constructions such as word class constructions or transitive constructions.

<table>
<thead>
<tr>
<th>Verse</th>
<th>Lexical morphemes</th>
<th>Predicate construction</th>
<th>Argument realisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>John 4:32</td>
<td>Nekem van eledel-em, ami-t egye-m, ami-röl ti</td>
<td>Nem tud-tok</td>
<td></td>
</tr>
<tr>
<td>Acts 23:34</td>
<td>és amikor meg-tud-t-a, hogy Ciliciá-ból</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matt 7:20</td>
<td>Tehát gyümölcs-ei-k-röl ismer-itek meg őket</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
For reasons explained above, it is both impractical and undesirable to specify all these constructions for each of the one hundred verses in each language, and the second delimiting criterion was introduced—that the construction is differentiating within the domain. To exemplify this, consider the list of Hungarian constructions above. We note that the Nominative Subject construction occurs across all three contexts and it is indeed the case that it occurs in all verses in the sample. Thus, there is no need to code the verses for this. The opposition between the object being in the Delative case as opposed to the Accusative, however, is relevant, since it is a differentiation within the domain. Applying this criterion of differentiation (this is of course done across the entire verse sample) we are left with the considerably shorter listing in Table 3.4.

<table>
<thead>
<tr>
<th>Verse</th>
<th>Lexical morphemes</th>
<th>Predicate</th>
<th>Argument realisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>John 4:32</td>
<td>tud-</td>
<td>-tok (=Person agreement), Present tense, nem-negation, Indefinite object marking</td>
<td>Nominative Subject, Delative Object, Relative Pronoun Object, Pronominal Subject</td>
</tr>
<tr>
<td>Acts 23:34</td>
<td>tud-</td>
<td></td>
<td>hogy-clause, null subject</td>
</tr>
<tr>
<td>Matt 7:20</td>
<td>ismer-</td>
<td>-itek (=Person agreement), Present tense, nem-negation, Indefinite object marking</td>
<td>Null subject, Accusative Object, Pronominal Object</td>
</tr>
</tbody>
</table>

**Table 3.3:** Constructions involved in the realisation of three verses in Hungarian

We are now approaching our goal. A final missing step, however, is how to represent these analyses in a way which is appropriate for further (automated)
<table>
<thead>
<tr>
<th>Verse</th>
<th>Coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acts 23:34</td>
<td>L: tud-, P: meg-, S: , O:</td>
</tr>
<tr>
<td>Matt 7:20</td>
<td>L: ismer- P: meg-, S: , O:</td>
</tr>
</tbody>
</table>

**Table 3.5:** Constructions involved in the realisation of three verses in Hungarian

analysis. Most importantly, we want to be able to compare how similar the expressions of different verses are. We also want to be able to say something about how meaning is expressed in the verses of the sample and, relatedly, *how* verses are different or similar in a given language. For example, it might be of interest to know that a verse is expressed with the same lexical root but with some additional derivational morpheme or that some contexts are more prone to non-canonical subject marking. In other words, we want a way of comparing like with like, which requires some kind of categorisation. On the other hand, this categorisation cannot be too fine. For example, we want to compare case markers to other case markers, but also to adpositions marking the arguments.

The solution applied here is to use a quite broad categorisation scheme, based on the analysis of predicating constructions. The constructions selected according to the criteria above are divided into four groups: lexical roots making up the predicate, any further marking of the predicate and, as a further refinement of the argument realisation constructions, marking of the Knower and marking of the Known. For the Hungarian verses, this gives the result in Table 3.5 where the letters are mnemonic for Lexical, Predicate, Subject and Object, the latter two being more convenient ways of briefly referring to Knower and Known.

I will refer to the elements of this coding as *markers*, distinguishing between *lexical* and *constructional* markers. By lexical markers I mean those lexical items forming the predicate, that is, the markers in L in the coding. The remaining markers (P, S and O) are referred to as constructional markers. This divide might seem artificial in view of the construction grammar arguments against a strict divide between lexicon and grammar, but is motivated by the need to distinguish between two groups that in the data generally show different behaviour.

This coding, however, only partially captures an essential part of language, namely that markers often occur in paradigms or paradigm-like groups. That is, markers A and B in a language are not best described as independently being present or absent if A and B occupy the same morphological or syntactic slot. In other words, the situation should not be described as two variables with the possible values absent or present but as one variable with the possible
values A or B (and possibly null). As an example (not particularly relevant to the investigation at hand), consider person marking. Clearly, some way of finding out which markers in the data are in a paradigmatic relationship to each other is desirable.

The only truly accurate way of doing this is to devise individual coding schemas for each language in which the paradigmatic relationships are specified. Devising such schema requires considerable analytical work and it cannot be assumed that the information necessary for it can be retrieved from grammars (although it undoubtedly will be in some). It is therefore not feasible for a project of this kind. Rather than ignoring the problem, I have tried to partially tackle it through an automated procedure where markers are sorted into paradigm slots based on their distribution.

The basic idea is that if two markers co-occur, they are not in a paradigmatic relationship to each other. Starting in each language with a list of all markers coded for and their distribution across the 100 verses, the co-occurrence patterns are easily found. A first sorting process is already done in the coding—markers are coded as either lexical, other markers belonging to the predicate and so on. For lexical markers, the notion of paradigmaticity is less relevant. The most common situation is for each verse to have exactly one lexical marker, namely the root of the predicate. In some cases, more than one lexical marker is relevant to code for, e.g. in cases of compounds, serial verb constructions or other complex verb constructions. It may be that there are cases where a paradigmatic structure can be found in these kinds of construction, but since they are peripheral to the investigation at hand, lexical markers have not been included in the following procedure of finding markers that appear to be in a paradigmatic relationship to one another.

For the remaining three categories the procedure used is as follows. In each language, we start with the group of all markers in each category. The goal is to find groups of markers which do not co-occur. This is done by continually checking co-occurrence for smaller and smaller groups, starting with a group of size \(i\) where \(i\) is the number of markers and continuing down through all possible groups of size \(i-1\), \(i-2\) etc. until \(i-x=1\) is reached. In this way, the largest possible groups of markers which do not co-occur are found. Groups of markers uncovered by this procedure are then grouped into a “slot”. Markers which remain unassigned at the end of this procedure, i.e. which do not stand in a relationship of non-co-occurrence with any other markers, are assigned to their own slots.

As an example of the procedure, consider again Hungarian. It turns out that the three predicate-level markers \(meg-\), \(fel-\) and \(vissza-\) do not co-occur, and they are thus sorted into one slot—i.e. slot 0. As mentioned, the subject constructions remain invariant across the verses in the sample, and thus no mark-
ers are coded for there. For the Known, we find various case markings as well as the infinitive complement. Accusative objects and hogy-complements are coded as unmarked, i.e. without markers. Note that this (to ignore the difference between nominal and clausal complements) is a methodological choice, in accordance with the criterion on domain-specific differentiation. All the markers on the Known in Hungarian are non co-occurring with each other and are thus sorted into a single slot—again slot 0, but this time within the Known category. Slots are only compared within categories. Using the contexts discussed above, the results can be illustrated as in table 3.6 where the subscript number indicates the slot.

<table>
<thead>
<tr>
<th>Nr</th>
<th>Verse</th>
<th>L_x</th>
<th>P_x</th>
<th>S_x</th>
<th>O_x</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>John 4:32</td>
<td>tud-</td>
<td>DEL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Acts 23:34</td>
<td>tud-</td>
<td>meg−0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Matt 7:20</td>
<td>ismer-</td>
<td>meg−0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.6: Examples of codings of verses from Hungarian

Note that the results obtained by this automated procedure bear great similarity to how the language is normally analysed manually. For instance, the markers mentioned as belonging to the predicate are described as coverbs or verbal prefixes and as a grammar states “[o]nly one coverb can be attached to a verb at a time” (Rounds 2009).

To summarise, this section has specified what the object of investigation is on the linguistic side. It was argued that a narrow focus on lexical items is not sufficient to fully capture the variation within many domains, including knowledge. A framework which allows lexicon and grammar to be integrated into a whole is construction grammar, which argues that language in its entirety is a collection of form-meaning pairs — constructions — of varying levels of complexity and generality. Therefore, the objects of investigation in this thesis are the constructions involved in a knowledge predication. Since construction grammar takes everything to be constructions, this is a very long list. Therefore, two criteria for delimiting what construction are considered are formulated: first, a criterion limiting the type of construction and second, a criterion that mainly constructions which differentiate within the domain are considered.

The selected constructions are coded in the form of grammatical markers, in four broad categories: lexical forms making up the predicate (the lexical markers), further grammatical markers on the predicate and grammatical markers on the Known and Knower respectively (the constructional markers). Within these categories (excluding lexical items for which it is largely irrelevant) the markers are automatically sorted into slots based on co-occurrence,
where markers in the same slot are assumed to stand in a paradigmatic relationship to one another. These ordered collections of markers are the basis for the computing of similarity on which this thesis is based, and which will be described in more detail in chapter 5.

Given the 100 verses and the 83 languages included in the investigation, there are 8300 potential codings. Some of these are for various reasons missing. Three types of missing data were used in the coding:

- <empty>—the relevant verse or passage is missing the particular translation.
- NA (not applicable)—the relevant verse or passage exists but there is no knowledge predication. A common example is for ‘learn’ or ‘find out’ to instead be given as speech (cf. the Fasu example above: 25).
- NC (not clear)—the analysis of the verse is not clear to me; I cannot identify the knowledge predication but also cannot with sufficient certainty say that there is no knowledge predication.

In table 3.7, the statistics regarding these values in the coding are shown, for the full verse sample as well as for the knowledge verses specifically (as demarcated in chapter 5).

Note that NA values are only in a very specific sense missing data points—they are analysed and found to lack knowledge predications. Further analysis of the alternative strategies lies outside the scope of this thesis. The largest contributor to the relatively high proportion of NC codings in the full verse sample is the ability verses. A combination of sometimes only morphological marking (i.e. no lexical item to identify) and the frequent occurrence of no apparent marking at all make these difficult to analyse satisfactorily. In cases where it is clear that there is no colexification with ‘know’, I have applied an NC value fairly early in a difficult analysis, considering the peripheral role of these contexts to the investigation. Note that this may have artificially increased the similarity between knowledge and ability (since missing data points are not included in the similarity measure, see chapter 5), but since the

<table>
<thead>
<tr>
<th>Value</th>
<th>Full</th>
<th>Knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>%</td>
</tr>
<tr>
<td>&lt;empty&gt;</td>
<td>54</td>
<td>0.7</td>
</tr>
<tr>
<td>NA</td>
<td>358</td>
<td>4.3</td>
</tr>
<tr>
<td>NC</td>
<td>445</td>
<td>5.4</td>
</tr>
</tbody>
</table>

Table 3.7: Number of occurrences of different coding values
end result is separate clusters anyway (see again chapter 5) this has no impact on the results of this thesis.

3.7 Summary

In this chapter, I have presented the theoretical frameworks and general methodology underlying this investigation.

The investigation is governed overall by the insight of **constructional typology** that just lexicon is not enough to describe how meaning is expressed cross-linguistically. Based on Koch’s (2012) three identified problems, I gave examples of the same problems regarding knowledge predication, and I described a coding schema which takes this insight into account. Constructional typology is based on **construction grammar** which was briefly described. The key tenet of construction grammar for this investigation is that meaning is not limited to the lexicon, necessitating the more extensive approach to coding taken.

The other central theoretical framework is **cognitive linguistics**, from which several important concepts are drawn to explain the results of the investigation. In particular, **scenes** and **frames**, **construal** and **prototype categories** will play central roles in the discussion of results. Furthermore, **domain** is an important concept to this investigation, as to semantic typology in general. I have argued in this chapter that (at least) three senses of **domain** need to be distinguished: **conceptual domains**, **comparative domains** and **generalised domains**.

In addition, the theories and results of Natural Semantic Metalanguage have been given some attention in this and the previous chapter. This thesis is not a work within NSM, nor is it a work against NSM. Rather, the relatively prominent place of NSM in the discussions reflects the weight of the framework’s contribution to our current understanding of both the typology of knowledge predication in particular and semantic typology in general. This makes it necessary and fruitful to at times engage with the claims of NSM.

Moving from theoretical frameworks to methodological issues, the main method of this thesis is **parallel text typology**. This chapter has described the theoretical basis of this method and has given an overview of some challenges particular to this method, drawing on work within translation studies.

I described the process of choosing the verse sample—the verses whose translations are included in the investigation—and summarised what the sample looks like. Following from the discussion of constructional typology mentioned above I then described the coding schema used for each verse in each language. The schema’s design was guided by the need to balance between, on
the one hand, cross-linguistic comparability and the individual characteristics of languages and, on the other hand, detail and coverage.

In short, the resulting schema describes the knowledge predication of each verse in each language as a collection of markers, ordered in the four categories of lexical markers making up the predicate (L), further constructions involved in predicate realisation (P), marking of the knower (S) and marking of the known (O). It is these collections of markers which form the basis of the empirical investigations of this thesis. The details are described in subsequent chapters, as summarised below.

In chapter 5, section 5.2, I describe how the coding is used as the basis for a paradigm-adjusted similarity measure, describing mathematically how similar any two verses are to one another. In the same chapter, the similarity measure is used as the input for Hierarchical Agglomerative Clustering (described in section 5.3) to investigate research question A1.

In chapter 7, a slightly modified version of the similarity measure is used as the input of a Classical Multi-Dimensional Scaling analysis (described in section 7.2) to begin the investigation of research question B and its sub-questions. The results of this analysis—in the form of dimensions and the verses’ location in them—are also the basis for the results presented in chapters 9 and 11.

In chapter 9, I investigate the semantics of the dimensions by considering the markers associated with them, determined by using Point-Biserial Correlations. Chapter 8 uses the clusters introduced in chapter 5 and investigates quantitative differences between them regarding complexity and the proportion of outside colexification, thereby addressing research question B1. In the chapter, I also introduce a quantitative measure of directionality of formal motivation, presented in section 8.2.1. In chapter 10, the similarity measure and clusters of chapter 5 are used to address research question B3 concerning the connections of knowledge to other concepts.

Unless otherwise stated, all methods and procedures are implemented in the programming language python (Python Software Foundation n.d.) by the author. The proximity semantic maps in this thesis (such as figure 7.4 on p. 152) were created using R (R Core Team 2021) with the libraries ggplot2 (Wickham 2016) and cowplot (Wilke 2021). When possible, colour as meaning bearer is avoided in illustrations, but when necessary a palette friendly for the colour-blind is used (colourblind).

Network graphs (such as figure 5.4 on p. 122) were create in R using the libraries igraph (Csárdi et al. 2023) and ggraph (Pedersen 2022). Raincloud plots (such as figure 8.4 on p. 173) were created using ggdist (Kay n.d.) in addition to ggplot2. Maps were created using the ArcGIS Pro 3.1 software by ESRI, with base map data from Esri & the GIS user community (2021).
4. Language sampling

4.1 Introduction

By most estimates, there are around 6000-7000 languages in the world, and far from all are sufficiently documented—Glottoscope currently (1.6.2023) counts 7837 languages of which only around 65% are described in at least one grammar (Hammarström et al. 2018). It is therefore impossible for any typological investigation into the languages of the world to study the entire population of interest. The issue is made worse if the object of investigation is seen as all languages that have ever existed, ever will exist or all possible languages, rather than just all currently extant languages (cf. Bakker 2010, Cysouw 2005). In this case, studying the entire population of interest is not only practically but principally impossible.

Therefore, typological studies generally use a sample. In this chapter, I will present the sampling method used in this thesis. The chapter begins with a brief background on sampling in typology and the method on which this thesis’s method builds, i.e. the Diversity Value method (section 4.2) and then, step by step, describes the in this thesis used development of the Diversity Value method—geographically clustered Diversity Value sampling (section 4.3). Section 4.4 summarises the chapter.

4.2 Sampling in typology and the Diversity Value method

Purely random language samples—that is, samples where each member of the population has an equal chance of being included in the sample—are rare in typology (Veselinova 2016a). Random samples can be useful, Bakker (2010: 106) argues, when “research parameters […] are not (directly) related to genetic affiliation or to areal considerations, and for variables which are highly unstable and show a high level of variation”. In most cases, however, this is not the case and we expect the distribution of variables not to be independent of these considerations. In other words, we expect certain biases.

Bakker (2010) distinguishes (i) genealogical (or genetic) bias, (ii) areal bias, (iii) cultural bias and (iv) typological bias. In the first three cases, related-
ness, closeness or similarity in a non-linguistic domain is assumed to correlate with similarity in linguistic structure. The fourth kind of bias is similar, but it is linguistic factors that are expected to correlate. For example, a bias in the sample toward SVO languages can lead to biases in noun-phrase order, since the two are known to correlate.

The assumptions of bias may, from the opposite viewpoint, be expressed as the assumption that there is correlation between different types of diversity. Nettle (1999) distinguishes five types of diversity which a collection of languages, e.g. as spoken in a country, can exhibit; (i) language diversity, the presence of a large number of individual languages, (ii) phylogenetic (genealogical) diversity, the presence of many different families or subfamilies, (iii) structural diversity, a high degree of variation in structural characteristics and (iv) cultural diversity, the presence of many different cultures among the speakers. In the case of not originally areal groups, there can also be (v) areal diversity—e.g. a language family spread over several continents.

The general rationale of typological language sampling is then that if we wish to maximise the structural diversity of a sample, we should maximise other types of diversity, as we expect a correlation. This gives what is known as a variety sample—a sample concerned with capturing the full range of variation in a phenomenon but not (primarily) in making claims about frequency or likelihoods. A sample with this goal is known as a probability sample. For overviews of sampling in typology see e.g. Bakker (2010) or Veselinova (2016b).

There exists two fully systematised and formalised—a formalised method of picking languages is a methodological strength from the point of view of reproducability—general methods specifically designed for variety sampling: the Diversity Value method of Rijkhoff & Bakker (1998), a further development of Rijkhoff et al. (1993) and the Genus-Macroarea method (Miestamo 2005). In evaluating the two methods based on World Atlas of Linguistic Structure data (Dryer & Haspelmath 2013, Miestamo, Bakker & Arppe 2016) found that they perform significantly better (detect more variety) than a random sample. They found also that with smaller samples (<ca 200 languages) the Diversity Value method performed better. This motivates the use of an adapted version of the Diversity Value method in this study.

The Diversity Value method is meant to maximise genealogical diversity in a sample. The main idea is that each language family is assigned a Diversity Value, based on the number of branchings in the genealogical tree. The levels of branching are weighted so that those “closer to the top”, i.e. those at the assumed greatest time depth, receive greater weight. A language family with a relatively higher Diversity Value will then be assigned a greater number of languages in the sample, with the motivation that a language family with a
greater number of splits at a greater time depth is more likely to display internal structural variety.

In the original method (Rijkhoff et al. 1993), it was recommended that the minimum sample size equal the number of families in the chosen classification, as “there is no a priori reason why one family would be more interesting than another” (Bakker 2010: 117). With modern classifications (Hammarström et al. 2019, Eberhard & Fennig 2023), this puts the minimum sample size at around 250 languages. For many investigations, this is unrealistically large. Thus, this demand has in some versions of the method been replaced by a method of choosing families based on a weighted random choice, where families with large Diversity Values are more likely to be included (Rijkhoff & Bakker 1998, Miestamo, Bakker & Arppe 2016).

It is not clear how this counters the objection of Bakker (2010). Say we are adding a final language to our sample. We are choosing between a language from a mid-sized family and an isolate. Why would we expect that a language from the mid-sized family would be more different from our already chosen languages than the isolate, simply on account of being part of a more complex family?

In the sampling procedure applied here, I approach the problem of more families than places in the sample in a different way. The Diversity Value method aims to maximise genealogical diversity. I propose a modification which solves the problem by adding maximisation of areal diversity to the method.

4.3 Geographically clustered Diversity Value sampling

In table [4.1] below, the procedure applied here is summarised and the steps where it departs from Rijkhoff & Bakker (1998) are marked. The steps are explained in greater detail in the following subsections. The procedure is also illustrated in figure [4.1] below, following the explanation of step 7.

The sampling procedure in this study focuses on maximising linguistic diversity by maximising genealogical and areal diversity. Cultural and structural diversity are not directly taken into account—it is assumed that the maximisation of areal and genealogical diversity will also work toward maximising these.

1The exact number depending on how one deals with pidgins and creoles, unclassified languages, sign languages etc.
Step 1. Defining the sampling frame

The first step of the sampling procedure is to establish the sample frame, using the terminology introduced in Bell (1978). The frame, in Bell’s terminology, is the set of language from which the sample is drawn and is the concrete reflection of the sampling universe—the entire population of languages that is the object of study. As discussed above, the universe may, also in a world-wide study, be conceived of in somewhat different ways—all existing languages, all languages that have ever existed or all possible languages.

In this investigation, the Glottolog (v.3.4) (Hammarström et al. 2019) serves as the sampling frame. The Glottolog aims at total comprehensiveness in coverage of linguistic varieties. Furthermore, the genealogical classification described by the Glottolog—an important part of a genealogically based sampling method—is conservative, only accepting solid evidence for family groupings (the exception being languages grouped into “pseudo-families” such as sign languages). Given that the method used here gives greater weight to small families, this “splitting” approach (as opposed to a “lumping” approach) is conducive to maximum genealogical variety. Finally, the Glottolog is freely available, ensuring maximum reproducibility.

For methodological reasons, a number of adjustments were made to the
frame. Extinct languages are excluded to minimise documentary bias—the complexity added by extinct languages and branches to, e.g., Indo-European means a higher Diversity Value and more representation in the final sample. However, this added complexity has little to do with the family itself, but is rather due to the status of documentation and research. A less documented and researched family of comparable complexity and size likely also has lost members, but without knowledge of them, they cannot be included in the tree—creating a bias toward well-documented and researched languages. This bias is, of course, also present in other ways which are not as easy to single out—making it all the more motivated to counter it where possible. Sign languages are also excluded. The reason for this is simply that there are extremely few New Testament versions in a sign language and they therefore cannot be studied with the chosen method of this study. Finally, pidgins and creoles are also excluded from the frame due to the difficulty of deciding where in and in what family tree to place them—for a sampling method so reliant on accurate family trees as the Diversity Value method, this is a problem.

These adjustments clearly influence what universe the sample can be expected to represent, moving it from all possible languages to all extant, non-sign, non-pidgin or creole languages. To in a small way counteract this, the sample is divided into a core sample of 95 languages—drawn from the sampling frame according to the procedure described below—and an extension sample—5 languages from the excluded groups (though, unfortunately, no sign languages due to lack of materials).

In practice, sampling in typological studies is generally constrained in another major way—many languages lack sufficient documentation and description. Based on this, a distinction can be made between a catalogue frame—a collection that includes all languages that exist or have ever existed to the best of our knowledge—and a corpus frame—the set of languages for which appropriate research materials for the question at hand are available. In the case of this study, the corpus frame is the set of languages for which there is a New Testament translation available.

There are then two approaches to drawing the sample; sampling from the catalogue frame—and then discarding those groups where there is no corpus representation—or sampling directly from the corpus frame. The first approach is taken here. The reason for this is methodological; it allows us to see where exactly the sample lacks representation.

1Concretely, those in the Pidgin family or a Creole subgroup.
Step 2. Dividing the language into macro-areas

The languages in the sampling frame are divided geographically into five macro-areas: North America, South America, Eurasia, Africa and Oceania. The method of sampling area by area is an old one, going back to Dryer (1989), the motivation being that languages show areal patterns even at a very large continent-scale level. This is not a feature of the original Diversity Value method and has been applied to maximise geographic variation and minimise macro-areal biases. However, note that, unlike in Dryer, it is the languages and their current physical locations which are considered, not families or genera.

Many different suggestions for what macro-areas the languages of the world should be divided into can be found in the literature. The division here follows that of the Glottolog, including Meso-America in North America. The areas Papunesia and Australia are here combined in the area Oceania—the reason for this is the very few Australian languages in the corpus. See the Glottolog database for details of this division (Hammarström et al. 2019).

A language family may be present in more than one area. If that is the case the two parts are treated separately: that is, separate family trees are constructed from the languages of each area. This means that the sampling procedure can remain agnostic in questions of the homeland of widely spread families. However, families with only a marginal presence in terms of the number of languages in an area, for instance Indo-European in the Americas, are excluded in those areas. This is determined by dividing the number of languages in a family in an area with the total number of languages in the family. If the number of languages from a family in an area is less than 10% of the total number of languages in the family, they are excluded. This is of course somewhat arbitrary and is done for convenience.

Each area is assigned an equal number of languages, specifically nineteen. The decision to give all areas equal weight comes from the exploratory nature of the study.

Step 3. Determining the Diversity Values

The algorithm of the Diversity Value method is then applied area by area, rather than as in the original method for the whole sample at once. The procedure is as follows. For each family, the genealogical tree as provided by the Glottolog is retrieved and the Diversity Value (DV) is calculated. The contribution of the uppermost level, immediately under the family level is equal to the number of nodes at that level. The contribution of the level under that is equal to the number of nodes in that level minus the number of nodes in the above level, together multiplied by the maximum number of levels in any family minus the number of levels removed from the top, in this case one. The
contribution of all levels is then added together and divided by the number of levels in the family tree, yielding the Diversity Value of the family. In this way, all splits in a family tree are weighed into the final Diversity Value, but the higher up in the tree they are—that is, the older—the greater weight they are given, reflecting the assumption that families with branches that are higher up in the tree and thus more likely to have split a long time ago are more likely to show variety than those with recent splits. For a more detailed description as well as examples see Rijkhoff & Bakker (1998).

Isolates and families with no intermediary levels between family and language are assigned the Diversity Value of 1.5. This gives them a Diversity Value smaller than the smallest value possible for families with internal branching (which is 2.0). In Rijkhoff & Bakker (1998) isolates are instead assigned a Diversity Value of 1.0. This is adjusted to 1.5, to allow for step 4 (see below). Note that this also better reflects the view that “we consider Language Isolates to be the last surviving members of previously existing larger families” (Rijkhoff et al. 1993: 179).

Step 4. Calculating the logarithms of the Diversity Values

The next step is one of two ways in which the sampling procedure of this study departs most sharply from that proposed by Rijkhoff & Bakker (1998). The base 1.5 logarithm of all Diversity Values is used for calculating the number of languages assigned rather than the original values. This is a weighting procedure, used to give greater weight to small families and isolates in the final sample (cf. the next section). The number 1.5 is the value assigned to isolates and minimal families and is therefore used as a baseline.

Step 5. Calculating the assigned number of languages

As discussed above, the original Diversity Values method deals with more families than languages in the sample by a weighted random choice. There are two disadvantages to this. The first is the already mentioned issue of there being no reason to assume that a language from a more complex family will be more distinct than a language from a small family, all else being equal. The second is that there is reason to assume that two geographically close languages from different families may in some respects be more similar than two languages from opposite ends of a large, complex family. Thus, it may be warranted to let e.g. Indo-European be represented by more than one language at the expense of some smaller family, given that this smaller family is, so to speak, represented by a geographical neighbor.

The procedure is adapted to capture these intuitions—formulated above as the goal of maximising both genealogical and geographical diversity. Thus,
the Diversity Values of all families are summed and the number of languages assigned to each family is calculated by multiplying the total number of language in the sample by the family’s Diversity Value divided by the sum of Diversity Values (area by area). The resulting value (the In-Sample Value) is the number of languages from that family that should be included in the sample. This is where step 4 is important. Using the raw Diversity Values gives too much weight to a few very complex families (Indo-European, Austronesian etc.). Weighting by means of logarithmic values gives a better balance between the small and large families.

With the In-Sample Values in place, the sampling proceeds as follows. If the rounded In-Sample Value of a family equals 1, it is added as a sample group—that is, a group from which one language is added to the sample. Otherwise, one of the following steps is applied.

Step 6. Determining subgroups

If the rounded In-Sample Value for a family is greater than 1—that is, if the family would be represented by more than one language in the sample—the method is applied again to the subgroupings of the family until a level in the tree is reached where the In-Sample Value is 1 or below.

Step 7. Geographical clustering

All families with an In-Sample Value smaller than one are grouped according to geography. The number of groups is that of the sum of the In-Sample Values, that is, the number of languages that the families together should be represented by in the sample. In the final sample, a language is chosen from each group.

The clustering is done using the \(k\)-means algorithm, as implemented in the scikit-learn Python-library \(\text{(Pedregosa et al. 2011)}\). To further take into account the constraint that groups should be equal in size with regards to Diversity Value, the algorithm is applied in two steps. In the first step, the algorithm is applied to all the families, represented as the midpoint of all languages in the family. In the second step, the algorithm is applied again to any cluster where the sum of In-Sample Value is greater than two and then re-applied to the remaining families. This gives relatively greater weight to geography than equal size of In-Sample Value for the clusters. A different approach could be taken, giving greater weight to In-Sample Values, but the risk then is that families that are in fact quite far apart, and thus have no actual chance of diffusion, are grouped together to satisfy this constraint. Therefore, a primarily geography-based clustering is to be preferred.
<table>
<thead>
<tr>
<th>Area</th>
<th>Nr of missing sample groups (/19)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>5</td>
</tr>
<tr>
<td>Eurasia</td>
<td>3</td>
</tr>
<tr>
<td>Oceania</td>
<td>6</td>
</tr>
<tr>
<td>North and Meso-America</td>
<td>7</td>
</tr>
<tr>
<td>South America</td>
<td>2</td>
</tr>
</tbody>
</table>

**Table 4.2:** Number of sample groups in each area without NT translations

The same thing is done for the subgroups from step 6, with the constraint that only subgroups from the same family are grouped.

Steps 2–7 results in 95 sample groups that between them are maximally genealogically and geographically diverse.

**Step 9. Choosing individual languages**

The choice of specific language from each grouping is not automated in my application of this method. Instead it is done manually, based on several factors. The main factor is the availability of a New Testament translation. Only languages where there is a sufficiently complete translation are included in this study. The criterion for sufficiently complete means that translations of, for example, only one gospel are excluded. This means that some groups decided on by the sampling procedure are not represented in the final study, since there are no translations for any language in that group.

Also important for the choice of language in each group is the availability of other resources, such as grammars and dictionaries. Languages where these exist are preferred over those where they do not (no sample group had to be excluded entirely on the grounds of insufficient description). Clearly, these two criteria mean that there is a certain bibliographic bias to the sample. However, applying the sampling procedure to an as-complete-as-possible catalogue (of course, any catalogue of the world’s languages suffers from some kind of documentary bias) means that the bibliographic gaps and resulting biases become apparent.

In table 4.2 the numbers of sampling groups without translations or with insufficient translations is given. Note also that in total four sample groups are represented by two languages (from different families) each—one in Africa, one in North and Meso-America and two in Oceania.

In addition to the practical considerations discussed above a number of

---

1The reason for this is that the analysis work began according to an earlier version of the sampling method with slightly different groups; from the perspective of variety sampling, excluding already analysed languages makes no sense.
**Figure 4.1**: Illustration of the sampling procedure
other considerations went into the choice of specific language. First of all, if
some interesting feature in the domain in the language was known from some
source, I was more likely to include it. Secondly, languages where I had access
to speakers were more likely to be chosen, though of course this was relevant
only for very few groups. Thirdly and fourthly, the number of speakers and
the literacy rate was taken into account, with an aim to maximise the variation.
This was not done in any systematic way, and the sample is not in any way
balanced in regard to these parameters, but there is variation. Fifthly, all other
things being equal, I preferred modern Bible translations to older ones.

As mentioned, an extension sample of 5 languages was also selected, one
from each macro-area, from the languages excluded in step 1. These are all
Germanic languages: a historical language (Middle English), two geographi-
cally “displaced” varieties (Afrikaans and Pennsylvania German) and two cre-
oles (Kriol and San Andres Creole English). Thus the full language sample
consists of 83 languages—the core language sample, genealogically and geo-
graphically stratified, of 78 languages and the extension sample of 5 languages.
The locations of the languages in the sample, as well as those languages which
are part of sample groups without materials[1] are shown in map 4.2.

The sample is given in full in Appendix A.

4.4 Summary

Due to the large number of languages in the world, most typological inves-
tigation requires sampling. Variety samples aim at maximising the linguistic
variation captured in a sample. One of the fully formalised methods for va-
riety sampling is the Diversity Value method, which, however, offers no way
to choose between language families in a smaller sample. In this thesis, I
therefore apply a development of that method wherein also geography is taken
into account. In addition to calculating Diversity Values, language families
are clustered according to their locations, which distinguishes sample groups
of assumed minimal internal variety from which languages can be drawn. In
this thesis, this method is used to yield 95 sample groups evenly distributed
over five macro-areas. After excluding groups lacking New Testament transla-
tions, a 78-language core language sample remains. This is supplemented by
a 5-language extended language sample.

[1] Note that this is not the same thing as individual languages for which there is no NT
translation—even if a language lacks an NT translation but is part of a sample group in which
at least one languages does not, it is not shown of this map
Figure 4.2: Map of the languages in the sample as well as sample groups without NT translations
5. Demarcating knowledge

5.1 Introduction

As stated in the introduction, this thesis is centred around two sets of research questions, each set specifying an overarching question, repeated below.

A What is the cross-linguistic status of knowledge predication as a distinct construction?

B What linguistic distinctions are made in knowledge predication—what meanings are distinguished?

The answer to question A has relevance for question B. As I argued in section 3.2, domains can be distinguished in different ways. I distinguished between comparative domains—constructs of the researcher—conceptual domains—coherent areas in semantic space—and generalised domains—abstractions from individual languages.

Semantic typological investigations often take as their starting point comparative domains. These are typically based on research and intuitions about how a corresponding conceptual domains, generalised domains or both look. That is, the researcher assumes that the domain which they are studying has sufficient internal coherence to be a reasonable object of study. In many cases, this seems a completely warranted assumption. In other cases, however, such assumptions seem more questionable, knowledge being a case in point. As I argued by example in the Introduction, we do not really know enough about the cross-linguistic expression of these meanings to make any firm statements about what the generalised domain of knowledge looks like. Identifying a conceptual knowledge domain is also difficult. Whereas some concepts have a clear basis in the physical world—such as the senses—and may thus safely be assumed to be shared across cultures, knowledge concerns social reality and it is thus a strong assumption that it is something that exists or looks the same way across cultures and societies.

For these reasons, the grounds for constructing a comparative knowledge domain as a basis for investigation are inadequate. The goal of this chapter is to remedy this, by presenting the results of an empirical investigation into
A1. Is knowledge predication generally distinguished in languages, and if so, what are the boundaries of it?

More concretely, I examine whether there is a group of contexts in the data that are regularly expressed in the same way as each other but differently from other contexts and which in meaning roughly corresponds to ‘know’, and if so, which these contexts are.

To answer this, we need two things: (i) a measure of similarity of expression (discussed in section 5.2) and (ii) a way of dividing the contexts into groups based on that similarity of expression (discussed in section 5.3). With these ingredients in place, I present the results of agglomerative clustering applied to the data in section 5.4 and some methodological challenges in taking the results as the basis for a generalised domain are discussed (section 5.4.1). Section 5.5 summarises and concludes the chapter. Table 5.1 give an methodological summary of the chapter.

### 5.2 Measuring similarity

There are countless ways in which similarity of expression can be described. I will focus here on methods aimed at describing the similarity between two data points $a$ and $b$ as a numerical value, which can serve as the input of further statistical methods. There are other ways of describing similarity, such as describing the relationship between words as identical, derivational or unrelated or describing one construction as an instance of another, but these are...
not considered here. The focus is furthermore limited to measures of the similarity of contexts—also here there are other possibilities, such as comparing abstract functions (e.g. Croft & Poole 2008) or whole languages (e.g. Hartmann, Haspelmath & Cysouw 2014).

A commonly used similarity measure to numerically describe the similarity of two contexts with multiple expressions—such as in this thesis—is some variation of what is usually referred to as the relative Hamming distance, which is the number of elements in which two sequences differ divided by the number of elements (Hamming 1950). Examples of studies using such a distance measure are Wälchli & Cysouw (2012) in which the sequences compared are the forms used to express motion in translations of New Testament verses or Croft & Poole (2008) in which the contexts are translations of questionnaire sentences (from Dahl 1985) and the sequences compared is the use of tense and aspect categories. Basically, each context is represented as a sequence consisting of the value it has in each language, and the distance between contexts is calculated by comparing the sequences. We might call this the binary application of the Hamming distance.

The main drawback of this method is that it only takes identity or non-identity between coding of contexts into consideration. We consider only whether two verses have the same value in a language or not. This same-ness can be approached in different ways—we might consider the codings as a whole or, as done in Wälchli & Cysouw (2012), consider the lexical roots used. Regardless of the approach chosen, the result is that contexts in any given language are either coded as identical or not similar at all. It is therefore not a good solution for an investigation explicitly interested also in constructional differences—i.e. in partial similarity.

As an illustration, consider the three verses in Hungarian which where discussed in section 3.6. The coding of these verses is repeated in 5.2 (see section 3.6.2 for a description of the notation used). Clearly, the codings for knowledge in verses 1 and 2 are similar to some degree (they share the same lexical root) but also different to some degree (one has the object in Delative case, and one uses the derivative morpheme meg-). We would like a distance measure which captures this kind of partial similarity or, in other words, a way to represent contexts as partially similar based on partially shared characteristics.

<table>
<thead>
<tr>
<th>Nr</th>
<th>Verse</th>
<th>$L_x$</th>
<th>$P_x$</th>
<th>$S_x$</th>
<th>$O_x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>John 4:32</td>
<td>tud-</td>
<td></td>
<td></td>
<td>DEL</td>
</tr>
<tr>
<td>2</td>
<td>Acts 23:34</td>
<td>tud-</td>
<td>meg-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Matt 7:20</td>
<td>ismer-</td>
<td>meg-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.2: Examples of codings of verses from Hungarian
More specifically, we are after a similarity measure where:

1. contexts in which all elements are shared are maximally similar
2. contexts in which no elements are shared are maximally dissimilar
3. contexts in which some elements are shared are partially similar

Several previous studies have implemented some version of such a similarity measure. For example, Wälchli (2010) in comparing locative markers, assigns an intermediate value to markers which are not identical but share morphological material. Another study which implements such a similarity measure is Levshina (2015) in which analytical causatives are investigated. The contexts compared are translation equivalents, coming not from the Bible, but from a corpus of movie subtitles. The contexts in each language are coded for 5 parameters: the causing-verb, the form of the caused event (e.g. infinitive, finite), transitivity, coreferentiality of causer and another participant in the causation event and finally active or passive form. Complete overlap yields a similarity of 1, no overlap a similarity of 0. When there is partial overlap, the different features are weighted differently, so that having the same verb counts as more similar than having the same value on the transitivity parameter. Levshina takes a data-driven approach to weighting and settles on “the one that led to results that could best be interpreted from a semantic perspective” (Levshina 2015: 498).

A similar approach is taken in Hartmann, Haspelmath & Cysouw 2014 in which the participant roles of different verbs (“microroles”) are compared based on two possible features for each microrole—marking on the argument and indexing on the verb. The similarity between two microroles is given by the number of shared features divided by the maximum number of features for either microrole.

The similarity measure used in this thesis is similar to these. They all can basically be thought of as using the relative Hamming distance between individual codings of items, where the coding is a sequence of elements. We might call this the coding-element comparison application.

As a relative Hamming distance is an average, we can also apply a weighted average, as done by for example Levshina, whose choice of weights is data-driven. The choice can also be a priori motivated, based on the view taken of what it means for contexts to be similar for the purposes of the particular study. In this study of demarcating knowledge within the larger cognition-perception field, lexical elements are more heavily weighted than constructional, capturing the following principle about similarity between contexts:
4. Shared lexical elements imply greater similarity than shared constructional elements.

This is motivated by the research questions concerning categorisation (cf. 3.2). As research on basic level categories (see e.g. Hajibayova 2013) supports, the most salient categories tend to be lexical. Thus, when investigating the categorisations of languages, it is reasonable to weight lexical distinctions more heavily. Note also that weighting lexical roots heavier than other elements is what Levshina (2015: 498) settles on as being “best [...] interpreted from a semantic perspective”.

One further principle is incorporated into the similarity measure used in the demarcation of the knowledge domain:

5. Contexts with no shared lexical elements but shared constructional elements are maximally dissimilar.

This last principle is perhaps less self-explanatory than the others. It means that if two contexts do not share lexical items, they are considered maximally dissimilar, even if they share some constructional elements. The are two motivations for applying this principle in the demarcation of knowledge. The first is that it makes the results more comparable with other studies on colexification, which generally concern shared lexical items. The second reason comes from the coding. Recall that the coding is done mainly based on contrast within the sample and that this means that mutual lack of coding for a given feature is considered similarity (see more on this below). This similarity is of course true in some sense also for contexts with different lexical roots—that two contexts share e.g. an unmarked subject means that they are similar in some sense. This kind of similarity, however, if shared with all contexts—which is the case when not explicitly coded for—is not relevant to partitioning the contexts and only serves to introduce unnecessary noise.

The principles 1-5 are implemented in a measure which is similar to the one used in Levshina (2015). The main difference between the approach taken here and that of Levshina is that it is not the values of predetermined features which are compared, but a set of markers collected through a general procedure (described in section 3.6.2). This means that the coding of contexts does not consists of specified values for a determined number of features, but rather of four parts: the lexical roots making up the predicate, further marking on the predicate (e.g. derivational morphology), marking on the knower and marking on the known. Of these four parts, only the lexical roots are coded for in the majority of cases; as discussed more thoroughly in section 3.6.2, only those markings which contrast are coded for. Lexical roots are missing only in those
relatively rare cases where the translation in a given language does not use a predicating construction, but rather e.g. evidential marking.

These markers then need to be compared across contexts. There are several possible ways of doing this. The simplest would be to consider each marker as a feature of its own, so to speak, and compare absences and presences across contexts. As mentioned in 3.6.2, however, there is a problem with this when computing the similarity of contexts. To illustrate the problem, consider an example. In Swedish, there are particle verbs. A verb is followed by a particle (usually a preposition forms) which gives a new meaning. Ignoring for the sake of convenience the possibility of particles consisting of more than one word and if these then should be analysed as two particles, we have, for instance, the following to compare—känna ‘know (person)’, känn till ’know of’ and känn igen ‘recognise’. The comparison of these constructions is shown in table 5.3. Note that the construction without a particle comes out as more similar to a form with a particle than two forms with different particles do to each other. This is not a desirable result, if we think of the various options (nothing, till and igen in this case) as parts of a paradigm. Then, the unmarked option should be the same distance to the others as those with overt marking. This is where the procedure for sorting markers into paradigm slots that was described in 3.6.2 comes in.

Taking this sorting into slots into account in the computation of similarity can be thought of as moving yet another step before comparing sequences. In the binary method, we compare contexts directly, by representing them as a sequence of single values and comparing these sequences. In the coding-element comparison method, it is not the sequences of values in each context which are directly compared. Rather, each context in each language is represented by a sequence of values, and it is these that are compared. In the procedure employed here, it is not pre-determined elements but rather a set or markers which are sorted into slots based on the procedure described in 3.6.2 which are compared. We might call this the paradigm-adjusted marker-based application of the Hamming distance. This is the similarity measure used in this thesis.

To summarise: if there are no lexical items shared by the two codings, but there are lexical items present in at least one of the codings, the similarity

<table>
<thead>
<tr>
<th>Forms</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>till</td>
</tr>
<tr>
<td>känn till–känna till</td>
<td>0</td>
</tr>
<tr>
<td>känn–känna igen</td>
<td>1</td>
</tr>
<tr>
<td>känn till–känna igen</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5.3: Example of comparison of markers in Swedish
between them is 0, i.e. maximally dissimilar. If neither coding contains any lexical items, the grammatical constructions are compared, equal weight being given to the three subtypes. If there are shared lexical items, the similarity is defined as a weighted mean of the proportions of shared markers, sorted into paradigm slots, with lexical items given the same weight as the other types combined. We may call this version of the measure the *lexically weighted similarity measure*.

The measure may be illustrated with an example. Consider again the following contexts from Hungarian, repeated in Table 5.4.

<table>
<thead>
<tr>
<th>Nr</th>
<th>Verse</th>
<th>Lx</th>
<th>Px</th>
<th>Sx</th>
<th>Ox</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>John 4:32</td>
<td>tud-</td>
<td></td>
<td></td>
<td>DEL0</td>
</tr>
<tr>
<td>2</td>
<td>Acts 23:34</td>
<td>tud-</td>
<td>meg−0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Matt 7:20</td>
<td>ismer-</td>
<td>meg−0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.4: Examples of codings of verses from Hungarian

We find two different lexical roots, *tud-* `know` and *ismer-* `know (person)`. In one context, 1, the object is marked with the Delative case, which contrasts with the Accusative (the default case for direct objects). In two contexts, 2 and 3, we find the prefix *meg-.* No context is coded for subject-construction, since this is invariant across the contexts of the sample.

Beginning with the first context in Table 5.4, the codings of 1 and 2 have their one lexical root in common—resulting in the value 1. There are no forms present for knower marking, giving the value of 1. For known marking, there is a coded marker in context 1 which belongs to slot 0 (marked by a subscript in Table 5.4). Context 2 has no marker belonging to slot 0, and in comparing this slot the value is thus 0. Since there is only one slot in the Known category, this is also the value for that category. The situation is the same, but reversed, for predicate marking; the prefix *meg-* is present in context 2 but not in context 1, also here giving a value of 0. The similarity of 1 and 2 is thus:

\[
similarity(1, 2) = 1 + \frac{0 + 1 + 0}{3} = \frac{2}{2} = 0.67
\]

Moving on to comparing 1 and 2 with 3, we employ principle 5—that is, since the lexical roots are different, the encodings have a similarity of 0, regardless of the shared derivational morpheme *meg-* between 2 and 3.

Note that the situation illustrated by Hungarian above—there only being one paradigm slot for the coded markers in each category—is quite common. This is to a large part an effect of the choices made in what markers to include.
in the coding. A less restrictive approach would likely have resulted in more slots for many languages.

These language-by-language measure of the shared linguistic forms between contexts are then averaged across the language sample—the sum of the comparisons is divided by the number of comparisons made. This means that contexts that for some reason are not comparable (i.e. <empty>, NA and NC values, see section 3.6.2) are not taken into account in calculating the average similarity. Only the languages of the core sample are included in the average similarity; this is to avoid skewing of the sample by including five additional basically Germanic languages. The average similarity of each pair of contexts is then combined into a similarity matrix, showing how similar on average across the sample each pair of verses in the sample is. This similarity matrix may equally well be represented as a distance matrix (by subtracting the similarity from 1), giving the same information but from the opposite point of view.

5.3 Grouping similar verses

With the procedure described in the previous section, we have the first part of what is needed for an answer to the question of the demarcation of a knowledge domain, i.e. a measure of similarity of expression. The second thing needed is a way to group our contexts based on this similarity. There are many methods for this—that is, the clustering of data points based on some distance measure. Previous applications in linguistic have used for example \( k \)-means clustering, specifically \( k \)-medoids (Wälchli 2018) and hierarchical clustering (Levshina 2022) see also van der Klis & Tellings (2022) for a survey of approaches.

Here, hierarchical agglomerative bottom-up clustering as implemented in scikit-learn is used (Pedregosa et al. 2011). This method has the advantages of i) allowing the clustering to be performed immediately on the distance matrix, ii) allowing for clusters of very different size and iii) being deterministic, i.e. giving the same result each time. It functions basically as follows. All data points—in this case contexts—start out as their own clusters. Points and subsequently clusters are then merged with the closest cluster until the desired number of clusters is reached. Note that the method requires specifying the number of clusters for the points to be divided into. This is determined here using what is known as silhouette scores (Pedregosa et al. 2011) [Rousseeuw 1987]. The silhouette score of a clustering is the average of the silhouette coefficient of all data points. The silhouette coefficient of a point is calculated by subtracting the mean distance of it to all points in the same cluster from the mean distance to all points in the nearest other cluster and dividing it by the largest of the two means. The silhouette score varies between -1 and 1, where
-1 indicates faulty clustering (on average, the distance between data points is greater within clusters than between) and 1 clearly defined clusters (the distance between clusters is greater than the distance within clusters) (Rousseeuw 1987).

In estimating the optimal number of clusters, the clustering is performed for all relevant numbers of clusters (generally 2 to \( n-1 \), where \( n \) is the number of data points to be clustered) and the number of clusters with the highest silhouette score is chosen. This represents the number of clusters at which the clustering is most well-defined—i.e. small distances within clusters and relative large distances between clusters (Kaufmann & Rousseeuw 1989, Rousseeuw 1987).

In summary, then, as an answer to the second methodological issue of dividing contexts into group based on similarity raised in the introduction to this chapter, a clustering algorithm—agglomerative hierarchical clustering—is applied, the number of clusters to be distinguished being determined by the silhouette score of the clustering.

5.4 Demarcating the knowledge domain

With this means of dividing the contexts into groups using a clustering algorithm after determining how similar they are using the measure described above, we are in a position to address the concretised research question posed at the beginning of the chapter by applying these methods to the data. The data considered here is the full verse sample—verses representing knowledge predication in addition to near-lying concepts (see 3.5 and appendix B). Their coding in the 78-language core language sample (see chapter 4) is considered. In the results presented below, 99 of the 100 verses in the full verse sample are included.

Applying agglomerative clustering to the distances obtained from the measure described above and comparing the silhouette scores of different numbers of clusters, we get the results shown in figure 5.1.

As we can see, there are two distinct peaks at eight and ten clusters (dashed and dotted line respectively), with eight clusters receiving a somewhat higher silhouette score. This suggests that there are eight clearly distinguishable clusters in the data but that one of them shows considerable within-cluster variation, so much that dividing it into clusters of its own gets an almost as high

---

1 One outlier was omitted on account of a deviantly high number of hapax legomena (forms used only once) vis à vis the verse sample (more than 1/3 of the languages using a hapax legomenon in the verse). The verse is Acts 25:25 “But I understood that he had done nothing deserving death himself [—]”. For this verse many translations use words meaning ‘find’, ‘grasp’ or words for which no meaning other than that of the given verse can be found.

115
Figure 5.1: Silhouette scores against number of clusters. Dashed line at n=8 and dotted line at n=10

silhouette score. We will therefore take both of these clusterings (eight and ten clusters) into account.

Table 5.5 shows the clusters distinguished\footnote{The clustering in full is available in the Online Supplementary Materials} with labels, examples of verses belonging to each cluster and the number of verses in each cluster. The verse with the highest silhouette coefficient in each cluster is given. The labels correspond to the English lexemes frequently used in the clusters and are given as approximate labels of convenience.

With eight clusters—the optimal number going by silhouette score—there is a large cluster which includes contexts which in English use the word *know*—I will refer to this as the extended knowledge domain. With ten clusters, the second best number by silhouette score, this cluster is broken up into three—one corresponding roughly to ‘know’, one to ‘understand’ and one with a somewhat more mixed expression in English but which can be characterised as the ‘find out’-cluster. The ‘know’-cluster at ten clusters will be referred to as the narrow knowledge domain.

5.4.1 Methodological issues of the clustering

Before taking this clustering as the basis for postulating a generalised domain, there are a number of questions that can be asked about it. First of all, to
<table>
<thead>
<tr>
<th>8 clusters</th>
<th>10 clusters</th>
<th>#</th>
<th>Verse example</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNOW</td>
<td>KNOW</td>
<td>36</td>
<td>Mark 1:24 “[—] I know who you are—the Holy One of God!”</td>
</tr>
<tr>
<td></td>
<td>FIND OUT</td>
<td>8</td>
<td>Luke 9:11 But when the crowds found out, [—]</td>
</tr>
<tr>
<td></td>
<td>UNDERSTAND</td>
<td>11</td>
<td>Matt 13:51 “Have you understood all these things?” [—]</td>
</tr>
<tr>
<td>SEE</td>
<td></td>
<td>8</td>
<td>John 5:6 Jesus, when he saw this one lying there [—]</td>
</tr>
<tr>
<td>REMEMBER</td>
<td></td>
<td>9</td>
<td>Mark 11:21 And Peter remembered and said to him, [—]</td>
</tr>
<tr>
<td>BE ABLE</td>
<td></td>
<td>11</td>
<td>Luke 1:22 And when he came out he was not able to speak to them, [—]</td>
</tr>
<tr>
<td>OBEY</td>
<td></td>
<td>4</td>
<td>Col 3:20 Children, obey your parents in everything, [—]</td>
</tr>
<tr>
<td>THINK</td>
<td></td>
<td>3</td>
<td>John 11:13 [—] but they thought that he was speaking about real sleep</td>
</tr>
<tr>
<td>HEAR</td>
<td></td>
<td>4</td>
<td>Matt 13:19 When anyone hears the word about the kingdom [—]</td>
</tr>
<tr>
<td>BELIEVE</td>
<td></td>
<td>5</td>
<td>John 11:27 [—] I have believed that you are the Christ, [—]</td>
</tr>
</tbody>
</table>

**Table 5.5:** Clustering of the verses at 8 and 10 clusters
what extent is it a product of the specific clustering algorithm chosen? Second, to what degree is this clustering an artefact of the initial categories used to choose verses? And finally, what impact does the number of clusters have on the results? These issues are discussed in turn below.

a) The choice of clustering algorithm

Starting with the first question—to what degree the clusters are an effect of the specific method used—it can first be said that in using clustering for exploratory data analysis, we are not working toward a specific result. That is, there is no gold standard to which to compare the clusters produced. Therefore, it cannot really be said that one clustering is better or worse than another (although one clustering may be more useful than another, or more intuitive). What we can say is that if the results of several methods for clustering converge on one and the same result, this result likely reflects important features of the data.

This is the case with the knowledge domain. In addition to agglomerative bottom-up clustering, I applied two other clustering algorithms to the data, namely Partitioning Around Medoids (PAM) and DBSCAN. PAM works by trying to minimise the distance between the points in a cluster and a central point (“medoid”) of that cluster. PAM is a variant of $k$-medoids clustering, which is related to $k$-means clustering, but uses actual data points as central points rather than an abstract centroid. This makes it less sensitive to outliers (Kaufmann & Rousseeuw 1989). $K$-means clustering in general tends to work best with even cluster sizes (Pedregosa et al. 2011). As with agglomerative clustering, the number of clusters needs to be specified; here the number with the highest silhouette score is chosen.

DBSCAN identifies areas with high density of data points surrounded by areas of lower density—the areas with high density end up as clusters. Unlike the other two algorithms, DBSCAN does not take the number of clusters as a pre-set argument, instead taking a value for the maximum distance allowed for two points to be close enough to count as density. Another difference from the other two algorithms is that DBSCAN does not necessarily assign all points to a cluster; points can also be labeled “noise”, i.e. as not clearly belonging to any cluster (Ester et al. 1996, Pedregosa et al. 2011). Both alternative clustering algorithm were implemented using scikit-learn (Pedregosa et al. 2011).

The results of the three clustering algorithms—agglomerative clustering with both eight and ten clusters—are shown in figure 5.2. The focus lies here

---

1Parameters used for DBSCAN were min_samples=3, eps [=maximum distance allowed] = mean($d$) - 2*(st.dev($d$)), where $d$ are all the distances in the input matrix

2To visualise the clustering, the verses are shown on dimensions 1 and 2 of the results of a
on the cluster in each result containing verses from the narrow knowledge domain. The verses included in this cluster are marked in black. The results are not identical, but can be interpreted as there being a set of core ‘know’-verses which are clustered together regardless of algorithm along with more peripheral contexts whose clustering varies. The core verses correspond roughly to what was above labelled the narrow knowledge domain, whereas the extended knowledge domain is more varied (cf. chapter 8).

In short, the results of three quite different clustering algorithms converge on the same core narrow knowledge domain with a more varying periphery, i.e. the extended knowledge domain. Variations of the details of the distance measure used yield similar results.

b) The effect of initial categorisations

As for the question of the effect of the initial categorisation on the results, there can of course be no direct effect—the clustering algorithm has no access to the manual categorisation of the verses, only the similarity measures based on the codings in the languages of the sample. The question is rather if the a priori categories have led to a choice of verses that gives a somehow skewed result. The main purpose of the verse sample is getting at the full variety of knowledge

Classical Multi-Dimensional Scaling analysis—for details, see chapter 7
predication and to investigate the demarcation toward and possible connections with neighbouring concepts. In practice, this means that many more verses were chosen to represent knowledge, whereas other concepts are represented by a smaller number of what is meant to be prototypical cases. There are two main ways in which this might have influenced the results presented here.

First of all, it clearly decreases the chance of finding variation in the, from the perspective of this thesis, peripheral concepts. Even with the relatively small number of contexts considered here, the expression of these peripheral concepts is not uniform. A view of this can be achieved by calculating the average similarity (as described by the measure above) between the verses in clusters (including the in-cluster similarity). The results are given in figure 5.3. The closer the score are to 1, the more alike are the codings in the compared clusters—1 is total uniformity.

The considerably larger KNOW-cluster has the most internal variation, but the other clusters also show internal variation. Some of these domains have also previously been shown to have great internal variation, e.g. perception (Viberg 1984, Wälchli 2016). It is therefore not unlikely that the relative lack of variation here is due to the small number of verses investigated, rather than some characteristic of these concepts. This, however, is not really a problem, as this thesis does not aim to address variation within concepts other than knowledge.
A potentially more relevant problem is the choice of contexts concealing or obscuring relationships between the domains investigated. As mentioned, verses in the peripheral domains were mainly chosen to be prototypical instances. Beyond being connected through straightforward colexification, there are many instances of domains and concepts being related in less clear-cut ways by overlapping at the edges. Whereas the data collected here should be sufficient for discovering relations of the first kind, it is more doubtful if the second kind can be uncovered, due to the sampling of the verses being concentrated to prototypical examples. It might also be noted that using translations by itself makes drawing conclusions about this kind of extension a bit tricky—there is a risk that the use of a given extension is simply carried over from the source language.

There are thus some limits to what conclusions can be drawn from the data at hand. What can be concluded, however, is that in most languages in the sample there is a linguistic differentiation between know and other concepts. That the line between these may be blurred if looking at less prototypical contexts and the divide therefore less sharp has no impact on this result.

c) The number of clusters

There is also the possibility that the number of clusters has an effect on the result. As discussed, the number is not chosen arbitrarily, but by picking the number that gives the highest silhouette score and thus the most well-defined clusters. However, this does not mean that it is this number of clusters that is the “correct” number, and it may therefore be of interest how the clustering looks with different numbers of clusters. Figure 5.4 below shows with coloring how the clustering develops from two to seven clusters, with the eight optimal clusters as a basis. Note that since the algorithm works bottom-up, i.e. from the data points as individual clusters to fewer and fewer cluster by merging them, once a context is assigned to a cluster it is, so to speak, stuck there.

Reducing the number of clusters by one past the optimal, the HEAR and OBEY clusters are merged. Moving to six clusters, REMEMBER and THINK are merged. At five clusters KNOW and SEE are merged while at four clusters, BELIEVE joins the THINK-REMEMBER cluster. At three clusters, the HEAR-OBEY cluster joins the KNOW-SEE clusters and at two clusters, BE ABLE joins this cluster as well, creating a divide between KNOW-SEE-HEAR-OBEY-BE ABLE and REMEMBER-THINK-BELIEVE. In other words, the most basic divide appears to be between what might be termed non-factive cognition and everything else. KNOW remains separate until at five clusters, where it is merged with SEE, after which it at three clusters is merged with HEAR-OBEY. Thus, while the separation of KNOW is quite robust, there are
Figure 5.4: The results of agglomerative clustering with different number of clusters. Nodes correspond to the clusters at $n = 8$, with the labels the same as in table 5.5. Size corresponds to the number of verses. Coloring corresponds to the clustering (in that same color means same cluster). Width of the edges corresponds to the average similarity between verses of that cluster.
clear connections especially to perception.

5.5 Summary and conclusions

In this chapter, I have shown that when applying agglomerative clustering to the similarity measure derived from the coding of the full verse sample, two ways of clustering emerge as almost equally good. We see this in the distinct peaks of the silhouette score at eight and ten clusters. At ten clusters, we have a cluster consisting of contexts roughly corresponding to English *know*, as well as one containing ‘understand’ contexts and one ‘find out’ contexts. At eight clusters these three are merged into a single cluster.

The results of the agglomerative clustering presented here thus show that knowledge can be distinguished as a generalised domain—when considering the categorisation of cognition and perception in a large number of languages, boundaries coalesce around a group of contexts which corresponds to knowledge. In fact, we can speak of two domains: the *extended knowledge domain*, which includes ‘understand’ and ‘find out’, and the *narrow knowledge domain*, where these are excluded. Either solution gives almost equally well defined clusters, with the extended knowledge domain being slightly better. This indicates that whereas there is considerable homogeneity within the extended knowledge domain, there is also considerable variation. It is the extended knowledge domain which will be taken as the domain of investigation in the remainder of this thesis.
6. The universality of ‘know’

6.1 Introduction

In the previous chapter, it was shown that when applying agglomerative clustering to the distance measure derived from the coding of the verses, two clusterings emerged as optimal going by silhouette scores—the extended knowledge domain and the narrow knowledge domain. The former includes meanings in English expressed by know, find out and understand whereas the latter includes only meanings roughly corresponding to know. I argued that these may be regarded as generalised domains—abstractions from the categorisations of individual languages. As such, they can be used to make inferences about knowledge as a conceptual domain. I argued, following de Haan (2010) and Dahl (2016) that their presence in many diverse languages suggests the cognitive salience of a domain.

With this, research question A1 is addressed. Question A2 remains—How universal is a distinct construction for knowledge predication? While it is of interest to consider the categorisations made in languages as generalisations (for example, with the use of clustering algorithms) it is, fundamentally, the individual languages which are the objects of investigation. It is therefore of interest to consider to what degree the individual languages conform to the general picture. Even more so since, as discussed in 2.1, the universality of knowledge is a hotly debated subject where, for instance, one of the leading approaches within semantic typology—Natural Semantic Metalanguage—holds KNOW to be a universal basic prime (Wierzbicka 2018b). What light can the data of investigation shed on the universality of ‘know’?

The chapter is structured as follows. First, section 6.2 discussed the two languages of the sample which may be said to lack a dedicated ‘know’ form and section 6.3 continues with languages where a ‘know’ form can be distinguished, but the lines between the knowledge domain and other domains are nonetheless unclear. Section 6.4 concludes the chapter with an extended discussion of what the data presented can tell us about the universality of ‘know’. The methodology of the chapter is summarised in table 6.1.
Research Question | A2. How universal is a distinct construction for knowledge predication?
Operationalisation | Are there languages in the sample which do not make a distinction between knowledge and one or more of the other concepts included in the verse sample?
Method(s) | Qualitative
Language sample | Full
Verse sample | Full

| Table 6.1: Methodological summary of chapter 6 |

6.2 Languages without a form for ‘know’

As we have seen evidence of, ‘knowing’ and perception are intimately connected linguistically. These connections are found in many languages. Somewhat less commonly, the knowledge domain is also connected to ‘thinking’, ‘remembering’ or modality. In this section, we will be concerned only with a very particular subset of languages with these connections, namely those where there is no form which is exclusively associated with part of the knowledge domain (narrow or extended), that is, languages which may be said to lack a form for ‘know’. I will call these cases absolute colexification, which, based on the definition of colexification given by François (2008: 170), I define as:

Definition 8. A language absolutely colexifies two functionally distinct senses A and B if, for any form associated with A, it is also associated with B and vice versa.

This definition excludes such examples as English, where for instance see can express both the concepts ‘see’ and ‘know/understand’, but the opposite (know expressing ‘see’) is not found.

Only two languages in the sample properly fall under this definition: Kalam (Nuclear Trans-New Guinea, PNG) where ‘know’ is absolutely colexified with perception and other cognition, and Fasu (Isolate, PNG) where ‘know’ is colexified with a range of cognitive meanings, including ‘think’. As we will see below, there are also a number of languages in which the lines between ‘know’ and other domains are blurred, without them falling under the definition of absolute colexification.

Given that the data used here are translations it might be suspected that patterns of colexification or a lack thereof are due to interference from the source language. However, this does not appear to be the case given a comparison...
with a different data set. In Table 6.2, some data on colexification from CLICS are shown (Rzymski et al. 2020). For the relevant pairs of concepts, the number of varieties with forms for both concepts is given, along with how many top-level families (using the Glottolog’s classification, Hammarström et al. 2019) they represent. The number and percentage of languages and families where there is absolute colexification—i.e. where any form with meaning A also has meaning B and vice versa—is then given.

As discussed in chapter 3, the meanings ‘feel’ and ‘taste’ were not included in the investigation of this thesis due to a lack of suitable contexts in the NT, despite the indications that these might be colexified with ‘know’. For the relevant concepts, the figures from CLICS are similar to the results of this thesis. The total number of languages in which any absolute colexification with ‘know’ is found is slightly higher—4% to 2.6% in my data—but this can be explained by the inclusion of more concepts in CLICS as well as the lack of consideration for constructional differences—as we will see, the latter is what distinguishes ‘know’ from ‘be able’ in several languages in the sample. That the numbers are fairly close speaks against interference of translation from a source language as the driving reason for the apparent near-universality of ‘know’ in the results presented here.

We will now consider the situation in those two languages where no distinct form for ‘know’ can be found. In Kalam, the form used for ‘know’ is the very general verb nipt: Pawley & Bulmer (2011: 416) give the gloss “[g]eneric for acts or processes of perception, sensing and cognition” together

\[1\] The numbers in the total row refer to the total number of languages/(families) for which data for at least one colexification pattern is found and in how many of these at least one case of absolute colexification is found.
with the following 22 translation equivalents, noting that “it is not certain that all of these are separate meanings for Kalam speakers” (Pawley & Bulmer 2011: 416):

1. Be conscious, aware.
2. Be awake.
3. Think, reason, cognize.
4. Perceive, sense something.
5. Know something, know how to do something.
6. Understand, comprehend, grasp something mentally.
7. Take notice of something, pay attention to s.th.
8. Realise, become fully aware of a situation, see that something is the case.
9. See something, perceive s.th. with the eyes.
10. Look at, observe something.
11. Hear something, perceive s.th. with the ears.
12. Listen to or for something.
13. Feel something by touching it.
14. Feel emotion (object an internal organ, seat of emotion).
15. Smell something.
16. Taste something.
17. Try or test something.
18. Have thoughts, ideas, opinions.
19. Learn something, acquire knowledge or understanding.
20. Discriminate, tell (in sense of discern), know the difference (between).
21. Be used to, familiar with, have previous knowledge or experience of something.
22. Believe, be under the impression, think that (something is the case).

Pawley & Bulmer (2011: 416)

In addition to these senses, the form niŋ can also mean ‘shine, give off light or heat (especially of the sun)’. According to some speakers, the two forms are cognate, and the sense of the sun shining comes from the sun ‘looking at’ the world (Pawley & Bulmer 2011: 419).

Note that these very broad verbs are a general feature of Kalam, and not limited to the cognition-perception domain. There are about 130 verb roots in Kalam, 15 of which make up almost 90% of the verb tokens in text, and 35 of which make up 98%. Other examples of these very generic verbs are g- ‘happen, act, do, make, work’, ag- ‘make a sound, say’, ay- ‘put, form, stabilise, become’, d- ‘touch, have, control, finish’—these generic verb roots commonly combine with either adjuncts or other verb roots into complex verb constructions (Pawley et al. 2000: 157–159). Examples are given in 44.

1 The orthographies used in the NT and Pawley’s differ in that, among other things, a predictable epenthetic <i> is excluded in Pawley’s orthography. I have here given the examples as they appear in the source. In the running text I will write niŋ- (as in the NT) rather than ny- (as in Pawley’s publication)
(44) Kalam (Nuclear Trans-New Guinea, Papua New Guinea)


\[ Cn \text{ tap kun ak tap tmey ak } n y - b - \text{un} \]
1PL thing such this thing bad this know-PF-1PL

‘We know that this sort of thing is bad’

b. Pawley et al. (2000: 177)

\[ Yad \text{ wdn } n y - b - \text{in} \]
1SG eye perceive-PF-1SG

‘I saw (it)’

c. Pawley et al. (2000: 177)

\[ Yad \text{ kaj kuy } n y - b - \text{in} \]
1SG pig odor perceive-PF-1SG

‘I smell pork/pig’

d. Pawley et al. (2000: 177)

\[ Yad \text{ gos } n y - \text{sp-in} \]
1SG thought perceive-PROG-1SG

‘I am thinking’

If we found a complex verb construction corresponding to ‘know’—basically parallel to the situation in Bauzi, see 50 below—there would be no problem; we would have a form corresponding to the concept of interest. However, no such complex construction meaning ‘know’ can be found (Pawley 1994: 394). Various kinds of perception can be distinguished via complex verb constructions, as can thinking (see the examples in 44), but not ‘know’.

Kalam has in fact been explicitly raised as a challenge to the Natural Semantic Metalanguage claim of ‘know’ as universal (Pawley 1994). The argument is that the verb \( n i i ? \) does not correspond to any of the primes KNOW, THINK or FEEL but to a more general concept subsuming all those meanings. Pawley argues from a general methodological principle of assuming unitary meanings for forms unless there are reasons not to do so. Since all the uses of \( n i i ? \) can be explained by a very general ‘cognise/perceive’ meaning, no polysemy analysis is necessary. Wierzbicka (1994), however, does not agree with Pawley’s analysis (see also Dessai (2016) for a development of these points).

\[^1\text{At the time of publication, SEE and HEAR were not yet suggested as semantic primes, but the argument can extend also to these meanings}\]
Wierzbicka argues that since there are complex forms corresponding to all uses of *niŋ*- except ‘know’, these complex forms can be assigned as the exponents of e.g. *THINK* and *FEEL* (and *SEE* and *HEAR*, continuing along Wierzbicka’s line of argument regarding these newer primes) and *niŋ-* on its own as an exponent of *KNOW*.

To accept Wierzbicka’s argument, we need to accept either that 1) when *niŋ-* occurs in the complex forms for ‘see’, ‘think’ etc. it does not mean ‘know’, or 2) that the meaning of ‘see’, ‘think’ etc. includes ‘know’. Wierzbicka argues for the first position by claiming that we can distinguish between compositional relationships and “‘recurring partials’ linked by an unanalysable family resemblance” (Wierzbicka 1994: 455). Essentially, as I understand her argument, formal compositionality need not correspond to semantic compositionality (cf. the discussion of isomorphism in chapter 8). That forms can lose their meaning in fossilising constructions is well known under the label desemanticisation in grammaticalisation theory (see e.g. Heine 2003). However, the fact that something can be the case does not mean that it is the case, and it remains to be shown—by native speaker judgement or experimental work—that the complex forms containing *niŋ-* are not understood by speakers as compositional before this analysis can be accepted.

It should be noted that in the verse sample, ‘think’ is expressed with the verb *ag*- ‘say’ and is thus not colexified with ‘know’. This possibility is also discussed in Pawley (1994), as in the examples in [45] below. The use of ‘say’-constructions to express thought or internal speech is a well-known phenomenon, see e.g. Sprenck & Casartelli (2021).

(45) Kalam (Nuclear Trans-New Guinea, Papua New Guinea)

a. **Pawley** (1994: 395)

“*Mb gobt*” *agt* *gos ak niŋ*
such we.do they.said(wondered) thought the having.perceived

*agt*, “*kapkap ŋluk  ꜀ mdut,*” *agl*... they.said(decided) secretly hiding put we.stay having.said(decided)

‘They wondered what to do, then having though it over they decided to conceal themselves...’
b. Pawley (1994: 396)

Yad **apin** “b-tud maj ma-ñbay”
I said(thought) men-white sweet:potato not-they.eat

**apin**
I said(thought)

‘I had always thought that white people don’t eat sweet potatoes.’

Moving on to the other language in the sample where there appears to be a lack of distinction of ‘know’, we remain in Papua New Guinea. In Fasu (Isolate), we find **hemakapu**- ‘know, think, love, hope, want, etc.’ used for a wide range of cognitive meanings, though we find separate forms for both ‘hear’ and ‘see’. The form **hemakapu**- might be complex; there is a verb **pu**- ‘walk’ which according to the grammar gives a meaning ‘of its own accord’ to compounds with it (Loeweke & May 1980: 45). No form like **hemaka**, however, is found in the descriptive sources (Loeweke & May 1980, May & Loeweke 1981, Loeweke & May 1966). Examples are given in 46. In 46a the ‘know’ use is illustrated. In 46b we can see, in the same clause, both the nonfactive use, here from the context most readily interpreted as something like ‘want, hope’ and the ‘know/understand’ sense. In 46c we see illustrated clearly that **hemakapu**- can have a false belief meaning.

(46) Fasu (Isolate, Papua New Guinea)

a. Mark 12:14

*Aiyaka isiapo hemakapu-raka-ri-po*
saying.like.thus 1PL.ERG cognise-NEUTR-STAT-IND

‘And like this we know [this]’ (AS)

b. 2 Cor 1:13

*Yumo nōmo hemakapu-sua-ne repo wakasema*
now 1SG.ERG cognise-PST-? 2PL.ERG little

**hemakapu-kua, keseke repo koteaka su-ane hemakapu-simo**
cognise-SEQ later 2PL.ERG well all-ACC cognise-PURPOSE

*nōmo hemakapu-raka-ri-po*
1SG.ERG cognise-NEUTR-STAT-IND

*Lit.: ‘I think (?know, hope, like) that you now understand a little, and I hope that you will understand fully’ (AS)*

‘But I hope that you will understand completely just as you have also understood us in part’
c. John 11:13

Ayaka Rásarase nena makata takamasa
saying.like.thus PR.NAME ? be.sick

ware-kena-paka-siki hemakapu-kua Rásarase ipi kakaro
lie-PRS-lest-? cognise-SEQ PR.NAME 3REFL truly

kusu-ane Yasu-no takumeaka some-raka-no
death-ACC Jesus-ERG code.speech speak-NEUTR-RESULT

ipu himete-sa-po
3PL.ERG not.know-PST-IND

Lit.: ‘(It being said like this), they thought Lazarus was ?sleeping, but Jesus was speaking about Lazarus’s death and they did not know’ (AS)
‘Now Jesus had been speaking about his death, but they thought that he was speaking about real sleep’

There do not appear to be any complex constructions involving hemakapu- corresponding to more specific meanings, which means that the solution offered for Kalam by Wierzbicka is not possible in Fasu. Instead, an analysis that preserves ‘know’ as a category in Fasu would need to show that hemakapu-is polysemous between several cognitive meanings, including ‘know’. As was the case in Kalam, this would require in-depth analysis of language data and perhaps new elicitations and judgements.

Note that in the final example (46c) we see a lexical negation form, himete- ‘not know’. As a matter of fact, a similar lexical negation form is found in Kalam: sakiy. Given this commonality, a question worth raising is whether perhaps these two languages do not distinguish ‘know’ in the positive, but do so in the negative.

For Kalam, this analysis seems doubtful. The translations given of the adjective sakiy are ‘lacking awareness, ignorant; lacking knowledge, stupid, mindless, deaf, unable to hear, mute, lacking speech’ as well as several other related meanings (Pawley & Bulmer 2011: 485–486) The form can be used with the verb ay- ‘put’ to form an impersonal verb meaning ‘be unthinking/ignorant, be deaf/mute, be crazy’ or with the verb g- ‘do’ to mean ‘forget, behave stupidly/recklessly, be unaware’ (Pawley & Bulmer 2011: 486). As a second meaning, Pawley 2011: 486 gives ‘scarce, rare, unusual’. It is possible that the two meanings are related. In the verse sample, sakiy occurs only once, with the meaning ‘not remember, not know’ (47a), but other meanings are found in the NT (47b). See also 47c for an example from Pawley et al. 2000.
Based on the translation given in Pawley & Bulmer (2011) and the examples in 47a-47c, it seems that sakiy is better characterised as a negation of the general meaning of niŋ rather than specifically of ‘know’. This appears also to be Pawley’s analysis, who writes that sakiy stands “in contrast to niŋep” (Pawley & Bulmer 2011: 484), where niŋep is the adjectival form of niŋ-.

For Fasu, the situation is different. As mentioned, the form found here is himete- and is verbal. The dictionary of Fasu gives the translation ‘ignorant of something, not knowing, not understanding’ (May & Loeweke 1981: 10). This seems to be borne out in the NT.

However, as we will see in section 9.6, it is not clear that ‘not know’ is the symmetrical counterpart of ‘know’. Forms meaning ‘not know’ frequently also express meanings like ‘not remember’ or even more general ones like ‘be unaware’. This might be explained by the fact that in the negative, these concepts which are differentiated in the positive, become indistinguishable—the important thing about ‘not knowing’ is a lack of presence in the mind, which is shared by nearby concepts such as ‘not remembering’. Thus, it may be premature to conclude that a ‘know’ category exists in Fasu based on the presence of a form apparently meaning ‘not know’. To draw this conclusion, we would need to show that himete- means ‘not know’ specifically, and is not a general predicate of non-awareness.
Kalam and Fasu, then, are the two languages in the sample which show absolute colexification between ‘know’ and a non-knowledge domain. There are, however, several languages where the connection between ‘know’ and some other domain is strong without there being absolute colexification; the next section presents examples of this.

6.3 Blurred lines between ‘know’ and other concepts

One case of a particularly strong connection between ‘know’ and other domains without absolute colexification is languages where some form within the knowledge domain (e.g. expressing understanding or realisation) is also used in another domain, but there is also a distinct word for ‘know’. Another example is languages where forms for perception extend into the knowledge domain, but coexist with a form specific to ‘know’. An example of the latter is Daga (Dagan, PNG) where the verb for ‘hear’ anu- (48a) is also used for ‘know’ (48b) but there is also an additional form with the noun asigi- ‘knowledge’ which is specific to knowing (48c). Note that anu- in various constructions also extends to ‘remember’—this is not the case for asigi-.

(48) Daga (Dagan, Papua New Guinea)


\[\text{sinao tarep doak ame anu-n}\]
\[\text{drum dance big DEM hear/know-PST}\]
‘he heard music and dancing’

b. Matt 25:26

\[\text{ge anu-ana ne nawan uon}\]
\[\text{2SG hear/know-PST 1SG my.work NEG}\]
‘You knew that I do not work’ (AS)

c. Mark 12:14

\[\text{nu asigi-nu-p ge waenapan muk anen}\]
\[\text{1PL knowledge-1PL.POSS-LOC 2SG person 3PL.EMPH what}\]
\[\text{wane taine ame menan ge imuga potaga ya wariniwa}\]
\[\text{say do that therefore 2SG reject}\]
‘We know that you do not care what people say’ (AS)
Languages like Daga cannot be used as counter-evidence to the universality of a ‘know’-category, any more than a language such as English, where perception can also be used for knowing. A separate form suggests a separate concept, even if that form is not the only possible way of expressing that concept.

As a related but different example, consider Trinitario-Javeriano-Loretano (Arawakan, Bolivia). Here, one form for ‘know’ -echo- (49a) is also used for ‘remember’ (49b). However, the other form for ‘know’, -imoti-, is not extended to ‘remember’. What separates this from Daga is that there is a clear difference in meaning between -echo- and -imoti-, where -echo- means (roughly) ‘know that’ and -imoti- (roughly) ‘know person/object’ (49c). Thus, only a part of the knowledge domain is colexified with ‘remember’. Additionally, there is a form -emtisco- ‘forget’ which is never used outside ‘remember/forget’ contexts and several time appears negated for ‘remember’. It therefore seems that a distinction between ‘know’ and ‘remember’ can be maintained in Trinitario-Javeriano-Loretano.

(49) Trinitario-Javeriano-Loretano (Arawakan, Bolivia)

a. John 6:6

   tjicho  ema  mu-echo-ripo  to  ma-jich-yore
   because  3SG.M  3SG.M-know-IAM  ART/NMLZ  3SG.M-do-FUT

   ‘because he knew what he was going to do’

b. Mark 11:21

   Ma  Pedro  mu-écho-po  ma-jich-po
   and  PR.NAME  3SG.M-know-PFV  3SG.M-say-PFV

   ‘And Peter remembered and said...’

c. John 7:29

   Nuti  n-imoti  ema
   1SG  1SG-know  3SG.M

   ‘I know him’

Additionally, languages where the form for ‘know’ is related to but distinct from a form from another domain do not fall under the definition of absolute colexification. As an example of this kind, consider Bauzi (Geelvink Bay, Indonesia) where the form for ‘know’ is ozobohu- (50a), clearly related to ozo- ‘think’ (50b, 50c). The -bohu is a form meaning ‘far’ and perhaps also ‘long’. The root ozo- is also used in contexts of ‘remember’, both with and without
However, it appears that ‘remember’ can be distinguished from ‘know’ by the use of the pre-verbal particles *vi* (a directional morpheme) and/or *fa* (with iterative meaning), see 50d.

(50) Bauzi (Geelvink Bay, Indonesia)

a. Mark 12:14

> Guru *oa, Oho lab meedam bak lam*  
> Teacher VOC 2SG.ERG that do-CONT-IRR NMLZ that.ABS  
> *im ahebu ozobohu bak*  
> 1PL.ABS all know DECL  

‘Teacher, we know that you do this.’ (AS)

b. **Briley** (1997: 122)

> “Labi *i-ho aii-me ozo, ‘O-m ote-he-da-t*  
> CONJ 1PL-ERG hear-SA think 2SG-ABS kill-REAL-PTS-ERG  
> *esu-hu bak la-ame?’ laha-m bak*  
> put.down.SA-REAL place there-RHQ like.that-IRR NOM  
> *ozobohu-de-m bak am tame*, laha-me ab  
> know-STATIVE-IRR NOM INDIC EMP like.that-SA INDIC  
> *bu-dda-m-am.* call.out-CONT-IRR-INDIC  

"Then we will hear it and think, ‘Oh is that the place where your murderer lives?’ that will be the way that we will know", like that they were calling out.

c. **Briley** (1997: 14)

> Labi Vadu-hat *ozo-ha*  
> CON Vadu-ERG think-DA  

‘When Vadu considered [the situation]’

d. Mark 11:21

> Làhàmu *Petrus-at vi ozobohu-di Yesus bake ab*  
> Therefore PR.NAME-ERG PRT know-SA PR.NAME to IND  
> gago-h-am  
> say-REAL-IND  

‘Peter remembered and said to Jesus’ (AS)
These languages also are not a challenge to the universality of ‘know’ although they might be taken as an argument against the view that it is primitive (as espoused by the Natural Semantic Metalanguage group, see chapters 2 and 3.3).

We will consider one more type of language where the lines between ‘know’ and some other domain are somewhat blurred, namely the type where ‘know’ shares lexical expression with some other domain but differs constructionally. This is the case in a number of languages where the same form is used for ‘know’ and ‘be able’ but with different constructions. As an example, consider North Tanna, where the form for ‘know’ -itun is also used for ‘be able to’, and interestingly, the form for ‘not know’ -oruruis used for ‘not be able’. Thus, we find colexification both in the positive and negative. However, when used for the ‘know’ meaning, the complementiser m@m@m is used and the complement is fully finite whereas when the meaning is ‘be able’, the complement is nominalised. It is therefore possible in languages like North Tanna to maintain a distinction between ‘know that’ and ‘be able’, although ‘know how’ and ‘be able’ must be seen as coexpressed.

(51) North Tanna (Austronesian>Malayo-Polynesian, Vanuatu)

a. Mark 12:14

I@tgatun, e-k-ot-itun m@w@ ik et@w m@uati@p
teacher 1PL.EXCL-NPST-PL-know COMP 2SG man straight
kon m@w@gin@n
and ?

‘Teacher, we know that you are truthful’

b. Mark 1:40

Ok@m@ n-ak-o@lkeike, kon n-@k-itun n-ol-wir-@n
If 2-?NPST-want CONJ 2-NPST-know NMLZ-make-good-NMLZ
io
1SG

‘If you are willing, you are able to make me clean’

1The languages are North Tanna, Culina, Hungarian, Uduk, Sanumá and perhaps Kotiria
c. Matt 7:18

\[ Nig \text{ \textit{itw}m \textit{t-wir} \textit{okol} \textit{t-\textit{oe}}-\textit{in-\textit{on}} \textit{noan} \text{ tree \ REL \ 3SG-good \ able \ 3SG-NEG-bear.fruit-TR-NEG \ fruit} \]

\[ \textit{itw}m \textit{t-\textit{rat}} \text{ REL \ 3SG-bad} \]

‘A good tree is not able to produce bad fruit’

Note also that some of the languages mentioned here appear to have an additional way of expressing ability which differs from ‘know’ — in North Tanna we find \textit{okol} for external ability and ability of non-humans. Such forms might be evidence in addition to the constructional differences for a distinction between ‘know’ and ‘be able’, but more needs to be known about the interaction of the different modality expressions.

6.4 Summary and conclusions

To summarise what has been said in this chapter regarding the universality of ‘know’ as a category, we noted that only two languages in the sample fulfil the criteria for absolute colexification between ‘know’ and some non-knowledge domain. This number is on a par with results from the CLICS database, indicating that it is not primarily an effect of the data of this thesis being translations. The two languages where ‘know’ is absolutely colexified with other categories are both found in Papua New Guinea; Kalam (Nuclear Trans-New Guinea) and Fasu (Isolate). In Kalam, a very general verb \textit{ni\textit{j}}- with the approximate meaning ‘perceive, be(come) aware of, think’ is found; very general verb meanings are not exclusive to \textit{ni\textit{j}}-, but represent a general feature of Kalam. To complicate the situation, \textit{ni\textit{j}}- may occur with adjuncts or in multi-verb constructions to specify the meaning, e.g. \textit{wdn ni\textit{j}}- ‘eye perceive’ = ‘see’. However, no such complex construction is found for ‘know’.

In Fasu, we find a verb \textit{hemakapu}- with a wide range of cognitive meanings, such as ‘think’, ‘hope’, ‘know’ etc. Unlike in Kalam, this colexification does not extend to perception.

In both languages, we find forms extending to the negation of ‘know’. In Kalam, the form is the adjective \textit{sakiy} and based on its use in the NT and the descriptions and examples given in Pawley (1994), it seems to be best characterised as a negation of \textit{ni\textit{j}}-; that is, as being the negative counterpart of the same very broad concept expressed by \textit{ni\textit{j}}-. The form in Faasu (\textit{himete}-) does not appear to have as wide a meaning as the Kalam form.

We saw that Kalam has been explicitly raised as a challenge to the universality of ‘know’, particularly in the form espoused by the Natural Semantic
Metalanguage—I add to this the data from Fasu. While these languages are somewhat problematic for NSM, it is not clear that they are counterexamples—see above for how the Kalam data has been addressed by proponents of NSM.

Whether the lack of a distinct ‘know’ form in these languages is indicative of a lack of a mental category ‘know’ for their speakers remains an open question.

The research question which led us to consider these languages was how universal the existence of a form for ‘know’ is (research question A2).

From the results presented here, we can make two observations that may help to answer this question. The first is that the existence of a word roughly corresponding to ‘know’ is overwhelmingly more common than the absence; only two examples of absolute colexification between ‘know’ and other domains are found in the sample, which is on a par with the values from CLICS. This tells us that regardless of whether it is completely universal or not, the category of knowing is one that is highly salient across cultures and languages.

This ubiquity is particularly interesting as knowing has no obvious correlate in the physical world—unlike, say, seeing, whose near-universality can be explained by the fact that across cultures and languages, people have the physical sense of sight—and yet ‘know’ appears to recur as a category almost everywhere. These results would also seem to lend support to the NSM claim of KNOW’s universality. In short, with regard to NSM, it can be concluded that the data presented here neither confirm nor deny the claim about KNOW as a universal; although it is challenged by examples like Kalam and Fasu, it is undoubtedly supported by the general pattern in the data.

The second observation is that the languages in which absolute colexification was found—as well as many languages where there is no absolute colexification but the connection between ‘know’ and some other domain is uncommonly close, e.g. 48 and 50—are spoken on New Guinea. This suggests that the lack or fuzziness of the knowledge category might be an areal feature, perhaps connected in some way to cultural practices or discourse tendencies of the area (see e.g. [Evans & Fedden forthc.] for an argument that a tendency to distribute information over smaller units has influenced discourse and grammar in the area).

It should be acknowledged that the identification of a ‘know’ category here rests on extension and translational equivalence. It cannot from this be concluded that the ‘know’ categories are intensionally the same. We are confronted here once again with the issue of extension vs. intension or etic vs. emic meaning, discussed in more detail in chapter 3.3. As discussed in section 2, there have been some that studies experimentally attempt to assess whether the concept ‘know’ is the same across cultures. Wierzbicka 2018, holds that the concept of ‘know’ is universal, but that cultural and pragmatic factors mean
that the appropriateness of applying it in various situations can vary—but this variation does not indicate that the meaning of ‘know’ itself varies.

Essentially, these issues can be boiled down to two questions:

I. Does the form for ‘know’ vary across cultures and languages as regards situations in which it can appropriately be applied?

II. If it exists, does this variation come from differences in the ‘know’-concept? If so, what are these differences?

It would perhaps be more surprising if the answer to question turned out to be negative. Wierzbicka (2018) gives the example of Martha’s statement to Jesus in the Gospel of John about her brother: “I know he will rise, in the resurrection on the last day” (John 11:23–24). As Wierzbicka points out, the use of ‘know’ in this context is probably not in line with modern Westerners’ idea of when we know things—Martha’s conviction is religious belief rather than knowledge. As another example, consider the experimental results presented in Machery et al. (2018), where the authors demonstrate that judgements on what counts as knowing differ between participants from Brazil, Japan, India, and the United States. Not only are there most likely differences between languages, there are also differences within languages, for example across discourse styles. ‘Know’ may be acceptable in persuasive discourse where it would not be otherwise, as evidenced in the following quote from a presidential inauguration; “We know our economy is stronger when our wives, our mothers, our daughters can live their lives free from discrimination in the workplace” (Kashiha 2022: 96).

These differences in the outer limits of the ‘know’ category are not particularly well suited to investigation using parallel corpora. Although it is possible for translators to depart from the source text and not use the form for ‘know’ in the target when it is used in the source or, vice versa, to use the form for ‘know’ when it is not used in the source text, it is intuitively clear that the use in the target text is strongly influenced by the source text. To illustrate the point, we can consider the more concrete meaning of ‘large in size’. Presumably, it varies across cultures and languages to what sizes precisely this description will be applied. Consider, for instance, a tall man. Whether someone is appropriately described as a tall man will presumably vary between Sweden (where the average male height is 182 cm) and Japan (where it is 172 cm). But when translating a Japanese text and encountering the Japanese equivalent of tall man, a translator into Swedish will probably generally not try to determine whether the man in question fits the Swedish criteria for being tall, but will apply the Japanese categorisation in their translation.

For this reason, it would be a mistake to look at parallel corpora, note that everyone who is described as the translation equivalent of tall in Japanese is
also described as lång ‘tall’ in Swedish and conclude that the criteria for these are identical. Similarly, it would be a mistake to try to chart the outer limits of ‘know’ categories using parallel corpora. Question I, therefore, remains largely unaddressed in this thesis and so does, consequently, the follow-up question II.

What I will focus on in the following chapters is rather differentiation within the knowledge domain. It can of course be asked whether this is not equally sensitive to influence from a source language. I do not believe that this is the case. To see why, we return to the meaning ‘large in size’. I have argued that parallel corpora would not be very good data for determining what in a culture counts as large since the target will probably simply apply the categorisation of the source. But once this categorisation is applied, that is, it has been decided that object X is large, it remains for the translator to choose the appropriate form within the ‘large in size’ domain. This choice is constrained by the grammar and the collocational patterns of the target language, and is thus less susceptible to interference from the source text. Thus, in Swedish, a person can be lång ‘tall’ but a building—unlike in English—cannot; it is hög. Similarly, in Swedish you vet a fact, but känner a person. It would be a mistake to think that these domain-internal differentiations are completely impervious to translational interference from a source text or other peculiarities of translations and there are thus no doubt details and variations within languages that require further attention in non-translated language from different genres, but parallel corpora can give a good first bird’s eye view of the knowledge domain.

It is to this bird’s eye view we turn in the following chapters.
7. The dimensions of knowledge

7.1 Introduction

In chapter 5, it was shown that languages in general separate out a group of contexts that contains meanings corresponding to English know. I argued that two generalised domains can be distinguished—the extended knowledge domain and the narrow knowledge domain. There is, however, considerable variation within the knowledge domains; as was seen, the extended knowledge domain contains both the narrow knowledge domain as well as the relatively well-defined clusters UNDERSTAND and FIND OUT. Not even the narrow knowledge domain, though, is homogeneous, and we have already seen examples of lexical distinctions made within this domain in chapter 2—e.g. Swedish veta ‘know (that)’ vs. känna ‘know (person)’ or the distinction in Russian between znat ‘know that’ and umet ‘know how’.

It is to this internal variation we turn in this and following chapters, addressing research question B:

B. What linguistic distinctions are made in the knowledge domain—what meanings are distinguished?

This chapter gives an overview of the internal structure of the extended knowledge domain as investigated by means of Classical Multi-Dimensional Scaling, a picture which is then developed in subsequent chapters. In section 7.2 network and proximity semantic maps are introduced and the advantages of the latter for an investigation such as this are discussed. I also describe the specific method used for constructing proximity maps in this thesis—Classical Multi-Dimensional Scaling—and motivate the choice of this method. Section 7.3 discusses the choice of the number of dimensions to be considered and sections 7.4-7.8 discuss the selected six dimensions in turn. Section 7.9 summarises and concludes the chapter. A methodological summary of the chapter is given in table 7.1.
7.2 From distances to semantic maps

In chapter 5, it was shown how the data collected for each verse can be represented by distances between the verses. These distances are very useful for some applications, such as the clustering applied in that chapter. However, a matrix of distances is not very informative to a human being who wants to know how some points are situated in relation to each other. Imagine someone who is interested in the geography of Europe. A matrix of distances between cities will be of limited use to them. Clearly, an ordinary map with the cities represented in their respective locations would be much better. It is clear how the relationship between the map and the distances looks—we simply measure the distance between two points, either physically based on some coordinate system. It is perhaps less intuitively clear how to make the reverse move—from distances to points in space—but a number of methods allow for precisely this.

Applying these methods we can move from a set of distances, for example the driving distances between selected European cities (table 7.2), to points in space (figure 7.1). The positions of the points in figure 7.1—produced entirely from the distances between them—are very similar to the actual locations, even with the distances being driving distances and not straight distances.

The basic idea of using these methods for analysis of linguistic data is that just as the location of cities can be represented in a two-dimensional space, linguistic variation can be modelled in a multidimensional space. This idea is what lies behind the use of semantic maps to represent linguistic variation. In a traditional semantic map, functions or meanings are the nodes in a network laid out in space, with the edges representing potential co-expression patterns.

---

1 Data from [Map Crow Info n.d.](Map Crow Info n.d.)

2 The x- and y-axes represent the first and second dimensions of the CMDS analysis respectively and have been flipped to align with north-south and west-east.
Table 7.2: Driving distances (km) between selected European cities (excerpt)

<table>
<thead>
<tr>
<th></th>
<th>Barcelona</th>
<th>Belgrade</th>
<th>Berlin</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barcelona</td>
<td>0</td>
<td>1528.13</td>
<td>1497.61</td>
<td>...</td>
</tr>
<tr>
<td>Belgrade</td>
<td>1528.13</td>
<td>0</td>
<td>999.25</td>
<td>...</td>
</tr>
<tr>
<td>Berlin</td>
<td>1497.61</td>
<td>999.25</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Figure 7.1: The first two dimensions of the results of an CMDS analysis of the driving distances between selected European cities
The meaning of a linguistic form can then be given as the extension in this space, i.e. which functions in the network are covered by it and which are not. For an overview of the use of semantic maps, see Georgakopoulos & Polis (2018).

The underlying space in which the extension of language specific forms is represented remains the same regardless of language, allowing for cross-linguistic comparison. There is disagreement on whether this space represents a cognitive reality or not. For example, Croft (2001: 105) argues that the space of semantic maps "represents a universal structure of conceptual knowledge for communication in human beings". Others criticise this view, for example Cristofaro (2010) who argues that there are other ways for multi-functionality patterns to arise than the similarity assumption underlying semantic maps (that co-expression implies conceptual similarity), for instance through historical coincidence. Cysouw (2007: 228) stresses the role of a semantic map as a "model of attested variation", remaining agnostic regarding the relation to a possible conceptual space.

These early semantics maps—made up of nodes and edges—have retroactively been called classical semantic maps (also implicational and connectivity, see Georgakopoulos 2019). I will refer to them as network maps:

Definition 9. A network semantic map is a semantic map which represents meanings and their coexpression in a network graph structure (consisting of nodes and edges)

There is no doubt that network maps have been hugely useful in linguistic typology. However, we face important analytical and methodological challenges in the construction of network maps, some of which are:

1. The node problem—how do we decide what the nodes in our maps are? We cannot know a priori what the primitives of a domain are—introspection, no matter how rigorous, inevitably runs the risk of being biased by the linguist’s native language(s) or languages of expertise.

2. The co-expression problem—assuming we can decide on a set of primitives, how do we decide on what counts as being co-expressed? To give an example relevant to the topic of this thesis, in many European languages ‘see’ may in some contexts be used for ‘know’, but we would hardly say that this means that ‘see’ and ‘know’ are co-expressed in these languages, for most purposes. But how do we draw the line?

1Though note that what is usually cited as the earliest example of a semantic map in linguistics, Anderson (1982) did not have a network structure
3. The network structure problem—how do we go from a set of co-expression patterns to a network structure representing this? Traditional semantic maps are implicational in the sense that edges between two nodes should only be added if their co-expression cannot be explained by other co-expressions and edges. How can this structure be reliably inferred?

In part as a response to these problems, network maps have lately been supplemented by proximity maps (also known as statistical or probabilistic), exemplified in works like Levinson et al. (2003), Croft & Poole (2008) and Wälchli & Cysouw (2012)

Whereas network maps are generally constructed manually—though recently methods for automatic creation have been developed; Regier, Khetarpal & Majid (2013)—proximity maps are produced through an appropriate statistical method applied to the data, with the following solutions to the three problems identified above:

1’ The node problem is solved by taking an exemplar approach (Wälchli & Cysouw 2012). No categories or conceptual primitives are given—instead every exemplar (in the case of this thesis a Bible verse in a language) is treated as an analytical primitive and as a data point or “node” in the semantic map.

2’ The solution to the co-expression problem is made possible by the exemplar approach—binary co-expression of abstract categories is replaced by the co-expression of concrete exemplars. Often this is represented as a similarity measure of the kind described in chapter 5.

3’ The network structure problem currently cannot be solved automatically (see Angluin, Aspnes & Reyzin 2010)—although a way to automatically approximate a solution has been put forth (Regier, Khetarpal & Majid 2013). Proximity maps get around this problem by doing away with the network structure of traditional maps and instead treating the semantic space as continuous in which objects may be more or less close to each other—corresponding to more or less similar.

Given a way of measuring the distance between data points, the task of constructing a proximity like this is identical to the task described above of going from distances to locations. Just as we can go from distances between cities to their approximate locations on a map, we can move from distances between data points to their relative locations in a semantic space.

1Since, as we see in point 3 below, proximity maps abandon the traditional network structure there are properly no nodes, just data points
In short then, network semantic maps typically display the abstracted extension of forms across a set of abstract functions located in a network-structured semantic space whereas proximity maps reflect the actual extension of forms in an actual data-set across a set of analytical primitives in a continuous semantic space constructed mathematically.

Definition 10. A **proximity semantic map** is a semantic map which represents meanings and their coexpression in a continuous space.

Just as there is disagreement about what the space of traditional semantic maps represents, it can of course be asked if a statistical procedure can really construct the semantic dimensions underlying a given domain. Most researchers acknowledge that this is unrealistic and, as in the quote by Cysouw (2007: 105), stress that the results reached are simply a model of the data. Thus, Wälchli & Cysouw (2012: 679) consider the space constructed by an application of CMDS to be a **probability space** in which the positions of contexts reflect the probability of them being expressed similarly. Semantic similarity is one of the factors affecting this, but there are others, like historical accident. It is up to the researcher to describe the probabilities’ connection with potential underlying semantic dimensions.

Note also that just as in network semantic maps, the positions of the data points in the probability space are the same regardless of language.

As mentioned, there are several methods that may be used for the construction of proximity maps. They vary in how they function, what kind of input they take and what the results optimise. Here what is known as Classical Multi-Dimensional Scaling, Torgerson scaling or Principal Co-Ordinate Analysis is used, implemented by the R function `cmdscale` (R Core Team 2021). This is the most basic form of metric MDS and works by applying matrix algebra. A symmetric matrix is always decomposable into what is known as eigenvalues and eigenvectors. An eigenvector of a matrix A is a vector which, when multiplied by some number—known as the eigenvalue—equals the matrix A multiplied by the eigenvector. For a distance matrix where the distances are Euclidean (i.e. straight lines between points) the values of the eigenvectors will perfectly correspond to the coordinates underlying those distances. If the distances are not Euclidean, the result is not a perfect fit and there are negative eigenvalues. In Classical MDS, these are simply ignored as errors, which is usually considered unproblematic if they are small enough. For a description of the mathematics behind CMDS, see e.g. van der Klis & Tellings (2022) or Borg & Groenen (2005 chapter 12).

---

1 There is no necessary connection between exemplar data and representation in continuous space or abstract functions and networks—exemplar data can also be represented in a network and abstract functions in continuous space (see e.g. Croft & Poole 2008).
The choice of Classical MDS rather than some other algorithm has a number of motivations. The first is that, like Principal Component Analysis to which it is under certain conditions identical (Timm 2002: 544), the dimensions of a CMDS analysis pick out the directions along which there is the most variation. That is, in a CMDS analysis, the first dimension positions the data points in such a way that as much of the input distances as possible are accounted for. The second dimension then tries to account for as much as possible of the remaining un-accounted for distances, and so on. The eigenvalue associated with the dimensions corresponds to how informative it is and can be used to choose an appropriate number of dimensions for analysis.

These results are somewhat different than results from many other algorithms, such as the non-parametric unfolding proposed by Poole (2000) and applied to linguistic data in Croft & Poole (2008) or UMAP (McInnes et al. 2018), applied to linguistic data in e.g. Georgakopoulos et al. (2022)—in that the goal is not to as accurately as possibly represent the original distances in as few dimensions as possible, but to accurately represent the variation along the most meaningful dimensions (see e.g. Timm 2002, Rencher & Christensen 2012). For an exploratory investigation of a domain where we expect there to be many relevant dimension, this is a suitable approach.

There are also other advantages of CMDS. Unlike many methods, it is non-iterative and deterministic, meaning that it yields the same result every time. It takes a distance matrix as input, meaning that a variant of the same distance measure as was described in chapter 5 can be used.

For visualization reasons, most traditional semantic maps are two- or at most three-dimensional. In constructing a proximity map, the researcher typically chooses the number of dimensions to consider. In some methods, this choice can have far-reaching consequences in that, as mentioned, the goal is to represent the original distances as faithfully as possible in the given number of dimensions. Classical MDS, on the the other hand, does not have this goal, and the choice of dimensions is merely a question of which dimensions to take into consideration (the difference can be thought of as similar to trying to fit a five-minute song into three minutes by skipping notes here and there or by just cutting it off after three minutes). In CMDS, the analysis of an $n \times n$ distance matrix results in a maximum of $n-1$ dimensions. These remain invariant regardless how many dimensions the researcher chooses to actually consider.

---

1Uniform Manifold Approximation and Projection
7.3 The dimensions of knowledge: Preliminaries

In the following, Classical MDS is applied to the similarity data (based on the 78 core language sample; as with the clustering in chapter 5, we want to avoid the bias of five additional Germanic languages) from the 55 contexts identified as the extended knowledge domain in chapter 5. The distance measure is very similar to the one described there, the difference being i) that principle 5 is not applied, that is, similarity is allowed between contexts without shared lexical items and ii) there is no weighting applied to the different parts of the codings. This is the all-equal similarity measure. The reason for using this measure is that we are equally interested in constructional means of expressing distinctions within the domain as lexical ones.

Figure 7.2 shows the eigenvalues of the dimensions of the CMDS analysis of the 55 considered contexts. There are negative values, which as mentioned shows that the original distances were not entirely Euclidean. However, the magnitude of the negative values is small, indicating that the distances are similar enough to Euclidean that treating them as such is unproblematic (Borg & Groenen 2005: 262). As mentioned, larger eigenvalues means a larger proportion of the variance explained. As can be seen, there is a so-called elbow at 6 dimensions. That is, after 6 dimensions, the informativeness of the dimensions begins to be quite small, resulting in a much flatter curve and giving the appearance of an elbow in the curve. This elbow method is often recommended for choosing the dimensionality of an MDS analysis (Borg & Groenen 2005: 48).

In the following sections, the first 6 dimensions of the CMDS analysis will be considered. As discussed in Wälchli & Cysouw (2012: 685) there are two factors that influence how early in a CMDS analysis variation patterns will show up: how many data points it is relevant for and how many languages it is relevant for. In other words, patterns of variation that show up in the higher—that is, later—dimensions are likely to be relevant only for a few verses or for a small number of languages. For the current purposes of trying to uncover what the main parameters underlying the linguistic expression of the knowledge domain are cross-linguistically they are therefore not particularly relevant. It can be noted that the 6 dimensions considered here account for 92% of the variation (calculated using an adaptation of Dawson 2015).

However, we can also choose to focus on the fact that this leaves almost a tenth of the variation unexplained. This tells us that even with these large-scale semantic features considered, there is room for considerable language-specific variation. The focus on the larger patterns should not be taken to suggest that the distribution and meaning of individual forms in individual languages is uninteresting; it is simply a result of the method chosen.

The suggested interpretations of the first six dimensions are summarised
Figure 7.2: Eigenvalues of the dimensions of the CMDS solution, negative values in red

in table 7.3, they will be discussed in turn below. Before that, two factors important for the interpretation should be emphasised. First, that the contexts investigated through CMDS were chosen through the clustering of data collected for the wider cognition-perception domain, in which they cluster together. This means that the contexts considered here are all closely connected linguistically, i.e. they are generally expressed in the same or similar ways. In the following, the focus will usually be on languages where there is a distinction made along one of the dimensions, but it should be kept in mind that the default, so to speak, is sameness of expression.

The second point is that the results described here need to be considered

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Low pole</th>
<th>High pole</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>States</td>
<td>Achievements</td>
</tr>
<tr>
<td>or</td>
<td>Know</td>
<td>Understand</td>
</tr>
<tr>
<td>2</td>
<td>Internal</td>
<td>External</td>
</tr>
<tr>
<td>3</td>
<td>Know that</td>
<td>Know person, Recognise</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Realise</td>
</tr>
<tr>
<td>5</td>
<td>Not know</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Know how</td>
</tr>
</tbody>
</table>

Table 7.3: Summary of the interpretation of the first six dimensions.
in light of the process of coding verses. As discussed in section 3.6.2, many things which could potentially have been included in the coding were not. This includes features that show no contrast within the domain but also distinctions within the domain that are simply instances of the same distinctions made in the language at large. Cases where this is likely to have affected the outcome of the CMDS analysis in relevant ways are pointed out and discussed throughout.

### 7.4 Dimensions 1 and 2: Understanding and finding out

We begin by considering the first and second dimensions, which together account for around 65% of the variation. The first two dimensions are closely intertwined and will therefore be considered together.

The position of the verses along Dimensions 1 and 2 are shown in figure 7.3. Three clusters are clearly distinguishable; they are marked with A, B and C. Cluster A (exemplified in table 7.4) is on the low end of both Dimensions 1 and 2. Cluster B (see table 7.5) is on the high end of Dimension 2, but is intermediate between A and C in Dimension 1. The cluster marked C (see table 7.6), finally, is on the high end in Dimension 1 but on the low end of Dimension 2. We will begin to approach the semantics of Dimensions 1 and 2 by considering the contexts in these regions.

---

1. All dimensions are available in the Online Supplementary Materials.
Beginning with cluster A, it consists of typical ‘know’ contexts, i.e. those that in chapter 5 were labelled KNOW. Table 7.4 lists the three verses with the lowest value for Dimension 1.

The group marked with B, contexts with high values in Dimension 2, corresponds to the group labelled FIND OUT in chapter 5. Table 7.5 gives the three contexts with the highest value in Dimension 2.

Moving on to cluster C we find that the contexts there correspond to what was in chapter 5 called the UNDERSTAND-cluster. Table 7.6 gives the three verses with the highest values in Dimension 1.

### Table 7.4: The three verses with the lowest values in Dimension 1, from low to high

<table>
<thead>
<tr>
<th>Cluster A</th>
<th>Verse</th>
<th>English (Lexham)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cor 1 8:26</td>
<td>[—] for we do not know how to pray as one ought [—]</td>
</tr>
<tr>
<td></td>
<td>John 14:4</td>
<td>And you know the way where I am going.</td>
</tr>
<tr>
<td></td>
<td>Matt 26:02</td>
<td>“You know that after two days the Passover takes place[—]”</td>
</tr>
</tbody>
</table>

### Table 7.5: The three verses with the highest values in Dimension 2, from high to low

<table>
<thead>
<tr>
<th>Cluster B</th>
<th>Verse</th>
<th>English (Lexham)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acts 23:34</td>
<td>So after reading the letter and asking what province he was from, and learning that he was from Cilicia</td>
</tr>
<tr>
<td></td>
<td>John 12:9</td>
<td>Now the large crowd of Jews found out that he was there</td>
</tr>
<tr>
<td></td>
<td>Mark 15:4</td>
<td>And when he learned of it from the centurion</td>
</tr>
</tbody>
</table>

### Cluster C

<table>
<thead>
<tr>
<th>Verse</th>
<th>English (Lexham)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matt 13:51</td>
<td>“Have you understood all these things?”</td>
</tr>
<tr>
<td>John 8:27</td>
<td>They did not know that he was speaking to them about the Father.</td>
</tr>
<tr>
<td>Mark 9:32</td>
<td>But they did not understand the statement</td>
</tr>
</tbody>
</table>

### Table 7.6: The three verses with the highest values in Dimension 1, from high to low
Since these distinctions appeared already in the global clustering (see chapter 5), it is not surprising that they appear in the first two dimensions of the CMDS solution.

The same two first dimensions are displayed in figure 7.4, this time with the Swedish forms used displayed. Note that, as mentioned, the position of the verses is exactly the same as in figure 7.3; only the appearance of the dots have been changed to reflect the Swedish expression. The lower right corner is taken up by förstå ‘understand’. The high end of the second dimension is more varied, but is dominated by the two forms få veta and få reda på, which consist of the in this context auxiliary verb få ‘receive, get’ (also ‘be allowed’) together with either veta ‘know’ or the mainly to this and some closely related constructions limited reda. Cluster A—the lower left corner—is taken up by two forms, namely veta and känna—the distinction between these is considered further below and is not relevant to Dimension 1 and 2. The figure also contains approximate English labels for the regions discussed; it should be kept in mind that these should be taken as convenient labels for groups of contexts rather than as some sort of semantic primitives (cf. section 3.3).

Is it possible to move beyond these three clusters to a semantic characterisation of the dimensions individually? As discussed in sections 2.2.4 and

---

1 What van der Klis & Tellings (2022: 641) call map coloring.

2 The noun reda is a highly polysemous noun which in modern Swedish occurs mainly in a number of fixed phrases, among them få reda på and ordning och reda lit. ‘order and reda’ ‘order’, see SAOB reda subst (1967)
2.2.5. ‘understand’ appears to be a complex concept with many components. One of these is the tendency to be an achievement, something which is clearly shared by the ‘find out’ contexts. Thus, Dimension 1 appears to correspond to a split between ‘know’ as a state and ‘know’ as achievement.

In support of this analysis, consider the data from Bengali displayed in figure 7.5. The domain is divided between three main lexemes: jana ‘know’, cina ‘know (person)’ and boojha ‘understand’. A further divide, however, is made between whether the auxiliary verb para approximately ‘DO’ is used or not; without it, we have states (left side) and with it achievements (right side). Examples of the use of para with this meaning are given in 52. In 52a an example is given of the general ‘know’-word in Bengali, jana, used as the finite verb of the clause. In 52b we find the same verb, but in a non-finite verb form in -te (sometimes called Infinitive, sometimes Imperfective participle) together with para; the meaning here is know-as-achievement.

(52) Bengali (Indo-European>Indo-Aryan, Bangladesh)

a. Mark 12:14

āmarā jāni āpani ēkajana sat lōka
1PL know.1.PRS 2PL.HON one person honest

‘we know that you are an honest man’ (AS)
The situation in Bengali is in fact representative of the whole sample; the division between ‘know’ and ‘understand’ is primarily lexical while the division between know-as-state and know-as-achievement is primarily constructional. Recall that all verses in the sample are coded both for lexical and constructional markers (see section 3.6.2). The similarity between verses is then calculated based on both. However, it is not necessarily the case that distinctions are made according to the same parameters lexically and constructionally. In figure 7.6, on the left, we see the same configuration of verses already displayed, with the distance measure taking both lexical and constructional coding into account. In figure 7.6, on the right, we see the results of taking only lexical information into account. As can be seen, the divide between ‘know’ and ‘understand’ remains primary. In figure 7.6 (left), the ‘find out’ cluster is intermediate between these two clusters, which I have explained as being the result of a shared achievement component. However, looking at figure 7.6 (right), it can be seen that the intermediacy of ‘find out’ is largely gone—instead ‘find out’ is positioned with ‘know’ in the first dimension, with ‘understand’ alone in being distinct.

This shows that the similarity between ‘find out’ and ‘understand’ is cross-linguistically mainly realised at the constructional level; if we consider only lexical items, ‘find out’ and ‘understand’ share no more expression than ‘know’ and ‘understand’.

To further illustrate this point, consider what happens when only constructional markers are considered, as shown in figure 7.7. Here, the first dimension appears to be split between states and achievement; note how the distribution of the auxiliary para is split down the middle.

In short then, the situation can be characterised as follows. Constructionally, languages tend to be oriented toward the distinction between know-as-state and know-as-achievement. This distinction, however, is rarely realised in lexical stems, where instead the main distinction found is between ‘know’ and ‘understand’. This split is reflected in the first dimension, which both shows the division between ‘know’ and ‘understand’ and state and achievement.

This leaves the question of how to characterise the second dimension. It appears that the distinction here is between immediately externally caused knowledge and knowledge that lacks this component. As an example, consider the data from North Tanna (Austronesian>Malayo-Polynesian, Vanuatu)
Figure 7.6: Left: Dimensions 1 and 2 of CMDS solution with lexical and grammatical markers in distance measure. Right: Dimensions 1 and 2 of CMDS solution with only lexical markers in distance measure. With selected Bengali forms

Figure 7.7: Dimensions 1 and 2 of CMDS solutions with only constructional markers in distance measures with selected Bengali forms
in figure 7.8. It shows that the knowledge domain in North Tanna is mainly covered by the verbs -itun ‘know’ and -o'ruru ‘not know’. However, contexts on the high end of Dimension 2 are instead expressed using -eto ‘hear’. In this, North Tanna—as well as many other languages—appears to be oriented toward the fact that at this pole, the contexts become more or less indistinguishable from perception. This is a distinction that reoccurs in Dimension 4, where ‘hear’-like contexts are on the opposite pole from ‘realise’-like contexts.

In chapter 8, the asymmetries between the meanings distinguished in Dimensions 1 and 2 are discussed and in section 9.3.2 the semantics of the know-as-state vs know-as-achievement is discussed, followed by the semantics of ‘understand’ in section 9.3.3 and of ‘find out’ in section 9.3.4. In chapter 11, the expression of these dimensions cross-linguistically is discussed.

7.5 Dimension 3: Knowing people

Moving on to Dimension 3, this dimension can be interpreted as corresponding to the difference between knowing people and knowing facts. To see this, consider the distribution of Swedish känna, cognate with and similar to German kennen, in figure 7.9. As was discussed in section 2.2.2, the meaning of känna can be approximately given as ‘know somebody (personally)’. The same root is used but with the verb particle igen ‘again’ at the highest end of the dimension, with the approximate meaning ‘recognise’. Table 11.2 gives
Figure 7.9: Dimensions 1 and 3 of the CMDS analysis with Swedish *känna* and *känna igen*

three representative contexts along Dimension 3; the verses shown are indicated with arrows in figure 7.9.

We are dealing here with a scale, moving from ‘recognise’ through ‘know person’ to ‘know who’—this will be considered in more detail in section 11.4. Note that in the verse with the lowest value given in table 11.2 (Mark 1:24) the Swedish example uses *veta* rather than *känna*. As can be seen from the closeness to the ‘know that’ contexts, this is a common pattern cross-linguistically. However, in some languages also contexts with a ‘know who’ meaning are subsumed under the form for knowing a person. An example is Tlachichilco Te-

<table>
<thead>
<tr>
<th>Verse</th>
<th>English (Lexham)</th>
<th>Swedish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mark 6:54</td>
<td>And as they were getting out of the boat, people <strong>recognized</strong> him immediately.</td>
<td>Så snart de steg ur båten <strong>kände</strong> folket <strong>igen</strong> honom</td>
</tr>
<tr>
<td>John 7:29</td>
<td>I <strong>know</strong> him, because I am from him and he sent me.</td>
<td>Jag <strong>känner</strong> honom därför att jag kommer från honom</td>
</tr>
<tr>
<td>Mark 1:24</td>
<td>I <strong>know</strong> who you are—the Holy One of God!</td>
<td>Jag <strong>vet</strong> vem du är, Guds helige</td>
</tr>
</tbody>
</table>

Table 7.7: Selected verses with high values in Dimension 3, from high to low
pehua in which we find the lexemes -k’atsa- ‘know (that)’ and -mispa- ‘know (person)’. The latter form is used with a person as the object (53a), but also with a who-clause as an object (53b), compare with the use of -k’atsa- in 53c.

(53) Tlachichilco Tepehua (Totonacan, Mexico)

a. Acts 7:18

\[\text{yu } jantu } \text{ix-} \text{mispa-} \text{y } \text{ni } \text{José} \]
REL NEG PST-know-IPF ART PR.NAME

‘who did not know Joseph’

b. Mark 1:34

\[\text{yu‘uncha } \text{ix-ta-mispa-y-cha } \text{ta’ayucha } \text{ix-juni-ta} \]
3PL PST-3PL-know-IPF-IAM who PST-become-IMPF
\[\text{ni } \text{Jesús} \]
ART PR.NAME

‘They knew who Jesus was’ (AS)

c. Luke 8:53

\[\text{ix-ta-k’atsa-y ni } \text{ix-ni-ta-cha} \]
PST-3PL-know-IPF COMP

‘Because they knew that she was dead’

In section 9.4 the semantics of Dimension 3 is investigated more closely and in section 11.4 the expression of these meanings cross-linguistically is discussed.

7.6 Dimension 4: Realisation

The fourth dimension of the CMDS analysis is closely connected to the first and second dimension and is best understood in relation to these. Figure 7.10 shows Dimension 4 together with Dimension 1 (left) and Dimension 2 (right), with Swedish forms. As can be seen, the dimension makes further distinctions primarily within the FIND OUT and UNDERSTAND clusters—i.e. among the achievement contexts. What appears to be distinguished semantically is realisation.

In table 7.8 the three contexts with the highest values in Dimension 4 are given. Note the variation in which form is used in the English translation; this hints at the semantic complexity of the meaning distinguished and underscores that using a single English form as a label cannot be the endpoint of the analysis.

158
Figure 7.10: Dimension 4 with Dimension 1 (left) and Dimension 2 (right), selected Swedish forms

<table>
<thead>
<tr>
<th>Verse</th>
<th>English (Lexham)</th>
</tr>
</thead>
<tbody>
<tr>
<td>John 1:22(b)</td>
<td>and they <strong>realized</strong> that he had seen a vision in the temple</td>
</tr>
<tr>
<td>Acts 19:34</td>
<td>But when they <strong>recognized</strong> that he was a Jew,</td>
</tr>
<tr>
<td>Matt 17:13</td>
<td>Then the disciples <strong>understood</strong> that he had spoken to them about John the Baptist.</td>
</tr>
</tbody>
</table>

Table 7.8: The three verses with the highest values in Dimension 4, from high to low
As an example of a lexical item clearly related to Dimension 4, consider Mapudungung (Araucanian, Chile) ngüeneduam- ‘realise’, the distribution of which is shown in figure 7.11. The form consists of the two roots ngüne ‘govern, direct’ and duam ‘attention’, i.e. literally roughly ‘to govern one’s attention (toward)’. An example is given in (54). As will be seen in section 9.5, the close connection to internal processing is characteristic of this dimension.

(54) Mapudungun (Araucanian, Chile)
   a. Acts 19:34

   Welu ngüeneduam-nge-lu kisu ta-ñi judiu-nge-n
   but realise-PASS-TEMP self DEF-3POSS Jew-VRBZ-NMLZ

   ‘But when they recognized that he was a Jew’

The semantics of Dimension 4 are addressed in section 9.5 its expression in the languages of the sample in section 11.5.

7.7 Dimension 5: Not knowing

Moving on to Dimension 5, lexical items from European languages give little insight into how it is to be interpreted. Considering the contexts, however, tells us that the distinction is mainly between not knowing (low values) and
Table 7.9: The three verses with the lowest values in Dimension 5, from low to high knowing (high values). The three verses with the lowest values in Dimension 5 are given in table 7.9.

As an example of a language in which linguistic form is clearly sensitive to this dimension, consider the data from Moro shown in figure 7.12. Here, the knowledge domain is mainly taken up by two forms: -iğoba which is used in contexts of not knowing (example 55a), and -liyet (example 55b) which is used in contexts of knowing.

(55) Moro (Heibanic, Sudan)

a. Mark 14:71

  i-g-a-iḡoba  eda  igi
  1SG-CL-FIN-not.know  man  DEM

  ‘I do not know this man’

b. Acts 19:34

  i-g-a-liyet-o  yesu
  1SG-CL-FIN-know-?too  PR.NAME

  ‘I know Jesus’ (AS)

I will refer to this phenomenon as lexical negation or lexically negated forms:

Definition 11. Lexical negation or lexically negated forms are forms which express the negation of concept X but do not contain a synchronically overt marker of negation.

Note that this is quite a broad definition, including for example expressions as English ‘be ignorant of’. The definition is in practice narrowed by the fact that only forms which occur in the verse sample are included. Of course it cannot
be guaranteed that this shows a general frequency or basicness in the language, but any further such considerations are impossible given the material at hand.

Dimension 5 is thus primarily sensitive to the lexical negation of knowing. The term ‘lexical negation’ is preferable to suppletive negation in this context, since, as we will see (section 11.6), the situation is only rarely something akin to suppletion. Lexical negation of knowing occurs in some form in a little less than a fourth of the languages of the sample (18 languages).

Note that the dimension is sensitive to lexical negation specifically. As discussed in section 3.6.2, standard negation is not included in the coding schema. If it was, we would expect it to show up very early in the CMDS analysis since all languages in the sample, not unexpectedly, in some way express the negation of knowledge.

Perhaps the most striking feature of Dimension 5 is the very clear separation of one verse from the others at the low end of the dimension. This is reflective of an interesting feature of the typology of lexical negation forms, namely that for many of the languages with lexical negation, only Luke 12:56(b) (the clearly separated verse) is expressed with a lexically negated form. Within this group of languages, what is found is actually a lexically negated modal form, i.e. a form expressing non-ability, extended to a context of ‘not knowing how’.

In section 9.6 the semantics of lexical negation are investigated—what can be said about these forms beyond the fact that they express ‘not knowing’. Section 11.6 considers the distribution of lexical negation in the verse sample and geographically.
7.8 Dimension 6: Knowing how

The last dimension to be considered, Dimension 6, clearly distinguishes three contexts. These contexts, shown in Table 7.10 below, can be characterised as ‘know how’ contexts.

There are indications that the Bible corpus is not ideal for investigating this kind of contrast, i.e., between ‘know that’ and ‘know how’. For example, for Southern Kisi (Atlantic-Congo>Mel, Sierra Leone) Childs (1995: 202) mentions the form caa ‘know how’. The form does appear in the NT (as chaa) but not in the verse sample and not in ‘know how’ contexts (it appears to be one of the components of a construction meaning something like ‘person who does X well’). In addition to these cases where a form (or a use of a form) noted in a grammar is missing in the verse sample, there are several examples of the form in the NT being more limited in its use than the description in the grammar would suggest. An example is Tlahuitoltepec Mixe where Romero-Méndez (2009) gives the form -jät- ‘know how’ which can occur with an embedded question as complement—see example 56a (cf. English). The form -jät- appears in the verse sample (as -jajt-) but only with a nominal object (example 56b).

(56) Tlahuitoltepec Mixe (Mixe-Zoque, Mexico)
   a.  
   \begin{flushleft}
   \textbf{Romero-Méndez} (2009: 550)
   \end{flushleft}
   \begin{flushright}
   ka’t=ja’a\quad t-ta-jät-y\quad xě’n\ y-ak-yo’oy-t
   \end{flushright}
   \begin{flushright}
   NEG=DEM.D\quad 3A-?-know.how\quad DEP\quad how\quad 3A-CAUS-walk-IRR;DEP
   \end{flushright}
   ‘He does not know how to drive it (a car)’
Regardless of these possible limitations of the data, there are several languages in the sample where special forms or constructions are found for ‘know how’. As examples, consider Car Nicobarese where the form *huveuēu* ‘know how’ is found or Swedish where we find *kunna* ‘know how, be able to’ (figure 7.13). As we will see in section 11.7 this pair also illustrates two of the main types of languages regarding the expression of ‘know how’ in the languages—whether the form is identical to the form for ‘be able to’ or not. This is not the case in Car Nicobarese, while it is in Swedish.

These types—as well as other relevant ones—are discussed in section 11.7, and in section 9.7, the semantics of ‘know how’ is considered.

### 7.9 Summary of the dimensions considered

In the preceding sections, a first survey of the six first dimensions of the CMDS analysis of the knowledge domain was given. Dimension 1 was identified as being sensitive to two semantic parameters—the distinction between state and achievement and the distinction between ‘know’ and ‘understand’. Dimension 2 makes a further distinction among the achievement contexts, distinguishing externally caused knowledge. Dimension 3 can be characterised as a scale from ‘recognise’ to ‘know person’ to ‘know that’. Dimension 4 further refines
the categorisation of achievements by distinguishing a group of ‘realise’ contexts. Dimension 5 is primarily sensitive to the lexical negation of knowledge, but interacts with ‘know how’, which is the primary parameter underlying Dimension 6.

The interpretation of the dimensions as well as the position of the verses along them are shown summarised in figure 7.14. Labels are placed in the locations of the verses to which they apply.

With this overview of the structure of the knowledge domain—addressing research question B—we now turn to the sub-questions of B, addressing in more detail various aspects of the knowledge domain’s internal divisions. The next chapter (8) discusses the asymmetries found in the expression of the different part of the knowledge domain and argues that these can be explained by positing a core-periphery structure of the knowledge domain. Chapter 9 more closely discusses the semantics of the dimensions identified. Chapter 10 considers the connections between knowledge and other concepts. Finally, chapter 11 deals with how the dimensions identified are expressed in the languages of the sample.
8. The core-periphery structure of the knowledge domain

8.1 Introduction

In the previous chapter, an overview of the dimensions of the knowledge domain based on a CMDS analysis was presented. This chapter addresses the relationships between the parts identified, research question B1:

B1. How do [the knowledge domains’] meanings relate to one another—are some more basic or central?

The meanings or functions included in the extended knowledge domain are not, I will argue, all equal. There are recurrent motivational asymmetries between what can be seen as the core of the domain—‘know that’—and its periphery—‘understand’, ‘find out’, ‘recognise’, ‘know person’ and ‘know how’. This chapter presents evidence of these asymmetries.

Section 8.2, after an introduction to lexical motivation, considers two kinds of motivational asymmetry in the extended knowledge domain—directionality of formal motivation (8.2.1) and proportion of colexification or domain-external motivation (8.2.2). Section 8.3 finally, argues that these asymmetries can be explained by positing a core-periphery structure of the knowledge domain, with ‘know that’ as the prototype or core. In table 8.1 the methodology of the chapter is summarised.

8.2 Asymmetries in lexical motivation

The arbitrariness of the sign is a fundament of linguistics. A linguistic form has no relationship to what it signifies beyond conventionality. As has been recognised for a long time, however (e.g. Saussure 1916) this is not true across the lexicon. In all languages we find a group of words which in their phonetic shape are iconically related to their meaning—phonetically motivated words. More relevant for this thesis, however, are those words which are relatively motivated—motivated in relation to some other form in the language. As an example, the French poirier ‘pear-tree’ is relatively motivated by poire ‘pear’.
B1. How do [the knowledge domain’s] meanings relate to one another—are some more basic or central?

Are some parts of the knowledge domain more often motivated by others, and are some parts more often motivated domain-externally?

Measure of directionality of formal motivation, colexification count

Core Knowledge, (Full)

Table 8.1: Methodological summary of chapter 8

While the relationship between the form poire and ‘pear’ is arbitrary and thus in a sense also between poirier and ‘pear-tree’ (there is nothing that directly connects the form to the meaning) the latter is relatively motivated by the link between poire and ‘pear’. The existing connection between the form poire and the meaning ‘pear’ is extended—motivates—a new form. Such cases of relative lexical motivation are ubiquitous in the world’s languages (see e.g. Koch & Marzo 2007, Urban 2016). Koch & Marzo (2007: 262) argue that lexical motivation always consists of a formal relationship and a cognitive relationship. The formal relationship includes things like derivation and compounding but also formal identity, i.e. polysemy or colexification. The cognitive relationship refers to the conceptual link between the two meanings that motivate their full or partial co-expression, for instance metonymy or metaphor (cf. section 10.4.1).

Importantly, the motivational relationship may in some cases be clearly directed—as in the poirier-poir example given above. There is a clear directionality to the formal part of the motivational relationship here, and it generally follows that also the cognitive part of the relationship has this direction. In cases of formal motivation consisting of non-identity it is generally easy to identify the directionality of the relationship, but when the formal relationship is colexification it is more difficult to do so without diachronic or comparative data (Koch & Marzo 2007: 282–283).

I suggest that if meaning A is regularly motivated by meaning B—but not the other way around—we may speak of a cross-linguistic asymmetry in the directionality of motivational relationships. Similarly, if meaning A is more frequently than meaning B motivated by domain-external meanings, we may speak of an asymmetry in the likelihood of external motivation. In the following sections, I will show that both these kinds of asymmetries are present in the extended knowledge domain.
8.2.1 Directionality of formal motivation

There are clear differences in the data regarding which forms are motivated by which. Starting with a zoomed out view of the extended knowledge domain, is can be observed that when forms for KNOW and UNDERSTAND are partially colexified, UNDERSTAND is always the more complex form. There is a similar tendency—though one with exceptions—for FIND OUT to be a more complex form (cf. section 11.2). An example from South Kisi is given in 57 (see p. 196 for further discussion of the example). ‘Know’ is expressed with the verb sina (example 57a). ‘Understand’ uses the same verb, but with the ideophone pèıgú added (example 57b).

(57) Southern Kisi (Atlantic-Congo>Mel, Sierra Leone)

a. Mark 12:14

蒂查, ị sina māa ã chō sàkpō
teacher(Eng.), 1PL know COMP 2SG be righteous/truthful

‘Teacher, we know that you are truthful’

b. Mark 14:68

1.sg sina [nyê ị chō dìmìòó wò] pèıgú le

1SG know thing 2SG AUX speaking REL IDEOPH NEG

‘I do not understand what you are talking about’ (AS)

In this section, I will present quantitative evidence for the asymmetry of the directionality of formal motivation. Of interest here are cases where there is a motivational relationship between contexts—that is, where context A and B share at least one lexical element.

To investigate this, the following procedure was used. In each language, the codings of verses are compared across the clusters—each verse in group A (the source group) and each verse in group B (the target group) are compared. As noted, we are only interested in those pairings where the verses share a lexical morpheme. For those cases, the average number of markers found in verses in group A but not group B is calculated, across languages.

The markers are then weighted according to how good representatives they are of the target cluster, i.e. in what proportion of verses in the target cluster they occurs. A marker which occurs in all contexts in the target cluster has a value of 1, while a marker which occurs in only 1 out of 5 contexts in the target cluster has a value of 0.2. Finally, the counts in both directions for each pair—for instance, A to B and B to A—are compared, and the smaller is subtracted from the larger. This corrects for the effect of markers which are more or less

168
evenly distributed across the clusters. As an example, consider the example of Yalunka (Mande, Sierra Leone) where the form kolon is used for most of the extended knowledge domain. In some contexts, a particle ba with unclear meaning but connected to second person subjects also occurs. There are cases where a verse from the UNDERSTAND cluster contains ba while the verse from the KNOW clusters does not, as well as vice versa. Clearly, ba does not bring about either the KNOW or the UNDERSTAND meaning, and the two steps described above correct the procedure for this.

Additionally, the second step of comparing the value in both directions accounts for the effect of cases where both forms in both clusters are motivated by an unknown third.

The procedure was applied to the core language sample, giving a single value for each pair of clusters, giving 1) the directionality of the formal motivation and 2) the strength of the relationship—this is the measure of the directionality of formal motivation.

The directionality measure values can then be visualised as a network, where the clusters are nodes and the edges the value of the measure. Note that what is counted is only contexts where there are shared lexical morphemes—this means that the figure tells us nothing about how common a formal motivational relationship is, only in what direction it typically moves when it actually exists (for a view of the former, see section 5.4.1 or chapter 10).

The resulting network graph is shown in figure 8.1. The direction of the edges indicates the directionality of the relationship and their width the strength of it. Included are the nodes of the extended knowledge domain—KNOW, UNDERSTAND and FIND OUT—as well as the perception nodes SEE and HEAR. The relationships are largely what would be expected. The asymmetry between perception and cognition is seen in a strong directionality from e.g. SEE to KNOW or HEAR to UNDERSTAND. An interesting exception is the link from KNOW to HEAR. The main reason for this appears to be two languages where ‘hear’ is expressed in terms of ‘see’ and the form for ‘see’ is also used (sometimes modified) as ‘know’. Thus, for Cuiba (Guahiboan, Colombia) the form ‘see’ is ta-, the form for ‘hear’ is jume ta- ‘word/voice see’ and the forms for ‘know’ are yaputa- and yaita-. The other language is Warlpiri (Pama-Nyungan, Australia), where the form nya- ‘see’ is sometimes used for ‘know’ and is the basis for purda-nya- ‘hear’. Related to this, note also the strong directnedness between SEE and HEAR—a result clearly in line with the hierarchy of sense modalities argued for already by Viberg (1984).

Focusing solely on the knowledge domain gives the figure in 8.2.

Interestingly, the directionality is somewhat stronger from KNOW to FIND

---

1 Note the scale adjustment of the edges vis á vis 8.1
Figure 8.1: Network graph of formal directionality between clusters—knowledge and perception

Figure 8.2: Network graph of formal directionality between clusters—knowledge
OUT than KNOW to UNDERSTAND—despite the observation of a unidirec-
tional relationship between ‘know’ and ‘understand’ as opposed to between
‘know’ and ‘find out’. A possible explanation for this is that full colexifica-
tion is relatively more common between KNOW and UNDERSTAND than
between KNOW and FIND OUT.
Within the narrow knowledge domain, the evidence regarding directionality
is not as clear; clearly directional formal motivational relationships are rarer
within the narrow knowledge domain than in the extended\footnote{The exception, of course, is the category of ‘not knowing’—when not expressed lexically, this is always in the sample marked relative to ‘know’, but as discussed in section 3.6.2, this is not included in the coding of the data.} Applying the
same procedure of calculating directionality as above gives the network graph
shown in figure 8.3 (the groups used are the categorisations of the initial verse
sampling, see 3.6.2 and Appendix B). When considering this figure, it is im-
portant to remember that the procedure for calculating directionality is based
on groups of contexts rather than form. What this means in practice is that the
line from e.g. Know how to Recognise does not indicate that forms which are
exclusive to Know how motivate Recognise forms, but that forms which occur
in the Know how group—but which are typically shared with Know that and
Know person—motivate Recognise.
As figure 8.3 shows, there are few links of particular interest within the
narrow knowledge domain. The exception is Recognise, toward which the
edges show directionality for all other meanings found in the domain—Know
that, Know person and Know other nominal. This is in part because ‘recog-
nise’ is often explicitly encoded as an achievement, similarly to ‘find out’ (see
further section 9.4). What is interesting to note is that unlike the relationship
between stative ‘know that’ and achievement ‘know that’ (i.e. ‘find out’), there
are no examples in the data of the opposite relationship for ‘know person’ and
‘recognise’.
The lack of clear directionality between the other meanings is mostly due to the
fact that meanings within the extended knowledge domain (‘recognise’ to some
extent excepted) tend to be either fully co-expressed or completely distinct.
The situation of clearly directed motivational relationships within the narrow
knowledge—Recognise excluded—domain are, in other words, rare—there is,
in fact, only one clear example in the sample, namely Thai, as discussed later
on (see p. 212).

8.2.2 Domain-external motivation

In addition to the asymmetries in the directionality of motivational relation-
ships within the knowledge domain, there are asymmetries in the data regard-
ing the likelihood of motivation external to the domain—in other words, how often contexts in a group are expressed by constructions which colexify another concept. We saw in chapter 6 that the absolute colexification of ‘know’ with other concepts is rare in the sample and found only in two Papua New Guinea languages. Connections with some part of the knowledge domain are much more frequent, and this frequency differs between the different parts (see also chapter 10).

I will here consider a context colexified with a non-knowledge concept—that is, be externally motivated—if it shares a lexical morpheme with forms from the SEE, HEAR, THINK, REMEMBER or BE ABLE. Note that we are considering partial colexification as well as full. In the raincloud plot in figure 8.4, the proportion of colexification in the core language sample (varying from 0, not colexified in any language, to 1, colexified in all languages) per verse is shown per cluster.

It is clear that KNOW is more rarely expressed in terms of other domains than FIND OUT and UNDERSTAND. In fact, the mean rate of colexification in the FIND OUT cluster varies from around 20% for the least colexified contexts—the same as UNDERSTAND—to over 50% for the most colexified contexts.

---

1 As discussed in chapter 10, BELIEVE and OBEY are never colexified with knowledge without a mediating concept.

2 A raincloud plot (named for the similarity to rainclouds when turned 90 degrees counterclockwise from the orientation in the plots here) shows, from left to right 1) the distribution of the datapoints in a given number of bins, 2) a traditional boxplot with mean and quartiles and 3) the density distribution.
contexts, which essentially means that although these contexts cluster with
the knowledge domain, they are in fact equally likely to expressed in terms
of other domains. In the case of FIND OUT this domain is primarily hear-
ing. The difference between KNOW and UNDERSTAND is not as extreme,
but nonetheless considerable. The majority of the verses in the large KNOW-
cluster show a clearly lower rate of colexification, while a smaller number are
on a par with UNDERSTAND or the lower end of FIND OUT.

The colexifications in the categories of the narrow knowledge domain also
offer some insight. The data are shown in figure 8.5 constructed in the same
way as figure 8.4 above but with categories in the narrow knowledge domain.
Know that is most rarely colexified with other domains, followed by Know
other nominal, Know how and Know person with very similar levels. Finally,
Recognise has the highest proportion of colexification. The colexifications in
the two latter categories are, as we will see 9.4 10.3.2, mainly with ‘see’.

8.3 The core-periphery structure of the knowledge domain

I have, in the previous section, presented evidence of asymmetries in the moti-
vational patterns of the knowledge domain. How are these asymmetries to be
explained? In this section, I will argue that by applying the frame based model
of meaning and the prototype model of categories presented in section 3.1,
an explanation can be offered. There are two basic asymmetries that require explanation:

i) In the extended knowledge domain, UNDERSTAND and FIND OUT are motivated by KNOW, but not the other way around; in the narrow knowledge domain, Recognise is motivated by Know that/person/other nominal, but not the other way around

ii) In the extended knowledge domain, UNDERSTAND and FIND OUT are more likely to be motivated by forms outside the knowledge domain than KNOW; in the narrow knowledge domain, Recognise, Know person, Know nominal and Know how are more likely to be motivated by forms outside the knowledge domain that Know that.

Starting with [i], we turn to the principle of isomorphism ([Haiman 1985] 19). We can distinguish between paradigmatic and syntagmatic isomorphism ([Croft 2003] 102–110). Paradigmatic isomorphism refers to the fact that in a language system, generally, one form will correspond to one meaning—synonyms as well as homonyms are avoided. Syntagmatic isomorphism refers to the fact that in a string of language, one form will have one meaning—empty or portmanteau morphs are avoided. This is a widely accepted principle in the organisation of language.
Thus, we expect that when form \( x \) occurs, it corresponds to meaning \( X \). And if we find form \( xy \), we still expect \( x \) to correspond to \( X \). Additionally, we expect \( y \) to correspond to some other meaning \( Y \). So if we have, for example, the forms *sina* ‘know’ and *sina pèyugù* ‘understand’ (as in Southern Kisi, see above ex. 57), we expect that the same form will refer to the same meaning (*sina* = ‘know’) and that forms beyond that will have additional meaning.

In the frame approach to meaning taken here, this can be understood as follows. There are cases where there is a more or less exact match between a scene \( S \) and a frame \( (X) \) which corresponds to construction \( x \). This means that this construction—\( x \)—is highly suitable to express \( S \). If, however, the scene to be expressed deviates from the frame, form \( x \) becomes less suitable. However, if it is still recognised that \( S \) has some component in common with \( X \), isomorphism motivates the inclusion of \( x \) in the expression of \( S \) in addition to additional morphological material, isomorphic to the difference between \( S \) and \( X \).

Relevant questions here are of course why not all languages then partially colexify the two meanings and why some colexify them fully. The answer to the latter question can, I believe, easily be given in terms of economy. Just as isomorphism is a motivating factor in language, so is economy - the “principle that the expressions should be minimized where possible” (Croft 2003: 102). Economy drives languages to subsume similar meanings under the same form, given that the communicative loss is not too great. Also the first question can be answered by reference to some kind of economical constraint; a language in which every commonality of meaning was reflected in form would be impossibly complex. It is not unexpected that languages should vary in which categories have composite forms and which do not.

Point ii) can be explained by the same account as i). Not sharing characteristics of one frame will often amount to sharing them with some other frame. That is, as a scene becomes less similar to frame \( X \) it often becomes more similar to frame \( Y \) or \( Z \). The situation is illustrated in figure 8.6.

We have three distinct frames—\( X \), \( Y \) and \( Z \). Scenes which are very similar to these will be easily be expressed via the corresponding form. Other scenes which do not precisely match any given frame but have similarities to all, can be expressed by either of the forms, with or without modification or by some combination of them. Which strategy is chosen may be a question of construal or conventionalisation in a given language.

The data presented above can thus be interpreted as ‘know that’ being the prototypical frame of the knowledge domain around which others centre. Meanings in the knowledge domain are understood in terms of this prototype—thus, the prototype may motivate more peripheral forms, explaining the asymmetry of directionality observed above. Similarly, deviation from the prototype
may lead to meanings being understood in terms of other meanings, and thus domain-external motivation.

The evidence is quite clear for the three clusters within the extended knowledge domain—according to the explanation given here, KNOW is clearly placed at the core as the prototype, while UNDERSTAND and FIND OUT are peripheral. The evidence regarding the parts of the narrow knowledge domain is not as clear; as mentioned, this is largely due to a rarity of directed motivational relationships within the narrow knowledge domain. The likelihood of external motivation as well as the one clear example of formal directionality within the narrow knowledge domain—that is, Thai—offer some evidence that it is ‘know that’ which is the core of the knowledge domain, but more research is necessary on this point.

8.4 Summary and conclusion

This chapter has presented evidence that there are systematic asymmetries in the motivational patterns of the different parts of the knowledge domain. These are most clear in the extended knowledge domain, but can also be observed in the narrow knowledge domain. I presented evidence of two asymmetries— asymmetry of directionality of formal motivation and asymmetry of domain-external motivation. I argue that we can understand these motivational asymmetries by applying a model of frame semantics and prototypes and placing
parts of the knowledge domain at its core—‘know that’—and other parts at the periphery.

This account does not offer an explanation of why ‘know that’ specifically should function as a prototype for the domain. Why not ‘understand’ or ‘know how’? Possible explanations include complexity or frequency. It might be argued that ‘know that’ is simple or basic in some way which makes it natural as a basis for comparison. This aligns well with the NSM claim of KNOW as a prime (though note that the prime is explicitly not ‘know that’, but something like what is found in ‘I know (it)’ [Wierzbicka (2018b)]. We might, of course, accept that the peripheral concepts (such as ‘understand’) are somehow conceptually dependent on the core concept ‘know that’ without accepting everything else from NSM. Another possibility is that ‘know that’-scenes are relatively more frequent, and good candidates for comparison by some combination of familiarity and saliency.
9. The dimensions of knowledge: Form and meaning

9.1 Introduction

In the previous two chapters the internal structure of the knowledge domain was considered; the distinction made were identified and the internal relations of the parts were described. In this chapter, the semantics underlying the identified distinctions will be investigated further, thereby addressing research question B2:

B2. How can the distinctions [in the knowledge domain] be characterised semantically?

I address this question by considering the forms used to express these distinctions; this way of approaching the problem is described in section 9.2. The purpose of this chapter, then, is twofold; it gives an inventory of the constructions which cross-linguistically are associated with the dimensions of knowledge and it uses the properties of these forms to make inferences about the semantics underlying the dimensions. The dimensions identified in chapter 7 are addressed one by one—section 9.3 concerns Dimensions 1 and 2, section 9.4 Dimension 3, section 9.5 Dimension 4, section 9.6 deals with Dimension 5 and 9.7 with Dimension 6. Finally, section 9.8 gives a summary of the chapter. Table 9.1 summarises the methodology of the chapter.

9.2 From forms to meaning

In analysing the semantics underlying a dimension of an MDS analysis, there are basically two ways to proceed. Firstly, we can consider the verses themselves and draw conclusions from their content in English or another familiar language—this was the main focus in chapter 7. Secondly, the distribution of forms across the contexts can be studied. This, in turn, can also be approached in two ways. The analysis can be based on forms whose meaning we already have some ideas about—e.g. because they are from our native languages. Alternatively, the composition of the forms themselves can be used
Research Question
B2. How can the distinctions [in the knowledge domain] be characterised semantically?

Operationalisation
What are the properties of the markers correlated with the dimensions of the CMDS analysis; what are their sources, colexification patterns, grammatical and combinatorial properties, and what does that tell us about the conceptualisations of the meanings distinguished?

Method(s)
CMDS coordinates, Point-Biserial Correlation

Language sample
Full

Verse sample
Knowledge

<table>
<thead>
<tr>
<th>Research Question</th>
<th>Operationalisation</th>
<th>Method(s)</th>
<th>Language sample</th>
<th>Verse sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>B2. How can the distinctions [in the knowledge domain] be characterised semantically?</td>
<td>What are the properties of the markers correlated with the dimensions of the CMDS analysis; what are their sources, colexification patterns, grammatical and combinatorial properties, and what does that tell us about the conceptualisations of the meanings distinguished?</td>
<td>CMDS coordinates, Point-Biserial Correlation</td>
<td>Full</td>
<td>Knowledge</td>
</tr>
</tbody>
</table>

Table 9.1: Methodological summary of chapter 9

to make inferences about meaning. As was discussed in section 3.1 the view of meaning taken here is one inspired by cognitive linguistics where meaning is conceptualisation. I sketched a process wherein content is understood by comparison to a collection of prototypical frames, which are connected with linguistic constructions.

Another crucial insight of cognitive linguistics for how forms can be used to infer meaning is that conceptualisation often involves metaphor, metonymy and other processes of characterising scenes in terms of other scenes—as discussed in the previous chapters, some expressions are motivated by other expressions. Lakoff & Johnson (1980) famously argued that a very large part of our everyday speaking and thinking is done through metaphor—TIME IS MONEY, GOOD IS UP etc. Put simply, we don’t simply take every new thing we encounter and make up a new word for it. Instead we compare the new experience to our previous experiences, draw analogies and make comparisons and then we express these new experiences in terms of old experiences. This means that many forms in language are more or less transparent regarding the conceptualisation underlying them in the sense that they are understood through their similarity to other concepts.

Note that a distinction needs to be made here between the level of system and speaker. It does not really matter whether the sources of the extensions are currently evident to language users—that the extensions at one point arose is evidence that the underlying conceptualisation has been present. The goal of using forms to infer semantics is not to give a synchronic picture of how language users conceptualise various experiences, but to use the range of attested conceptualisations—distributed in space and time—to make inferences about the underlying experience. Fortescue (2001: 23) explains a similar method of inferring semantics from cross-linguistic data in his study of the language of
cognition by comparing it to the fable of a group of blindfolded people who try to describe an elephant (an animal which they have no previous experience of).

One person touches the side of the elephant, and reports that an elephant is like a wall. Another touches the tusk, and says that an elephant is like a spear. A third touches the trunk and concludes that an elephant is like a snake. In other words, they go about describing the elephant in the same way language users conceptualise experiences—by comparing it to previous experience and understanding it in terms of these. The analogy lies in the fact that just as the blindfolded people examining the elephant are limited to a single impression, language users are generally limited to one comparison at a time in their expressions. No single description from the blindfolded people is a particularly accurate description of an elephant, since they focus only on a single part of the animal. Importantly, however, the characterisations are not wrong—only incomplete. If the group were to pool their descriptions into a composite, they could get a fairly accurate description of an elephant. Note, however, that we would probably not say that these individual experiences of parts of the elephant actually add up to the full experience of the elephant (cf. section 3.1 and Croft 2007). The same is true of the approach taken here; it should not be understood as component analysis, but as understanding complex and basically irreducible experience by comparison to other kinds of experience.

The approach described here requires that we know which forms are used to express the meanings associated with the dimensions. The association between forms and dimensions can in fact be quantified, and this fact will be used to delimit what forms are considered. The distribution of a form can be represented as values of a binary variable and the co-ordinates on each dimension as values of a continuous variable. The association between a binary and a continuous variable can be measured with the Point-Biserial Correlation (e.g. Kornbrot 2014). A value of 1 represents absolute positive correlation, a value of -1 absolute negative correlation and 0 no correlation. The correlation can also be evaluated for statistical significance, i.e. the likelihood that the distribution of the form in question would come about by chance. The Point-Biserial Correlation of the distribution of forms and dimensions will be used as a heuristic tool to approach the question of what forms are associated with a dimension. To ensure that the markers considered really are connected with the dimensions in question, I have limited the selection to markers with a p-value of less than 0.01—i.e. a chance of less than 1% that the markers’ distribution on the dimension have come about by chance. The exact cut-off point is arbitrary, but seems to work well in regards to neither over- nor under-
including. Additionally, markers are limited to the dimension with which they correlate most strongly (the exception to this is when correlations with the values from alternative CMDS analyses are considered). In the following sections, the markers meeting these criteria in the languages of the full sample are summarised (though note that the correlations concerns the dimensions as established in chapter 7, that is, based on the core language sample). The markers are available in full in the Online Supplementary Materials.

9.3 Knowing, finding out and understanding

In this section, the semantic characterisation of the first two dimensions identified in chapter 7 will be considered again. As was outlined there, there appears to be two main—partially intertwined—semantic parameters underlying Dimension 1, namely the state-achievement distinction and the distinction between ‘know’ and ‘understand’. Furthermore, we saw there that there was a differentiation in how these meanings are expressed linguistically; the state-achievement distinction is associated with constructional differences whereas the ‘understand’–‘know’ distinction is associated with lexical differences. This observation is the basis for the attempt in this section to tease apart these two semantic divisions. Section 9.3.2 characterises the state-achievement division and section 9.3.3 the distinction between ‘know’ and ‘understand’ semantically. Closely connected to the state-achievement division, section 9.3.4 characterises the semantics of the FIND OUT cluster identified as being singled out in Dimension 2. Before the exploration of semantics begins, however, a methodological remark is necessary—this is the topic of section 9.3.1.

9.3.1 A methodological remark

In this section, I will discuss the relationship between the expression of ‘know’ as state vs. as achievement and aspectual distinctions in the rest of the language. As we will see, there are cases where the main means of expressing the distinction between know-as-state and know-as-achievement is a productive part of the language, such as the use of the Perfective in Trinitario-Javeriano-Loretano to express know-as-achievement.

It might seem that allowing for these kinds of forms goes against the methodological principles established in section 3.6.2. Recall that in the final demarcation of what to code for, a criterion was imposed that the distinction expressed was specific to the knowledge domain. With this, things like the standard expression of negation and tense were excluded, as well as the distinction between clausal and nominal objects. Why shouldn’t this also apply to things like perfectives?
The reason for this is methodological. It might, in principle, be of interest for coding purposes to distinguish those languages where we find a dedicated expression of the state-achievement distinction in ‘knowing’ and code only such constructions which are not merely instantiations of more general aspectual constructions in the language. In some cases this is unproblematic. On one end of the scale we have fully productive construction like the Trinitario-Javeriano-Loretano Perfective. On the other end of the extreme, we have constructions clearly limited to the knowledge domain—the clearest examples are when the achievement sense of ‘know’ is expressed by a distinct lexeme, such as Turkish öğrenmek ‘find out, come to know’. But in many cases, the situation is intermediate between these extremes. Consider, for example, Swedish, where we find stative vetta and achievement få vetta. This få ‘get, receive, may’ is not productive in producing achievements across the lexicon but is also not limited to knowledge, and extends to a few contexts of cognition and perception.

While identifying the cases at either extreme would probably be feasible, this is not the case for the many intermediate cases. Some measure of productivity could be applied to the constructions with some cut-off point of what constructions are to be included or not, but this would lie far beyond the bounds of this investigation. The remaining options are to include all aspectual constructions or to exclude all. Since exclusion would mean also excluding several constructions that are without a doubt relevant to the expression of knowledge, I have chosen inclusion. The results in this thesis should be considered in light of this methodological choice. In practice, the inclusion of constructions not specific to the knowledge domain has probably contributed to the position of the state-achievement distinction early in the CMDS analysis.

With these preliminary remarks, we move on to considering the achievement-state division in greater detail.

9.3.2 Achievement and state

As we saw in the section 7.4, there are two main semantic distinctions underlying the first dimension of the CMDS analysis—the distinction between states and achievements, which is mainly realised with constructional markers, and the distinction between ‘know’ and ‘understand’, which is mainly realised lexically (for the details of this division in the coding see section 3.6.2). In this section, forms correlated with a CMDS analysis of only the constructional data will therefore be considered, and in section 9.3.3 the lexical data is considered. Note that although the distribution of lexical forms is not included in the computation of the dimensions here, they are included in the Point-Biserial correlation analysis, which means that also those languages in which we do find lexical forms associated with the state-achievement distinctions can be
identified.

Thus, the forms positively correlated with the first dimension of a CMDS analysis of the constructional data only (cf. section 7.4)—henceforth Constructional Dimension 1—with a p-value of less than 0.01% are considered here. The 76 markers from 51 languages are summarised in tables 9.2 and 9.3. They should be interpreted as follows. Table 9.2 summarises the lexical stems associated with the dimension, according to their inter-domain connections. Lexical stems which occur associated with a dimension may either occur on their own or with various constructional markers, which might in turn show a correlation with the dimension. These constructional markers are presented in table 9.3. To the left, they are classified according to function and to the right according to which lexical forms they combine with (note that these forms may or may not themselves be correlated with the same dimension). The two parts of 9.3 (as well as following tables of the same type) should be read as independent of each other. Note that the combinations sum up to more than the total number of constructional markers if a marker combines with more than one lexeme.

To give some concrete examples, Bengali para ‘do, can’ correlates with Constructional Dimension 1 quite strongly (correlation coefficient 0.77, p=4 × 10^{-11}). Since para is as auxiliary modal verb, it is constructional marker, thus found in table 9.3 under the function “Modal”. As we saw in section 9.3, Bengali para combines with jana ‘know that’, cina ‘know person’ and boojha ‘understand’—see examples in 52. This means it falls under these categories under “Combines with” in table 9.3. These Bengali lexemes, however, are not found among the forms correlated with Constructional Dimension 1—as expected, given their wider meaning described in section 7.4. As another example, consider Yuracaré, where both the lexical item huëshë- ‘hear’ (found under ‘hear’ in table 9.2) and the suffix -të ‘middle voice’ which combines with huëshë- (“Other” and ‘hear’ respectively in table 9.3) correlate with the dimension, to somewhat different degrees. This shows that not only is the form huëshëtë-, roughly ‘realise’, associated with the dimension, so is huëshë- on its own—i.e. it rarely occurs outside of these achievement contexts (recall that the CMDS analysis includes only the extended knowledge domain—not perception).

Several interesting findings can be noted about the forms summarised in tables 9.2 and 9.3. The first is the high frequency of constructional markers with aspectual meaning, most often combining with a general form for ‘know’ (21 aspectual markers, and 41 markers which combine with a form for ‘know’, table 9.3).

As an example of a language where a general aspectual construction is used to express the know-as-achievement meaning, consider Trinitario-Javeriano-
### Table 9.2: Summary of lexical markers positively correlated with Constructional Dimension 1: number of markers (number of languages)

<table>
<thead>
<tr>
<th>Function</th>
<th>Combines with</th>
<th>Number of Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connections</td>
<td></td>
<td></td>
</tr>
<tr>
<td>‘hear’</td>
<td>‘know’</td>
<td>11 (11)</td>
</tr>
<tr>
<td>(Opaque)</td>
<td>‘understand’</td>
<td>10 (10)</td>
</tr>
<tr>
<td>‘see’</td>
<td>‘hear’</td>
<td>3 (3)</td>
</tr>
<tr>
<td>‘say’</td>
<td>‘know’ (person)</td>
<td>2 (2)</td>
</tr>
<tr>
<td>Other</td>
<td>‘see’</td>
<td>3 (3)</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>3 (3)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>29 (27)</strong></td>
</tr>
</tbody>
</table>

### Table 9.3: Summary of constructional markers positively correlated with Constructional Dimension 1: number of markers (number of languages)

<table>
<thead>
<tr>
<th>Function</th>
<th>Combines with</th>
<th>Number of Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspect</td>
<td>‘know’</td>
<td>17 (16)</td>
</tr>
<tr>
<td><strong>Known marker</strong></td>
<td>‘understand’</td>
<td>10 (7)</td>
</tr>
<tr>
<td>Directional</td>
<td>‘hear’</td>
<td>6 (5)</td>
</tr>
<tr>
<td>Modal</td>
<td>‘know (person)’</td>
<td>3 (3)</td>
</tr>
<tr>
<td>Other</td>
<td>‘see’</td>
<td>11 (11)</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>3 (3)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>47 (37)</strong></td>
</tr>
</tbody>
</table>
Loretano. In this language, two main verbs take up the knowledge domain—
*echo-* ‘know (that)’ and *-imoti-* ‘know (person)’. Trinitario-Javeriano-Loretano
also has rich verbal morphology with a large set of (non-obligatory) TAM-
suffixes (Rose 2014: 80). In the verse sample the distinction between state and
achievement is usually expressed by a contrast between a verb unmarked for
TAM (see 58a) and the use of the Perfective morpheme *-po* (58b). To see that
the difference is not just one of tense (present vs. past) consider 58c in which
the verb is marked with the iamitive *-ripo* ‘already’ (perhaps diachronically re-
lated to *-po* but not synchronically) instead of *-po*—here there is past meaning
but not achievement meaning.

(58) Trinitario-Javeriano-Loretano (Arawakan, Bolivia)

a. Mark 12:14

*Tata, v-echo te ’to p-echji-s’o to
father 1PL-know PREP COMP 2SG-speak-APPL.ACT ART
je’chugne
truth

‘Teacher, we know that you speak the truth’ (AS)

b. Acts 19:34

*Etotsero te na-écho-po to jurion-ira-ri’i ma
but PREP 2PL-know-PFV COMP jew-NMLZ-? ART
Alejandro
PR.NAME

‘but when they found out that Alexander was a Jew’ (AS)

c. John 6:6

*tjicho ema mu-echo-ripo to ma-jich-yore
because 3SG.M 3SG.M-know-IAM NMLZ 3SG.M-do-FUT

‘because he knew what he was going to do’

As another example, consider Hungarian, where we find three verb-roots dom-
inating the knowledge domain—*tud-* ‘know (that)’, *ismer-* ‘know (person)’
and *ért-* ‘understand’. All of these combine with the verbal prefixes *meg-

*Rounds (2009)* refers to these as *coverbs*, “verbal prefixes that can be attached to (and
subsequently separated from) verbs. Coverbs modify verbs in a variety of ways, including
indicating the direction of motion, the manner of an action, and the beginning or completion of
an action (aspect). By no means is this the extent of their semantic reach; sometimes coverbs
change the meaning of the verb entirely” (Rounds 2009: 60).
and ismer- in addition combines with fel-. Of the two prefixes, meg- is mostly clearly aspectual in meaning, and “indicate[s] the completion of an action in the past, present or future.” (Rounds 2009: 64). Fel-, on the other hand, is basically spatial, meaning ‘up’ (Rounds 2009: 70). The use of these prefixes is illustrated with ismer- ‘know (person)’ below—in 59a we see the un-prefixed verb expressing a state, in 59b we see the usage with fel- and in 59c the use of meg-.

(59) Hungarian (Uralic, Hungary)

a. John 7:29

Én ismer-em őt, hiszen tőle
1SG.NOM know-PRS.1SG 3SG.ACC since 3SG.ABL
származ-om
come.from-PRS.1SG

‘I know him, because I am from him’

b. Mark 6:54

Amikor ki-száll-tak a hajó-ból, azonnal
When out-get-PST.3PL ART ship-ELA immediately
fel-ismer-t-ék őt
UP-know-PST-3PL.DEF 3SG.ACC

‘And as they were getting out of the boat, people recognized him immediately’

c. Acts 12:14

Amikor meg-ismer-t-e Péter hang-já-t
When PFV-know-PST-3SG.DEF PR.NAME voice-3SG-ACC

‘[And] recognizing Peter’s voice’

There are two further important points that can be made regarding the Hungarian examples here. Firstly, that we also here see an example of the difference in lexical and constructional differentiation—lexically, ért- clearly singles out the ‘understand’ region, whereas the prefixes meg- and fel- are oriented toward the state-achievement distinction. The second point is the use of a fundamentally spatial form to create aspectual meaning, i.e. fel- ‘up’. This is found in several other languages of the sample (“Directional” in table 9.3). We can understand the use of general aspectual markers here in view of Croft’s theory of aspectual construal introduced in section 2.2.7. Different
aspectual constructions interact with lexical semantics to impose a particular aspectual construal on an event. This is illustrated, using the schemas of Croft (2012) introduced in 2.2.7, with Trinitario-Javeriano-Loretano as an example in 9.1. The use of -echo- unmarked for TAM profiles the know-phase of the event, whereas the use of the same verb but marked with the Perfective -po profiles the shift from not knowing to knowing.

As another example of the second point above—the use of spatial forms—consider Teso. Here the main form in the knowledge domain is -jen-. This verb may take the suffix -u(n) which the grammar describes as a ventive marker, i.e. a marker expressing motion toward the speaker (Barasa 2017: 160–163). The extended use of -u(n) and its itive (motion away from the speaker) counterpart -a(r) as providing evidential information, by placing the speaker relative to the reported event and thus as a witness is discussed in Barasa (2017). No other semantic extensions are discussed there, but in a grammar of the closely related language Turkana, Dimmendaal (1983: 104) notes that “the semantics of the process [=derivation with the ventive and similar suffixes] is not always transparent” and gives examples of aspectual meaning resulting from the Turkana cognate to -u(n). Teso examples are given in 60.

(60) Teso (Nilotic, Uganda)

a. Mark 12:14

\[
\text{ki-jen-i ebe } i-ra-i \text{ ijo lok-abeit}
\]

1PL-know-IMPF COMP 2-be-IMPF 2SG ?-truth

‘we know that you are truthful’
b. Mark 6:54

\[ Na-ke-dok-un-ia-ta, \quad atipet \quad ko-jen-u-t \]
when/at-climb-VENT-PST-PL immediately DEF-know-VENT-?PL
\[ it\eta \quad yes \]
people 3SG

‘And as they were getting out of the boat, people recognized him immediately’

It would appear that we are dealing here with some kind of metaphorical extension, wherein the dynamic aspect of know-as-achievement is understood as movement (cf. below). Perhaps there is also an element of the metaphor KNOWING IS IDEAS ENTERING THE MIND, discussed in more detail in section 9.3.3—this would explain why it is generally ventive markers found expressing this.

There is also a strong connection between know-as-achievement and perception, in particular hearing, in the relevant forms. Two types of connection can be distinguished. Firstly, where a form for ‘hear’ (or more rarely ‘see’) is used also for know-as-achievement and thus correlates with the Constructionsal Dimension 1 (‘hear’ with 11 markers and ‘see’ with 3 in table 9.2), and secondly where the form for ‘hear’ is the basis of a separate construction expressing know-as-achievement, which corresponds to the correlation of a constructional marker with Constructional Dimension 1 (‘hear’ and ‘see’ in table 9.3). Of these, the first is the most common. This phenomenon—perception for knowledge—is discussed in more detail in chapter 10, but an example is given in 61.

(61) North Tanna (Austronesian, Vanuatu)


\[ Moto \ e \ nian \ net\text{\textcopyright{\textregistered}} \ mim \ k-\text{\textcopyright{\textregistered}}-ot-o \to \ namnus\text{\textcopyright{\textregistered}} \ m\text{\textcopyright{\textregistered}} \]
but PREP day people 3PL-PST-PL-hear news COMP
\[ Iesu \ \ t-in-uwa \]
PR.NAME 3SG-PF-COME

‘but when the people found out that Jesus had come’ (AS)

Among the language where a construction for know-as-achievement involves only perception verbs and not ‘know’ forms, there is San Blas Kuna, where we find the construction mag/aku dake-/ito- ‘can/cannot see/hear’ which means ‘to (not) be able to see/hear’ but also ‘find out’ (with ‘hear’) and ‘recognise’ (with ‘see’). Particularly interesting is that the forms mag and aku are not the
general modality forms in the language, but are limited to perception verbs. A grammar of the closely related language Paya Kuna mentions that the same is true of that language (Forster 2011). This is interesting in view of the fact that in many contexts see and can see—and similarly hear and can hear—are basically equivalent (Wälchli 2016: 71), an argument also made by Vendler (1957: 148).

(62) San Blas Kuna (Chibchan, Panama)
   a. Mark 8:23
      Emigindi sunna mag be dake
today truly can 2 see
‘Can you see now?’ (AS)
   b. Luke 24:31
      Geb degi e=war-bogwad mag=Jesús=dak-de-mala-d
      CONN CONN DEM-CL:LONG-two can=PR.NAME=see-INCH-PL-TVF
      ‘and they [two] recognized [Jesus]’
   c. 1 Cor 14:16
      Ar aku be ito-ge-d, ibi be sog-sa
      because cannot 2 hear?-TVF what 2 say-PFV
      ‘because you do not understand what you are saying’ or ‘because he cannot hear what you are saying’ (AS)

In 62a the modal sense of mag is evident, but in 62b the meaning is best understood as achievement; note that we are dealing with two developments here—from modality to aspect and from perception to cognition. Example 62c offers an interesting candidate for what bridging contexts might have looked like—an extension of the ‘not able to hear’ to ‘not able to understand’.

The close connection between perception and know-as-achievement is also visible in the combination of forms for perception and knowledge to express know-as-achievement. An example is Naro. The general word for ‘know’ in this language is q’äa. This form combines with the form for ‘see’ to express a general know-as-achievement meaning (see 63a). The same construction but with ‘hear’ rather than ‘see’ is used for ‘understand’ (63b).
(63) Naro (Khoe-Kwadi, Botswana)


\[
A \ ba \ a \ ēem \ ko \ bóò-a \ q'āa
\]

and 3SG.M DEM when PTCP see-CONN know

‘And when he found out...’

b. Mark 14:68

\[
C'úùa-r \ hāa, \ a \ ra \ a \ kó-a \ q’āa \ tama
\]

not.know-PFV IPFV and 1SG DEM hear-CONN know NEG

‘I neither know nor understand’

In the case of Naro, this is clearly a conventionalised construction, as evidenced by it being negated with standard negation rather than lexical negation (the form c’úùa ‘not know’), as demonstrated in (63b).

The combination of perception and cognition to express achievement ‘know’ is also found in what appears to be less conventionalised constructions. As an example, consider the Meyah verses in (64). In these verses, the situation of coming-to-know-through-perception is explicitly decomposed into a perception component and a cognition component.

(64) Meyah (East Bird’s Head, Indonesia)

a. Acts 12:14

\[
Beda \ ofa \ eg \ oga \ egens \ koma. \ Beda \ ofa \ osujo hu \ oida. \ DEM \ hear \ voice \ INDF \ DEM \ then \ 3SG \ think \ COMP \ koma \ bera \ erek \ Pet rus \ oga. \ DEM \ TOP \ COP \ PR.NAME \ voice
\]

‘And recognizing Peter’s voice’  
Lit.: ‘then she heard a voice. Then she thought/realised that it was Peter’s voice’ (AS)

b. Mark 6:54

\[
beda \ rusnom \ jah \ suma \ rik \ rua \ noba \ rijgin ag a \ rot. \ COMP \ 3SG \ TOP \ PR.NAME \ and \ 3SG.POSS \ disciples
\]

‘people recognized him immediately’

Lit.: ‘then the people from there saw them and knew that they were Jesus and his disciples’ (AS)
There are also some clues as to the semantics of the achievement-state divide in the forms negatively correlated with the value of Constructional Dimension 1, i.e. those associated with the low end of the dimension. An overview is given in tables 9.4 and 9.5—79 markers from 56 languages are negatively correlated with the first dimension of the constructional CMDS analysis with a p-value of less than 0.01. The majority of forms are what we might call general ‘know’-predicates. They are generally transitive verbs, whose etymology is not synchronically apparent (“(Opaque)” in table 9.4).

There are, however, a few forms which are clearly aspectual (“Aspectual” in table 9.5). We find, for example, the Mapudungun (Araucanian, Chile) suffix -nie which combines with the verb for ‘know’ kim- to give the stative sense. The suffix -nie is analysed in Smeets (2008) as a ‘progressive persistent’ and is described as “[—] denotes an ongoing event which may or may not continue into the present. [—] [T]he morpheme -nie- [—] indicate[s] that a state, whether an ongoing event or the result of an event, persists” (Smeets 2008: 295). An example is given in 65a. When used without -nie, kim- has an achievement meaning, as in 65b (the -lu suffix marks the clause as a temporal adverbial clause).
(65) Mapudungun (Araucanian, Chile)

a. Smeets (2008: 194)

kim-nie-ke-n  kuyfí  ňi  nge-nu-n
know-PRS.PROG-CF-1SG.IND  formerly  3POSS  be-NEG-PVN
traktor
tractor

‘I know that in the old days there were no tractors’

b. Luke 9:11

Welu  ti  pu  che  kim-lu  Jesus  ňi  amu-n
but  DEF  COLL  person  know-TEMP  PR.NAME  3POSS  go-NMLZ
kañpüle
other.side

‘But when the crowds found out [that Jesus was going there]’

Another interesting type of form correlated with Constructional Dimension 1 are nouns for ‘knowledge’ and used in some construction to predicate knowledge as opposed to the cross-linguistically more common verbs. As an example, consider Yele. The general construction for stative ‘know’ in this language is as a locative metaphor, where what is known “sits” or “stands” in the subject’s knowledge. Both ‘stand’ and ‘sit’ are quite general verbs in Yele, and their distribution is determined by the semantics of the subject—examples of this construction are given in 66. In the verse sample, this construction is limited to stative ‘know’, and the achievement sense is expressed either with distinct lexical items or perception forms.

(66) Yele (Isolate, Papua New Guinea)


Ndêndê  ngê  u  lama  dê-noo  ya
true  M  his knowledge  NEG-1SG  sitting

‘He really didn’t know me’
Lit.: ‘I wasn’t sitting in his knowledge’


Post  Office  a  lama  daa  kwoo
post  office  my  knowledge  3.NEG  standing

‘I don’t know where the post office is
Lit.: ‘The post office isn’t standing in my knowledge’
Another example is Daga (Dagan, Papua New Guinea) where one of the constructions for knowledge predication is the possession of the noun asigi- ‘knowledge’. This too is largely limited to stative meanings. The occurrence of nominal forms in the state-sense of ‘know’ can be understood in relation to the semantics often argued to underlie parts of speech. For example, Givón (2001a: 50) claims that the foremost defining semantic feature distinguishing word classes is time stability. Prototypical nouns are very stable over time whereas prototypical verbs have short duration. Time-stability is also an important part of Croft’s approach to parts of speech, which are defined in part through the three semantic classes OBJECTS, PROPERTIES and ACTIONS. These in turn are defined through the four features Relationality, Stativity, Transitoriness and Gradability (Croft 2001: 87).

Although ‘know’ is most often expressed as a verb cross-linguistically, it actually deviates from the prototypical action in several ways. This is shown in table 9.6 where the distribution of Croft’s defining features over the three classes (from Croft 2001 87) are shown, with ‘know’ added, showing how it differs from the three prototypes. It is a state, similarly to Objects and Properties, and not a process, as Actions. Additionally, it is—at least arguably—gradable, similarly to Properties rather than Objects and Actions. However, KNOW shares with Actions and Properties the very important property of being relational; that is, it cannot be conceived of without reference to some other entity. In this it differs from Objects, which are non-relational and can be conceived of independently.

These semantic classes are one fundament of Croft’s theory of parts of speech, the other being propositional acts. What is of interest here, however, is that we cross-linguistically find forms of knowledge predication which reflect the fact that knowing is not a prototypical Action but is in some ways more like an Object or a Property.

To summarise, then, one of the fundamental divisions in the knowledge domain is between know-as-achievement and know-as-state. This division interacts with the distinction between ‘understand’ and ‘know’, but whereas the latter distinction is commonly made lexically the achievement-state distinction

<table>
<thead>
<tr>
<th>Class</th>
<th>Relationality</th>
<th>Stativity</th>
<th>Transitoriness</th>
<th>Gradability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objects</td>
<td>nonrelational</td>
<td>state</td>
<td>permanent</td>
<td>nongradable</td>
</tr>
<tr>
<td>Properties</td>
<td>relational</td>
<td>state</td>
<td>permanent</td>
<td>gradable</td>
</tr>
<tr>
<td>Actions</td>
<td>relational</td>
<td>process</td>
<td>transitory</td>
<td>nongradable</td>
</tr>
<tr>
<td>‘know’</td>
<td>relational</td>
<td>state</td>
<td>transitory</td>
<td>?gradable</td>
</tr>
</tbody>
</table>

Table 9.6: Features of the three types of entities according to Croft (2001) with ‘know’ added
is commonly made constructionally. This allows for teasing apart the two main aspects of Dimension 1 by applying CMDS to the constructional data and the lexical data separately. In this section, the forms correlated with the constructional analysis were examined, revealing the following about the semantics of know-as-achievement versus know-as-state.

**Know-as-state** is a largely opaque concept—it generally shows no synchronically clear connection to other concepts. It is generally not understood in terms of other concepts. The constructional forms found are mainly aspectual forms combining with an opaque root for ‘know’. There are some exceptions, and we see examples of know-as-state being expressed both in terms of seeing and hearing. Among forms associated with know-as-state nouns (or sometimes adjectives) with a meaning like ‘knowledge’ deserve special mention in that they provide evidence for the conceptualisation of knowing not as an event but as something thing-like. Particularly interesting is the Yele example where knowledge is expressed through a location metaphor.

**Know-as-achievement** is often understood in terms of know-as-state. This can be understood in terms of aspectual construal, where the same event can be construed in different ways by profiling different parts of the event’s internal timeline. However, the expression of know-as-achievement is not limited to this—it is also strongly connected with perception, in particular ‘hear’. Furthermore, to know-as-achievement has a dynamic component to it, as evidenced by the use of directional markers which would appear to be the expressions of some kind of metaphors of movement.

### 9.3.3 Understanding

In this section the semantics of ‘understand’ will be investigated—what I suggested as the other parameter, in addition to the state-achievement division, underlying Dimension 1. In the previous section, the distribution of forms vis-à-vis the first dimension of a CMDS analysis based only on constructional markers was considered. In this section, the distribution of forms vis-à-vis the remainder of the variation will be considered, i.e. the first dimension of a CMDS analysis based only on lexical data. I will refer to this as the Lexical Dimension 1. This allows us to tease apart the two major signals in Dimension 1 of the complete CMDS analysis—state-achievement and ‘know’—‘understand’. Note that as with the analysis in section [9.3.2](#) while the distances used for the CMDS analysis are limited to either only lexical or constructional data, the forms investigated for correlation are not.

A summary of the forms positively correlating with the first dimension of the CMDS analysis of the lexical-based distances is given in tables [9.7](#) and [9.8](#)—113 forms from 68 languages.
### Table 9.7: Summary of lexical markers positively correlated with Lexical Dimension 1: number of markers (number of languages)

<table>
<thead>
<tr>
<th>Function/meaning</th>
<th>Combines with</th>
<th>Total: 68 (60)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connections</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Opaque)</td>
<td>28 (26)</td>
<td></td>
</tr>
<tr>
<td>Motion</td>
<td>14 (14)</td>
<td></td>
</tr>
<tr>
<td>‘hear’</td>
<td>12 (23)</td>
<td></td>
</tr>
<tr>
<td>Loan</td>
<td>4 (4)</td>
<td></td>
</tr>
<tr>
<td>‘think’/‘remember’</td>
<td>3 (3)</td>
<td></td>
</tr>
<tr>
<td>‘light’/‘clear’</td>
<td>2 (2)</td>
<td></td>
</tr>
<tr>
<td>‘meaning’</td>
<td>1 (1)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>4 (4)</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>68 (60)</strong></td>
<td></td>
</tr>
</tbody>
</table>

### Table 9.8: Summary of constructional markers positively correlated with Lexical Dimension 1: number of markers (number of languages)

<table>
<thead>
<tr>
<th>Function/meaning</th>
<th>Combines with</th>
<th>Total: 45 (34)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Body-part</td>
<td>8 (8)</td>
<td>‘know’ 14 (13)</td>
</tr>
<tr>
<td>Aspectual</td>
<td>5 (5)</td>
<td>‘understand’ 10 (10)</td>
</tr>
<tr>
<td>Auxiliary</td>
<td>5 (3)</td>
<td>‘hear’ 9 (7)</td>
</tr>
<tr>
<td>Emphasis</td>
<td>4 (4)</td>
<td>Motion 8 (7)</td>
</tr>
<tr>
<td>Directional</td>
<td>2 (2)</td>
<td>‘meaning’ 3 (1)</td>
</tr>
<tr>
<td>‘meaning, word’</td>
<td>2 (2)</td>
<td>‘see’ 3 (3)</td>
</tr>
<tr>
<td>Knower marker</td>
<td>1 (1)</td>
<td>Other 9 (8)</td>
</tr>
<tr>
<td>Known marker</td>
<td>1 (1)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>21 (18)</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>45 (34)</strong></td>
<td></td>
</tr>
</tbody>
</table>
To begin with, it can be noted that several languages express ‘understand’ by a combination of the form for ‘know’ and some constructional form. Note that this in fact underestimates the close relationship between ‘know’ and ‘understand’; as we will see in chapter [11] in many languages the form for ‘know’ is simply extended also to ‘understand’. In these cases, no marker will be associated with high values in Lexical Dimension 1. As an example of a language in which the form for ‘know’ is the basis for ‘understand’, consider Southern Kisi where the general word for ‘know’ is *sina* (example 67a). This form is also used for ‘understand’, but then with what Childs (1995) labels an ideophone *pèygù* (example 67b, see also example 57 on p. 168).

(67) Southern Kisi (Atlantic-Congo>Mel, Sierra Leone)

a. Mark 12:14

\[ Títchá, i já síná māā à chō sákpō \]
\[ teacher(Eng.), 1PL know COMP 2SG be righteous/truthful \]

‘Teacher, we know that you are truthful’

b. Mark 14:68

\[ í siñá [nyè i já chō dímíóó wó] pèygù lè \]
\[ 1SG know thing 2SG AUX speaking REL IDEOPH NEG \]

‘I do not understand what you are talking about’ (AS)

Another example is the Koman language Uduk, where ‘understand’ is expressed with what Killian (2015: 272) terms partargument constructions: “Uduk makes frequent use of Partarguments, a term coined here to refer to the satellites of a complex predicate construction. Partarguments are typically body parts incorporated into the verbal predicate which changes the basic meaning of the verb, or classifies the argument of the verb.” To express ‘understand’ in Uduk the general word for ‘know’ *mish-* may occur together with ‘*ban* ‘temple’, gwo ‘word’ or both.

(68) Uduk (Koman, Ethiopia)

a. Acts 8:30

\[ Gwo jin deyi /e e mo tani mish-a /e ’ban \]
\[ word REL read 2SG ? PRT PHR.MRK know-? 2SG TEMPLE \]
\[ mo ’taa? \]
\[ PRT Q \]

‘do you understand what you are reading?’
In addition to the connection to ‘know’, the Uduk forms give two further insights into the conceptualisation of understanding—understanding has to do with what we can assume is the seat of the intellect, i.e. ‘temple’ ‘ban as well as with ‘word’ or ‘meaning’ gwo. Both types of connections will be addressed below.

Moving on to other concepts which are the basis for forms for ‘understand’, ‘hear’ is the most common. We find a number of verb stems meaning ‘hear’ associated with with this meaning (table 9.7), as well as a number of constructional markers which combine with forms meaning ‘hear’ (9.8). This corresponds to two strategies of expression—full and partial colexification of ‘hear’ and ‘find out’. In some languages, it seems that both strategies are possible. In Murui Huitoto, for example, ‘hear’ is caca-, and this is also used to mean ‘understand’, with or without the morpheme -i, which Wojtylak (2017: 318) labels emphatic (“Emphasis” in table 9.8). Note that the position of this morpheme close to the verb stem suggests a derivational function. In the verse sample, this -i occurs only in the understand sense and not the ‘hear’ sense. The uses are illustrated in example 69. In 69a, the ‘hear’ sense is contrasted with the ‘understand’ sense, which is marked with -i. In 69b, however, the meaning is clearly ‘understand’, but the verb does not receive the emphatic marker.

(69) Murui Huitoto (Huitotoan, Colombia)

a. Matt 13:19

\[\text{Jusiñamui-die comini-na seda-ja}\]
\[\text{God-S/A.TOP Conn people-NON.S/A.TOP take.care-NMLZ}\]
\[\text{rafue-na caca-di-maci ia-dedi ie-na}\]
\[\text{story-NON.S/A.TOP hear-LNK-3PL CONN-? CONN-NON.S/A.TOP}\]
\[\text{caca-iñe-di-maci}\]
\[\text{hear-EMPH-NEG-LNK-3PL}\]

‘[they] hear[] the story of God’s kingdom but do not understand it’

b. Matt 13:51

\[\text{¿Nana cue llua-na caca-d-omoi?}\]
\[\text{all 1SG tell-NMLZ hear-LNK-2PL}\]

‘Do you understand what I am telling you?’ (AS)

In some languages, the connection between ‘understand’ and ‘hear’ is evident only in the specific context of understanding a language. In a number of languages in the sample, this is the only cognitive context where we find ‘hear’. An example is Carib, where the form for ‘know’ -uku- covers the entire ex-
tended knowledge domain (i.e. ‘know’, ‘understand’ and ‘find out’) with or without achievement marking. The exception is one context where the object of the relation is a language and -eta- ‘hear’ is used. Note that in this example the meaning appears to extend to production (the question is from a soldier to Paul in response to Paul addressing him and asking for permission for speak, presumably in Greek)—more commonly, we find this use regarding the comprehension of language. This is interesting in light of Sweetser’s observation that understanding concerns in particular the communicative (see section 2.2.5).

(70) Carib (Cariban, Suriname)

a. Acts 21:37

Griek auran ka’tu m-eta-no-n
Greek language maybe 2-hearPRS-UNCERTAINTY

‘Do you know Greek?’

Note also the Naro example above (63) where forms for ‘know’ and ‘hear’ combine to express the meaning ‘understand’.

‘Understand’, then, shows a close connection to hearing. This is not surprising given previous research on the topic; already Sweetser (1990) discussed this relationship between ‘hear’ and ‘understand’, the connection between cognition and hearing was further developed by Evans & Wilkins (2000) and the recent large-scale work from Georgakopoulos et al. (2022) show similar results.

It seems that an important part of how ‘understand’ comes to be construed in terms of hearing is shared objects—that is, that the things which are typically understood are also typically heard. We saw this above regarding languages (which are typically both heard and understood) but there is also a connection in the languages of the sample between ‘understand’ and forms for ‘truth’, ‘meaning’ or more generally ‘words’ (‘meaning’ in table 9.8). This was seen above with Uduk, in which mish- ‘know’ may occur with so called part-arguments to expressed the meaning ‘understand’: ban ‘temple’, gwo ‘word (meaning, name)’ or both.

Another interesting example is Kannada in which we find two main forms for ‘understand’, both based on the root artha- ‘meaning, sense’ (this is the one marker under ‘meaning’ in table 9.7). One takes the understander as an argument in the dative—artha-āgu- ‘meaning-become’ (example 71a). The other, more common, form is arthamādikōṣṣu, consisting of artha, mādu ‘do, make’ and the auxiliary koṣṣu, where the subjects is nominative (example 71b).
Kannada (Dravidian, India)

a. Acts 8:30

\[
nínu ōduttirùṣudu ninage artha-ṣ-āguttide-yô
\]

2SG reading 2SG.DAT meaning-EP-become-Q

‘do you understand what your are reading?’

b. Matt 16:12

\[
āga ziqyaru yēsu hēsiddannu
\]

then disciple PR.NAME say.NMLZ.ACC

\[
artha-māḍi-koṇḍaru
\]

meaning-make-AUX.PST.3PL

‘Then the disciples understood what Jesus had said’ (AS)

In addition to facilitating a connection with hearing, these forms seem to reflect the fact that one aspect of the meaning of ‘understand’ lies in the object of the cognitive relation. One of the components of ‘understanding’ suggested in section 2.2.5 was that it takes as its objects complex things and concerns the “substantive qualities” of these. It seems that the use of a root which means ‘meaning’ is a reflection of this. Kannada is perhaps the clearest example, in which the predicate itself is based on the meaning-root, but there are other examples in the sample. Already mentioned is Uduk with gwo ‘word’. Another example is ehwi ‘meaning’ which combines with eri ‘hear’ to mean ‘understand’ in Ma’di (Central Sudanic, South Sudan).

With the rise of literacy, the connection between what is meaningful, hearing and understanding might perhaps be thought to be weakening in favor of the connection between sight and understanding through reading (cf. Evans & Wilkins 2000) but this does not seem to be reflected in the language of the NT. Note that the verse sample does contain a context where a man is asked if he ‘understands what he is reading’ (Acts 8:30)—forms for ‘see’ are not used here, but form for ‘hear’ occasionally are.

In contrast to ‘hear’, there is only one clear example of ‘see’ being extended to ‘understand’, namely in Comaltepec Chinantec, where verbs are either animate or inanimate, depending on the animacy of the subject (in the case of intransitive verbs) or object (in the case of transitive verbs) they take. In many cases, an animate form may be derived from an inanimate form by the addition of a final -ng (in the orthography of the NT). The inanimate form for ‘see’ is ngá-, and the animate form predictably ngáng-. It is this latter form which is also used to mean ‘understand’. This appears to be a fossilized
form and occurs with inanimate objects, see example 72 where \( júu^\) ‘word’ is inanimate, as is explicitly marked by the plural and definiteness markers (see also Anderson 2018: 20). Particularly interesting about this form is that it appears to suggest that understanding is something directed toward animates, presumably humans. This—although it is reflected by the use of ‘see’ rather than ‘hear’—would seem to be a reflection of the intersubjective nature of understanding pointed out by e.g. Sweetser (1990).

(72) Comaltepec Chinantec (Otomanguean, Mexico)

a. John 13:17

\[
\begin{array}{l}
\text{Jo} \quad \text{song} \quad \text{ngám}^-\text{b-a'} \quad \text{jalé''} \quad e \quad \text{júu}^\quad \\
\text{SEQ} \quad \text{if} \quad \text{see.AN.OBJ-EMPH-2SG} \quad \text{PL.INAN} \quad \text{DEF.INAN} \quad \text{word} \\
\text{na} \quad \text{DEM} \\
\end{array}
\]

‘If you understand these things [words]’

Perhaps also belonging to the conceptualisation of understanding in terms of sight are the languages where it is expressed by forms deriving from words for ‘light’ or ‘clear’ (‘light’/‘clear’ in table 9.7). An example is Mandarin Chinese where the form \( \text{ming}^2\text{bai} \) is used for ‘understand’ while also having meanings like ‘clear, open, obvious’. Fortescue (2001: 21) gives the translation ‘bright white’ and classifies the form as an example of the metaphor ‘understanding as (seeing the) light’. An example is given in 73.

(73) Mandarin Chinese (Sino-Tibetan, China)

a. Mark 9:32

\[
\begin{array}{l}
\text{men}^2\text{tu}^3 \quad \text{que}^4 \quad \text{bu}^4 \quad \text{ming}^2\text{bai} \quad \text{zhe}^4 \quad \text{hua}^4 \\
\text{disciple} \quad \text{but} \quad \text{NEG} \quad \text{understand} \quad \text{DEM} \quad \text{word} \\
\end{array}
\]

‘But [the disciples] did not understand the statement’

We find in the sample also three forms which colexify ‘understand’ with non-factive cognition, i.e. ‘think’ and in some cases ‘remember’. One of them is in Aymara, where the form \( \text{amuya}^- \) is used for both ‘think’ and ‘understand’. Interestingly, however, the two senses appear to be distinguished by which complementation strategy is used. When \( \text{amuya}^- \) is used with the meaning ‘understand’, the complement is nominalised and marked with a possessive suffix, which is one of the two main complementation strategies used for

\[\text{Note that the -ng assimilates to the following emphatic marker to -m}\]
‘know’ (example (74a). When *amuya-* is used for ‘think’ the complement is instead marked with *ukham*, roughly ‘like that’ (example (74b).

(74) Central Aymara (Aymaran, Bolivia)

a. Matt 17:13

_Ukapachaw Jesus-an yatiqiri-naka-pax Bautisiri_

then PR.NAME-POSS disciple-PL-SENT.SUFF PR.NAME

_Juanat parla-s-ka-ta-p_

PR.NAME talk-CONT-ASP-NMLZ-3POSS

*amuya-p-x-äna*

understand-PL-COMPL-3>3NEAR.REM

‘Then the disciples understood that he had spoken to them about John the Baptist’

b. John 11:13

_yatiqiri-nakapa-sti samart’añ ikiitsa parl-kaspa ukham_

disciple-PL-CONT rest sleep speak-? like.that

*amuya-p-x-äna*

think-PL-COMPL-3>3NEAR.REM

‘[the disciples] thought that he was speaking about real sleep’

The connection with thinking appears to reflect the distinction between internal and external know-as-achievement. Whereas some know-as-achievement events are clearly grounded in and thus understandable in terms of perception (such as hearing for finding out), understanding to some degree lacks this external connection and is instead grounded in internal processes—i.e. similar to thinking. In section 9.5 below, we will see that this connection is even more characteristic of ‘realise’.

Continuing on to other forms associated with the ‘understand’ contexts, expressions clearly grounded in motion metaphors are quite common (“Motion”, with 14 markers, in table 9.7). Three main types can be distinguished: understanding as coming upon/standing before (cf. Fortescue 2001: 21), understanding as grasping (again, compare Fortescue 2001: 21) and understanding as ideas entering the person/mind. The latter two forms of expression can be generalised as a combination of the metaphors IDEAS ARE OBJECTS and MENTAL ACTIVITY IS MANIPULATION (e.g. Lakoff & Johnson 1980, Jäkel 1995).

For the first type of metaphor—understanding is coming upon—the only examples in the data are from Germanic languages: a form for ‘stand’ with a
prefix cognate with either under- (English and the English-based creole San Andres Creole English) or for- (Swedish, Afrikaans and Pennsylvania German). Fortescue (2001: 43) suggests that the two types of forms may be linked by a form analogous to Danish underforstå with the historical meaning ‘recognise, understand as implied’.

Considerably more widespread in the sample is the metaphor of IDEAS AS OBJECTS which are either grasped or enter (some part of) the person understanding. Both models can be subsumed under the metaphor Jäkel (1995) terms UNDERSTANDING AN IDEA IS ESTABLISHING PHYSICAL CLOSENESS.

Note that the latter metaphor naturally places the cogniser as a non-subject participant in the clause, since the actor role is taken by the idea which is understood. This is clearly the case in Wolaytta, in which one of the ways to express ‘understand’ is with gel- ‘enter’, with the understood as the subject (note that the nominative nominalizer is used) and the person(s) understanding marked with a dative case-marker (example 75). It is also this metaphor which seems to be reflected in the marker associated with the knower correlating with Lexical Dimension 1 (“Knower marker” in table 9.8); the Wolaytta Dative is the marker in question.

(75) Wolaytta (Ta-Ne-Omotic, Ethiopia)

a. Mark 9:32

\\| SHin  eta-u  i  g-iyoo-gee \\
but  3PL-DAT  3SG.M.NOM  say-REL.IPF.NONSBJ-NMLZ.NOM

\\| gel-ibeenna \\
enter-NEG.PFV.3M.SG

‘But they did not understand the statement’

In the Wolaytta case case both verb indexing and overt argument marking make it clear that the understood thing is the subject, but this is not always as clear. In some cases it appears that there may be a reanalysis, and what appears to originally be a metaphor of ideas entering take on the argument structure properties of other cognition verbs with the knower as subject. An example is Thai, where the form for ‘understand’ is khâw-cay ‘enter heart’. The most likely analysis is that this is originally the same kind of metaphor as seen in Wolaytta, i.e. an idea entering the container that is the mind (in this case the ‘heart’). In Thai, grammatical relations are marked neither by nominal marking or verb indexing, so the only clue to subjecthood is word-order. Since

---

1In Kriol, the same form, sabi covers both ‘know’ and ‘understand’.
Thai is an SVO language, this would make the experiencer the subject—see example [76]

(76) Thai (Tai-Kadai, Thailand)

a. Acts 8:30

\[\text{th}â\text{an k}hâ\text{w-cay si}ŋ thîi thâ\text{an ?}â\text{an r}û\text{u} \]
\[2\text{HON enter-heart thing REL 2HON read Q} \]
‘Do you understand what you are reading?’

The metaphor \text{UNDERSTANDING IS GRASPING} is somewhat less common in the sample than \text{UNDERSTANDING IS IDEAS ENTERING THE MIND} but is attested in Eurasia and Africa. As an example, consider Igbo, where the verb \text{-gho- ‘catch’} is used with the suffix \text{-ta} for ‘understand’, see example [77]. The characterisation of the suffix \text{-ta} is not clear at this point, but it appears to have aspectual properties and combines also with \text{-ma- ‘know’}.

(77) Igbo (Atlantic-Congo, Nigeria)

a. Mark 9:32

\[Ma ha a-\text{gho-ta-ghi okwu ahu} \]
\[\text{but 3PL NEG-catch-ASP-NEG word DEF} \]
‘But they did not understand the statement’

What does the frequent use of motion metaphors tell us about the semantics of ‘understanding’? First of all, it seems that the metaphor suggested by \text{Jäkel (1995)} as central to our understanding of cognition—\text{UNDERSTANDING AN IDEA IS ESTABLISHING PHYSICAL CLOSENESSE}—finds considerable support cross-linguistically. Also the Germanic metaphor of standing under or before would seem to be explainable in these terms. The dynamic character of most of the metaphors (entering, grasping etc.) is also evidence of the achievement-semantics of ‘understand’.

In several of the languages making use of entering metaphors, the part of the experiencer which is viewed as being entered by the idea is explicit in the expression (“Body-part” in table 9.8). An example is Thai, as seen above in [76] where ‘heart’ is the location of understanding. As was seen with the case of Uduk, body-parts sometimes also enter into other types of expressions (68)—in that case the temple, perhaps as a metonym for the head. Other body-parts associated with ‘understanding’ are the belly, the liver, the ear and the intelligence, attention or mind (these last are perhaps properly not body-parts,
but person-parts). This offers further evidence of ‘understand’ having to do with the internal. Perhaps this metaphor of entering can also be seen as an expression of ‘understanding’ as connecting new ideas with old information (see section 2.2.5)—where the new ideas are understood as objects coming into contact with a person’s old ideas which are conceived of as contained within some part of the person.

To summarise what the forms found cross-linguistically can reveal about how understanding is conceptualised, then, to understand something is like both knowing and hearing, but includes something more. Part of the reason that understanding is like hearing is that the two concern the same kinds of objects—such as words and languages. These objects are themselves complex and concern the important features of things—meaning, truth etc. However, understanding is also like seeing, or like the things you see being brightly and clearly lit. Understanding is also like thinking in that it is based on internal processing rather than being caused by external stimuli. There is an intersubjective component to understanding—understanding has to do with understanding other people. Understanding is dynamic, both in the sense that it is often seen as an achievement and that it is understood as movement—either of the person understanding or of the things to be understood, which are like objects, into the container which contains the ideas and knowledge of the person understanding. To have understood something is to have incorporated an idea into your knowledge-container.

It can be noted that many of the features described above agree with the characterisations of ‘understand’ discussed in sections 2.2.4 and 2.2.5. Understanding is similar to knowing, but includes something more—there is a depth to it. Furthermore, the component of Apresjan’s (2000) explication allowing understanding to have its basis in imagination is echoed in the connections between forms for understanding and thinking. Also of interest is the component of Boguslavsky’s (1984) explication concerning substantive features—we can see how this is oriented toward cross-linguistically by forms for concepts like ‘truth’ or ‘meaning’ being part of ‘understand’ forms. Finally, also the intersubjective component of understanding appears to be reflected cross-linguistically.

9.3.4 Finding out

Having considered Dimension 1 from the lexical and constructional perspective, we now move on to the Dimension 2. Recall that much of the similarity between the FIND OUT cluster and the UNDERSTAND cluster was due to constructional markers—these were considered in section 9.3.2 In this section, the investigation will therefore be limited to those markers which cor-
Table 9.9: Summary of lexical markers positively correlated with Dimension 2: number of markers (number of languages)

<table>
<thead>
<tr>
<th>Function/meaning</th>
<th>Combines with</th>
<th>Known marker 8 (8)</th>
<th>‘know’ 8 (8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspectual</td>
<td>‘hear’ 7 (7)</td>
<td>1 (1)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Modal</td>
<td>Other 5 (5)</td>
<td>1 (1)</td>
<td>4 (4)</td>
</tr>
<tr>
<td>Directional</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td></td>
<td>14 (14)</td>
<td></td>
</tr>
</tbody>
</table>

Table 9.10: Summary of constructional markers positively correlated with Dimension 2: number of markers (number of languages)

relate with Dimension 2 but were not addressed there—that is, those forms which can be assumed to be associated specifically with the FIND OUT cluster and not with know-as-achievement in general. In this way, this section will investigate what semantically characterises the ‘find out’ contexts.

In tables 9.9 and 9.10 a summary of the forms correlating with Dimension 2, but which were not considered in 9.3.2 above, is shown.

What distinguishes the contexts at the high end of Dimension 2—the ‘find out’ group—from the more general achievement meaning is the stronger tendency to be expressed in terms of ‘hearing’ (more than two thirds of the lexical markers in table 9.9). This separates ‘find out’ from other contexts within the achievement group—it is characterised by being brought about by something external rather than an internal process (cf. sections 9.3.3 and 9.5). We saw an example of this above in 61 and, as mentioned, the co-expression of knowledge and perception is discussed in more detail in chapter 10.

In general, it can be noted that there is considerable overlap in the type of markers associated with ‘find out’ and with achievement in general. This suggests that an important part of the semantics of ‘find out’ is that it’s an achievement, and what was discussed in section 9.3.2 applies also here.

An interesting marker found here which seems to represent a specialisation of the ‘find out’-meaning is the Iu Mien gorngv ‘say’, which combines
with *haiž* ‘hear’ to expressed something like ‘find out through hearsay’ (one of the two markers found under ‘say’ in table 9.2 is involved in a similar construction). In Iu Mien, the construction with *gorngv* ‘say’ contrasts with *hiuv duqv* ‘know GET’, which appears to have a more general know-as-achievement meaning. The involvement of a form meaning ‘say’ would seem to reflect the second-hand nature of the acquired knowledge here—we hear what someone says happens, not what actually happens. This connection is observed also in the English term *hearsay*.

(78)  Iu Mien (Hmong-Mien, Myanmar)


\[ \text{Mienh camv } \text{hiuv duqv, ziouc gan } \text{jienv} \]
person many know GET, then follow CONT

‘But when the crowds found out, they followed him’

b. Luke 7:37

\[ \text{Ninh } \text{haiž gorngv } \text{Yesu } \text{yiem wuov dauh Faalîši } \text{Mienh} \]
3 hear say PR.NAME be.at DEM CLF pharisee person
\[ \text{nyei biauv nyanc nnaangx} \]
GEN house eat rice

‘she learned [through hearsay] that [Jesus] was dining in the Pharisee’s house’

To summarise, then, to find out something is a type of knowing-as-achievement. To find out something is also like hearing something. As a special type of finding out, we have finding out by hearsay, which is like finding out about something by hearing what someone says about it; it explicitly results in second-hand knowledge.

9.3.5  Interim summary

The preceding sections have attempted to give a semantic characterisation of Dimension 1 and 2 of the CMDS analysis of the extended knowledge domain by considering the markers correlated with the dimensions.

I argue that two semantic parameters underlie Dimension 1. First the state–achievement distinctions which divides the dimension into know-as-state—a largely opaque concept which is mostly not like anything else, but is in

---

1Note that the very exhaustive grammar of Iu Mien brings up the *haiž gorngv* construction and makes no mention of it being a calque (Arisawa 2016).
some ways more like an object or property than an action—and know-as-achievement—which is both like knowing-as-state and like perceiving. The second parameter underlying Dimension 1 is the distinction of an ‘understand’ group. Consideration of forms associated with these contexts reveal that ‘understanding’ is quite complex. It is like knowing, but more, just as it is like hearing, but more. An important component of ‘understanding’ appears to be its objects—these are things typically associated with hearing, like languages, words and meanings. Furthermore, ‘understanding’ is often dynamic rather than static—an achievement rather than a state. This is what explains the intertwining of ‘understanding’ and achievement in Dimension 1. In Dimension 2, the ‘find out’-group of contexts is distinguished. These are achievements, but are distinguished within that group by being externally caused by hearing.

9.4 Recognising and knowing people

Before starting with the form-based analysis of meaning in Dimension 3, a general observation can be made, namely that one of the clearest differences in form between verses at the high end of the dimension and at the low end of the dimension is the form of the object—nominal in the former case and clausal in the latter. For methodological reasons, this difference was not included in the coding of verses (see section 3.6.2), but it might nevertheless be of interest to note this distinction in form—it raises the question of to what extent a relation is defined by its objects. Can the same kind of relation hold between a subject and objects which are of fundamentally different kinds? These questions fall outside the scope of this chapter.

With these preliminary remarks, we will move on to consider the semantics of Dimension 3, beginning with the group of contexts at the extreme high end of the dimension identified in the previous section as having the meaning ‘recognise’. For this purpose, the Point-Biserial correlation introduced above is not the ideal tool. With this measure, we consider whether there is a correlation between a binary variable (the presence or absence of a marker) and a continuous variable (the values of a dimension), essentially telling us whether a marker is more likely to be found on either end of the dimension. The measure is less helpful for inferring whether a marker is found only on the extreme end of a dimension or across a greater part of it. Given the preliminary analysis of Dimension 3 as a scale of meanings presented above, this information is necessary. Is a form associated only with the very high end of Dimension 3—i.e. only the ‘recognise’-meaning—or with the whole positive side of the dimension?

I will approach this in a very straightforward manner and in the investigation of forms for ‘recognise’ limit the markers considered to those where 1) the
correlation with Dimension 3 is positive and the p-values less than 0.01 and 2) the value on the dimension of the verse lowest on Dimension 3 which is coded with the marker is greater than that of the lowest verse characterised as ‘recognise’ (Acts 19:34—this is in fact a somewhat atypical ‘recognise’ context, but note that all the markers included which had this as the lowest-values verse also included several higher-valued, i.e. more typical, ‘recognise’ contexts). This gives 51 markers from 35 languages. They are summarised in tables 9.11 and 9.12.

The first thing to be noted is that two thirds of the markers are constructional (remember that this includes both morphological markers as well as markers which are constructionally bound but not necessarily morphologically so). This tells us that the meanings expressed in these verses are often understood in terms of other meanings. The right part of table 9.12 summarizes what these meanings are—most common are ‘know’, ‘know (person)’ and ‘see’.

It is clear from the relevant forms that one component of the meaning of ‘recognise’ is aspectual (“Aspect” in table 9.12). In this, there are clear similarities to what was discussed in section 9.3.2 above and, indeed, many forms considered there have among their uses ‘recognise’ (see e.g. 60). Another common situation is when a form which on its own means something like

---

**Table 9.11:** Summary of lexical markers positively correlated with Dimension 3 only occurring in ‘recognise’ contexts: number of markers (number of languages)

<table>
<thead>
<tr>
<th>Function</th>
<th>Combines with</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Aspect</strong></td>
<td>‘know’</td>
<td>11 (11)</td>
</tr>
<tr>
<td>‘again’</td>
<td>‘see’</td>
<td>10 (5)</td>
</tr>
<tr>
<td><strong>Motion</strong></td>
<td>‘know (person)’</td>
<td>9 (8)</td>
</tr>
<tr>
<td><strong>Intensifier</strong></td>
<td>‘recognise’</td>
<td>2 (2)</td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td>Other</td>
<td>3 (3)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>33 (25)</td>
</tr>
</tbody>
</table>

**Table 9.12:** Summary of constructional markers correlated with Dimension 3 only occurring in ‘recognise’ contexts: number of markers (number of languages)

<table>
<thead>
<tr>
<th>Connections</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(Opaque)</td>
<td>10 (10)</td>
</tr>
<tr>
<td>‘see’</td>
<td>4 (4)</td>
</tr>
<tr>
<td>‘identity’/‘similarity’</td>
<td>2 (2)</td>
</tr>
<tr>
<td>‘hear’</td>
<td>1 (1)</td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td>1 (1)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>18 (16)</td>
</tr>
</tbody>
</table>
’know person’ combines with some aspectual marker to express ‘recognise’. We saw an example of this above in [59] where the Hungarian ismer- ‘know (person)’ combines with coverbs to express ‘recognise’.

Also recurrent are forms meaning something like ‘again, repetition, return’. They are all subsumed under the label ‘again’ in table [9.12]. We have already seen an example of this in Swedish, where känna ‘know (person)’ combines with igen ‘again’ to give the meaning ‘recognise’—see example [9.16]

(79) Swedish (Indo-European>Germanic, Sweden)

a. John 7:29

\[
\text{jag känner honom därför att jag kommer från honom.}
\]

1SG know.PRS 3SG.M.O because COMP 1SG come.PRS from honom

3SG.M.OBJ

‘I know him because I am from him’

b. Mark 6:54

\[
\text{så snart de steg ur båt-en kände folk-et igen honom.}
\]

as soon 3PL step.PST out.of boat-DEF know.PST people-DEF igen 3SG.M.O

‘as soon as they stepped out of the boat, the people recognised him’

Other examples of repetition or return playing a role in the expression of ‘recognise’ are the iterative particle fa in Bauzi (Geelvink Bay, Indonesia) or the particle rapa ‘again’ in Carib (Cariban, Guyana). The English recognise, too, has an again-component in the prefix re-. Forms like this suggest that the event of recognition is a repetition of a previous event.

In addition to these more aspectual components, it is also clear from tables [9.11] and [9.12] that recognition is tightly connected with sight. An example of this is Doromu-Koki where the construction used for ‘recognise’ is literally ‘see fully do’:
Doromu-Koki (Manubaran, Papua New Guinea)

a. Mark 6:54

amiye evade vei mama re-yadi
person quickly see fully do-2/3PST

‘people recognised him immediately’

As another example, see the San Blas Kuna example above (62b). The connection to sight is not particularly surprising if we consider the prototypical frame of ‘recognition’ as including a person as object—this point is returned to below. Note also the parallel to the discussion of understanding and hearing above—just like understanding is like hearing but more, recognising is like seeing but more.

One final type of form offers clues about the semantic characterisation of ‘recognise’—those with a meaning of similarity or identity. We find this type of construction in two languages in the sample (‘identity’/‘similarity’ in table 9.11). There is Kannada where gurutu ‘identity’ is involved in a number of constructions meaning ‘recognise’ and Wubuy (Gunwinyguan, Australia) where nguynju ‘similar’ is used to form nguynju-ga ‘similar-factative’ lit. make similar ‘recognise, identify’. An example of one the Kannada forms is given in 81.

(81) Kannada (Dravidian, India)


āga aṣaru yēsūs-annu gurut-isi-daru
then 3PL PR.NAME-ACC identity-AUX-3PL.PST

‘and they recognised [Jesus]’

What these forms reveal is that recognition is linked to identification (note the full colexification in Wubuy). This is connected with the semantics of repetition—to identify someone requires sorting them into previously held categories, which will frequently have come about by a previous similar process (i.e. sight).

The semantics of ‘recognise’ can thus be characterised as being similar both to knowing and seeing. In several languages, it is viewed as a sub-type of ‘know person’ in general, often being specified for achievement meaning. There is also a component of return or repetition in several languages. Also of interest for the characterisation of meaning is the connection to similarity or identity. In short, to recognise is to by sight identify a person by connecting them to a previously held representation.
Moving on to the characterisation of the rest of the dimension, the remaining markers (not associated only with the very high end of the dimension, i.e. ‘recognise’) will be considered—they are summarised in tables 9.13 and 9.14. Additionally, as mentioned, many expressions for ‘recognise’ are based on more general forms for ‘know person’—this allows us to make inferences about ‘knowing person’ based on ‘recognise’.

Immediately noticeable is the high proportion of synchronically opaque lexical stems (“(Opaque)” in table 9.13), comparable to the high proportion of opaque general forms for ‘know’. This means that the approach of investigating forms to reveal patterns of conceptualisation cross-linguistically is of limited applicability. To know a person is largely not understood in terms of others contexts. The two exceptions here are of course the many languages where knowing a person shares expression with knowing a fact—clearly these two are very frequently conceptualised as being the same kind of relation. It is rare in the sample that one of the senses is clearly understood in terms of the other (similar to how understanding is frequently understood in terms of knowing). The one clear exception is Thai, where the form for ‘know that’ is riūu. This is part of the form for ‘know person’, in addition to câk. This câk is an obsolete word for ‘know’, now only used in this compound (Jenny 2019: 584).
(82) Thai (Tai-Kadai, Thailand)

a. Mark 12:14

\[ \text{raw ŭ́u wàa thàan pen khon sùutroŋ} \]
1PL know COMP 2HON be person upright

‘we know that you are truthful’

b. John 7:29

\[ \text{tèe raw ŭ́u-càk phraá??ŋ} \]
but 1PL know-know 2/3.GOD

‘[But] I (lit. we) know him’

There are in addition several examples where the forms within the domain share a lexical morpheme, but it is difficult to judge that either form motivates the other. An example is Nopala Chatino. The forms for ‘know that’ and ‘know person’ share the root -lyo-. For ‘know that’ the form is j-lyo- ti’—causative jy-, -lyo- ‘knowledge’ and ti’ ‘heart, essence of a person’ which is used in the expression of several cognitive and emotional concepts (Pride & Pride 2010: 19, McIntosh 2015: 219). For ‘know person’, the construction is tyu’wi ‘be/have’ + lyo(o) ‘knowledge’ + ‘in ‘indirect object’. Examples are given in (83).

(83) Nopala Chatino (Otomanguean>Eastern Otomanguean, Mexico)

a. Mark 12:14

\[ \text{huare’ jlyo=tì’ hua si’yana ‘un lca um sca nu} \]
1PL know 1PL that 2HON be 2HON INDEF REL
\[ \text{nchcui’ cha’ nu ñì} \]
speak.HAB word REL true

‘we know that you are truthful’

b. John 7:29

\[ \text{Una nan’ ndyu’hui lyoo-n ‘in ñì si’yana} \]
? 1SG have.HAB knowledge-1SG PREP 3HON because
\[ \text{mdo’o-n se’en nducua ñì} \]
come.from-1SG place ?of 3HON

‘I know him, because I am from him’
The other exception to the opacity of forms is the connection, found also with ‘recognise’, to sight (‘see’ in table 9.13). As an example of this, we can consider Hopi. In Hopi, five main forms carve up the extended knowledge domain: navota, navoti’yta, tuwi’yta, mămatsi and mātsi’yta. Note that there is also the form tuwa which does not appear in these verses—this is the form for ‘see’. It is clear that there is a relationship between navota and navoti’yta and tuwa and tuwi’yta, presumably a derivational one.

In figure 9.2, the distribution of these forms in the verse sample is shown. Note that the forms mămatsi and mātsi’yta are used both for ‘recognise’ verses and ‘understand’ verses. This pattern of co-expression is limited to Hopi in the language sample. What is of interest here, however, is the other three forms—navoti’yta is used for ‘know that’, navota for ‘find out’ and tuwi’yta is used for ‘know person’.

Hopi thus illustrates the connection both between knowing people and sight and knowing facts and hearing. As a clue to why knowing people is so closely connected with sight, we can consider Car Nicobarese, where the same verb root is used for both knowing facts and people as well as for understanding. The precise meaning is determined by what the root akaha ‘know’ combines with—lōn ‘mind, think’ for ‘know that’, kuö ‘face’ for ‘know person’ and elnang for ‘ear’. Examples are given in 84.
To know a person is thus connected with knowing a face. Given the historical connection between forms for ‘face’ and forms for ‘see’ or related meaning such as ‘appearance’, ‘image’ (Steinberg 2010: 280) it can be suggested that this is also the explanation behind the connection with sight. There is much evidence that the face holds a special place in how humans relate to each other—research in psychology suggests that face perception is special in several ways (e.g. Tsao & Livingstone 2008) and face is an important concept in many cultures (e.g. Strecker 1993). It thus seems likely that the connection between knowing people and sight is in part a reflection of the privileged status of the face in social relations.

To summarise, then, to know a person is similar to knowing a fact. To know a person is also similar to seeing a person—in particular, seeing their face—but involves something more.

This, it might seem, is not a particularly satisfactory characterisation of what it means to know a person. As we saw in section 2.2.2 some philosophers have suggested that there is a distinct kind of knowledge of people—interpersonal knowledge, which is the result of reciprocal causal contact. I argued in that section that Swedish käamma (unmodified by any particles) is limited to this sense of ‘know (person)’. It seems like this sense is poorly captured by the likeness to seeing—knowing a person in this interpersonal sense does not follow from seeing them. A possible explanation for this is that the extension to personal knowledge proper is secondary, and that forms connected with seeing enter the domain via recognition—where, as we saw, sight is clearly crucial (cf. section 10.4.2).

Note that apart from the forms focused on sight, there are few transparent forms found correlating with Dimension 3. This might be taken to suggest that other than its similarity to knowing that—which is evident from the high degree of co-expression cross-linguistically—to know a person is a largely
opaque concept—similarly to knowing in general it is not understood in terms of other concepts, but is treated as a unique whole.

9.5 Realising

The forms correlated with Dimension 4 reveal that the semantics of it are closely tied to the internal—i.e. the mind and thought. There are 61 markers which are positively correlated with Dimension 4 with a p-value of less than 0.01, from 36 languages. In tables 9.15 and 9.16 the categories into which these constructions fall are shown.

The majority of the forms found are synchronically opaque, and thus cannot be used to explore the semantics of the dimension. Among the constructions which are analysable, a great degree of similarity with the forms for ‘understand’ can be noted—markers combining both with ‘hear’ ‘know’ are found (table 9.16). Given the analysis of this dimension as making a further distinction among the ‘understand’ and ‘find out’ contexts, this is not surprising.

What distinguishes the forms correlated with this dimension is, as mentioned, a connection to the internal and to thinking. An illustrative example is the Mapudungun ngünduam- ‘realise’. The form ngünduam consists of con-
sists of two parts—ngüne- ‘govern, direct’ and duam ‘attention’, i.e. literally roughly ‘to govern one’s attention (toward)’. Example[54] is repeated below.

(85) Mapudungun (Araucanian, Chile)

a. Acts 19:34

\[\text{Welu } \text{ngüne} \d\text{duam-nge-lu } \text{kisu } \text{ta-ñi } \text{judiu-nge-n}\]

‘But realise-PASS-TEMP self DEF-3POSS Jew-VRBZ-NMLZ

‘But when they recognised that he was a Jew’

Another example can be found in Sanumá (Yanomamic, Brazil) where the construction pi ku-, lit. ‘inside say’, is used for some ‘realise’ contexts as well as ‘think’ and ‘remember’. As another example of this pattern of colexification, consider Doromu-Koki, where moke- appears to be used for both ‘think’ and ‘realise’, see example[86]

(86) Doromu-Koki (Manubaran, Papua New Guinea)

a. Luke 1:22

\[\text{dada } \text{moke-yadi } \text{ina } \text{bi } \text{Sei } \text{ni } \text{vake-na } \text{yava } \text{ide-ri}\]
\[\text{so think-2/3.PST 3 TOP god AUX } \text{?-NMLZ house inside-at}\]
\[\text{rebe } \text{yokoi ve-yo}\]
\[\text{vision/ghost one see-2-3.PST}\]

‘and they realised that he had seen a vision in the temple’

We can thus summarise the semantics of this dimension as being similar to ‘find out’ and ‘understand’ in being achievements, but separate from ‘find out’ and similar to ‘understand’ in being mainly internal rather than external. The focus on this internal aspect, however, is stronger than with ‘understand’, which can be seen both in the origins of forms—which in several cases contain forms for the mind or the inside—and in colexifications—where these forms are in several cases also used for ‘think’ and ‘remember’.

To realise something, then, is a kind of know-as-achievement, sharing many characteristics with understanding. It is similar to thinking about something or remembering something and it involves the parts of you responsible for intellect.

9.6 Not knowing

Although it is not included in the coding schema of this investigation, the frequent use of standardly negated positive ‘know’-words give clear evidence that
'not knowing' is frequently conceptualised in terms of 'knowing'. This is not very surprising. The semantics of negation is a heavily researched area which will not be dwelt on further here. Instead, the focus will be on what the forms representing a conceptualisation not in terms of positive knowing can reveal about these contexts.

As was seen in section 7.7 one of the striking properties of Dimension 5 is the clear separation of one verse—Luke 12:56(b)—from the others. It was noted that this is due to forms with a basic meaning ‘not be able’ extending into the knowledge domain. The placement of this verse at the very low end of the dimension gives rise to some interactions in which also some contexts which have a meaning of ability rather than negation are placed relatively low on the dimension (since they too are similar to Luke 12:56(b)). This results in some forms which are clearly not negative in meaning correlating with low values in Dimension 5.

The forms presented in table 9.17 are therefore manually compiled, with forms judged not representing lexical negation excluded and a number of forms added. In table 9.17 forms which correlated with Dimension 5 with a p-value <0.01 are marked with two asterisks and those correlating with it but with a p-value >0.01 are marked with one asterisk.

As has already been touched upon, there is a considerable overlap between forms for ‘not knowing’ and ‘not being able’—this seems to reflect a general overlap between modality and ‘know how’, see section 9.7.

There is also an interesting connection where several of the forms are more general than just negation of knowledge, but also of memory and in some cases even consciousness entirely. How are these colexifications best understood? Recall from chapter 2 that the traditional analysis of knowledge posits three criteria for knowing something, formulated by Gettier (1963) 121 as:

S knows that P iff

(i) P is true,

(ii) S believes that P, and

(iii) S is justified in believing

If any of these criteria is not fulfilled, it is not true that S knows P, i.e. S does not know P. In other words, whereas ‘know’ includes all three criteria, ‘not know’ need only assert that one of the criteria is false. If this is the second criterion—for S to believe that B—it is easy to see that such a form’s meaning will also extend to ‘not remembering’ (as in Moro and Doromu Koki) and, in

---

1For unknown reasons, the well-known Korean moreuda ‘not know’ does not occur in the verse sample.
<table>
<thead>
<tr>
<th>Language</th>
<th>Form</th>
<th>Other meanings</th>
<th>Count (/15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. Alaskan Yupik</td>
<td>nallu-</td>
<td>used negated for ‘know’</td>
<td>34</td>
</tr>
<tr>
<td>Moro</td>
<td>-i生命力</td>
<td>a derivation is used for ‘forget’</td>
<td>14</td>
</tr>
<tr>
<td>Bauzi</td>
<td>bi-, bie-</td>
<td>unclear what the relationship between the forms is; bi and vabi are used for ability, vabi also for ‘forget’</td>
<td>12</td>
</tr>
<tr>
<td>Moro</td>
<td>-i生命力</td>
<td>a derivation is used for ‘forget’</td>
<td>14</td>
</tr>
<tr>
<td>Bauzi</td>
<td>bi-, bie-</td>
<td>unclear what the relationship between the forms is; bi and vabi are used for ability, vabi also for ‘forget’</td>
<td>12</td>
</tr>
<tr>
<td>Olo</td>
<td>ironom**</td>
<td>‘be unconscious’</td>
<td>8</td>
</tr>
<tr>
<td>Olo</td>
<td>wouye</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Nahro</td>
<td>c’úù**</td>
<td>‘not be able’</td>
<td>10</td>
</tr>
<tr>
<td>Culina</td>
<td>ssamo-**</td>
<td>‘not be able’</td>
<td>9</td>
</tr>
<tr>
<td>Toro So Dogon</td>
<td>ine**</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>Wubuy</td>
<td>malaladi**</td>
<td>also ‘not be able’</td>
<td>9</td>
</tr>
<tr>
<td>Daga</td>
<td>oare**</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>North Tanna</td>
<td>- órguru**</td>
<td>also ‘not be able’</td>
<td>8</td>
</tr>
<tr>
<td>San Blas Kuna</td>
<td>aku**</td>
<td>‘not be able’ re perception; extended to achievement cognition uses</td>
<td>7</td>
</tr>
<tr>
<td>Dogrib</td>
<td>ha dìì</td>
<td>‘be too much (for someone)’, ‘not be able’</td>
<td>2</td>
</tr>
<tr>
<td>Kriol</td>
<td>kaan**</td>
<td>‘not be able’</td>
<td>2</td>
</tr>
<tr>
<td>Kire</td>
<td>kakagi**</td>
<td>‘mistake, err’</td>
<td>1</td>
</tr>
<tr>
<td>Ma’di</td>
<td>kpe**</td>
<td>‘not be able’</td>
<td>1</td>
</tr>
<tr>
<td>Tamasheq</td>
<td>indär-**</td>
<td>‘not be able’</td>
<td>1</td>
</tr>
<tr>
<td>Uduk</td>
<td>mol-</td>
<td>‘not be able’</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table 9.17:** Lexically negated forms with correlation to Dimension 5 marked
some cases, to not believing anything at all, that is, being unconscious, as in Olo-ironom. If, on the other hand, we consider a form as asserting the negation of the first criterion, we see how the colexification with ‘mistake’ might occur (as in Kire). Note that such false belief expressions are not entirely uncommon cross-linguistically (Spronck & Vuillermet 2019, McGregor 2023).

An interesting question for further research is if we ever find lexicalised expressions for negating the third criterion—i.e. a form which expresses that S believes P, that P is true but that S has not come to this belief in the right way.

Another interesting issue concerning the lexical negation of knowledge which deserves mention is the existence of particle or interjection-like forms meaning something like ‘I don’t know’. We find such forms in several European languages, which are usually eroded forms of the standard negation, e.g. English dunno (<(I) don’t know), Swedish venne (<(jag) vet inte) or French chepa (<je ne sais pas). These kinds of forms do not generally appear in the NT, probably due to register.

Clearly, there are many issues left to explore regarding how not knowing is talked about in the languages of the world, but the following preliminary semantic characterisation can be formulated. To not know something is related to knowing something, but it’s the opposite. To not know something may be like being unaware of something or even unconscious or it may be like being mistaken about something. Furthermore it would seem that not knowing, though this is not really reflected in the NT, is a conversational and pragmatic thing, associated with the first person.

### 9.7 Knowing how

As we saw in the previous section, there is a strong connection between ‘know how’ and ability in the negative. The same true in the positive, as is clear from the markers correlated with Dimension 6, summarised in tables 9.18 and 9.19. A large portion of the markers are forms which mean something like ‘can, be able’. Given the limited number of relevant contexts, it is difficult to say anything about how far this co-expression of meanings extends in the languages. In section 2.2.1 I suggested that a continuum from pure physical ability to factual knowledge via ‘know how’ can be posited. It is clear from the data that forms associated with the ability-end of the continuum sometimes extends to contexts more like ‘know how’, but whether this represents complete co-expression of ability and ‘know how’ is impossible to tell based on the current data. It does, however, tell us that ability and ‘know how’ are closely connected conceptually. Note that also the verbal morphology in table 9.19 falls under this category—morphologically marked modality which is extended into the knowledge domain.
Connections

<table>
<thead>
<tr>
<th>Function/meaning</th>
<th>Combines with</th>
<th>Number of Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘be able’</td>
<td>‘know’</td>
<td>13 (13)</td>
</tr>
<tr>
<td>(Opaque)</td>
<td>‘be able’</td>
<td>4 (4)</td>
</tr>
<tr>
<td>‘see’</td>
<td>‘see’</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>18 (17)</td>
</tr>
</tbody>
</table>

Table 9.18: Summary of lexical markers positively correlated with Dimension 6: number of markers (number of languages)

Table 9.19: Summary of constructional markers positively correlated with Dimension 6: number of markers (number of languages)

As an example of a language where an ability form is found among the correlated markers and where, I would argue, this in fact represents a colexification of ability and ‘know how’, consider Swedish. In Swedish, the modal auxiliary verb *kunna* means ‘can’ (110a) but also ‘know how’, both in constructions with a verbal complement, where it is arguably an auxiliary verb, (110b) and as the main verb with a nominal object (110c).

(87) Swedish (Indo-European>Germanic, Sweden)

a. Matt 26:61

\[
jag \text{ kan } \text{ riva ner \textit{Gud-s tempel och bygga upp det det}}
\]
\[
\text{igen inom tre dag-ar}
\]

‘I am able to destroy the temple of God and rebuild it within three days’
b. Luke 12:56(a)

_Hycklare, jord-en-s och himl-en-s utseende_

hypocrite.PL earth-DEF-GEN and sky-DEF-GEN appearance

_kan ni tyda [—]

can.PRS 2PL interpret

‘Hypocrites, you know how to evaluate the appearance of the earth and sky’

c. Acts 21:37

_Jaså, du kan grekiska_

really 2SG can.PRS Greek

‘Oh, so you know Greek?’ (AS)

We find also a number of forms apparently specific to the ‘know how’ meaning. An example of this was seen in section 7.8 Car Nicobarese _huve ülke_ ‘know how’ (cf. example 109, p. 279). A dictionary of Car Nicobarese gives the meaning as ‘to be right, to understand, to succeed, to do rightly’ (Whitehead 1925: 91). Another form, _arandu_ in Kaiwá (Tupian, Paraguay) also has meaning connected to wisdom, according to a dictionary (Loman de Barros 2014: 148).

In addition to expression with a specific form or coexpression with ability, we find several languages where the general form for ‘know’ is used in these contexts, in several cases in a particular construction—this is reflected by the presence of _known_ markers in table 9.19. As an example of this kind of construction, consider Basque. Complements of _jakin_ ‘know’ are usually finite with a complementiser suffixed to the finite verb (88). In another construction, however, _jakin_ combines with a nominalized form of a verb, expressing ‘know how’.

(88) Basque (Isolate, Spain)

a. Artiagoitia et al. (2003: 646)

_Banekien_ Mikel berandu etorriko zena.

BA.know.1SG>3SG.PST Mikel late arrive.FUT AUX.COMP

‘knew that Mikel would arrive late’
These Known markers of table 9.19 are thus de-finitizing morphology of various kinds. There are, I believe, two ways to understand the connection of these ‘know’ + less-finite complement constructions with ‘know how’ meanings. It may be that they are formed by analogy with constructions with modal forms. Alternatively, it could be considered a case of iconicity in syntax, where the higher syntactic integration of a verb + non-finite verb combination reflects the semantic integration of the two in a ‘know how’ situation (cf. e.g. Givón 2001b: 40–58). (Arguably, also the first explanation ultimately comes down to the second, if the syntactic form of modal sentences is explained by the principles of event integration).

The correlated forms also reveal a connection between knowing how and habituality (“Habitual” in table 9.19), though it seems that this connection is only via ability. As an example, consider the Kuot example in 89.

(89) Kuot (Isolate, Papua New Guinea)

a. Luke 12:56(a)

\[\begin{align*}
\text{Mi-me maset a-gima kaguma onim na kimanam} \\
\text{2PL.S-HAB well 3SG.M.O-see sign origin PREP earth} \\
\text{ga panbinim} \\
\text{and sky}
\end{align*}\]

‘You know how to evaluate the appearance of the earth and the sky’

Lit.: ‘you see well the signs of the earth and sky’ (AS)

Knowing how to interpret—which is the meaning in the source text—is here expressed as (or replaced by) by habitually seeing well. The expression of knowing in terms of perception and something more has been discussed several times above, and in addition we have here a habitual marker. The sample seems to contain a number of languages where ability is connected to habituality. This is understandable, given that ability to do X follows both causally and by implication from habitually doing X.

In summary, then, to know how to do something is similar both to knowing that something is the case and being able to do something. It has to do with cognitive success (being wise, succeeding, doing rightly etc.) and relates, via ability, to habituality—it is like (or the result of?) doing something often.
9.8 Summary

In this chapter, the semantics underlying the dimensions identified in chapter 7 has been investigated in more detail, with a method based on the view that linguistic forms reflect speakers’ conceptualisations.

To begin with, the basic constructional divide in the domain is between know-as-state and know-as-achievement. To know something as a state is largely opaque—it is not expressed in terms of other concepts. The presence of constructions based on nominal forms, does, however, reveal that knowing deviates from prototypical actions in being more like a thing or perhaps a property. To know something as an achievement is much like knowing something like a state and is often expressed by means of an alternative aspectual construal of the same basic event. However, know-as-achievement is also understood as involving motion and as being like perceiving.

The basic lexical divide singles out the ‘understand’ meaning. Understanding is complex—it is like knowing, but more, just as it is like hearing, but more. Understanding has to do with things like languages, words, meaning and truth—but it also has to do with other people. As evidenced by their sharing Dimension 1, understanding also has an dynamic component—it is often more like an achievement than a state. Interestingly, the characterisation of ‘understand’ given here based on cross-linguistic data appears to largely agree with those based on particular languages discussed in section 2.2.5.

‘Understand’ contexts, then, are one part of the larger know-as-achievement group. Another group are the ‘find out’ contexts. These are often conceptualised as the achievement counterpart of stative know. What distinguishes them from the other types of know-as-achievement is that they are also conceptualised as being the same as hearing—to hear something is to find something out. In this, finding out represents externally caused knowledge. Additionally, forms expressing a meaning roughly like finding out through hearsay are found in some languages. These forms explicitly marks the knowledge produced as second-hand—coming from what someone says about something, not from direct access.

Recognition is, formally, like the achievement counterpart of knowing a person. It is like seeing someone, but involves an element of connecting the sight to previous experience. To know a person is similar to knowing a fact but it too, like recognition, is similar to seeing a person—in particular, seeing their face. But just like understanding is like hearing but more, knowing a person is like seeing, but more. The relatively many opaque forms, however, suggests that to know a person is something conceptualised only in terms of itself.

To realise something is similar both to finding out and understanding in being achievements. What distinguishes realisation is the stronger connection
with the internal as well as with ‘thinking’ and ‘remembering’. In contrast to finding out, realisation is internally caused know-as-achievement. Understanding, with its connections both to hearing and internal processing, seems to be somewhere in between. Thus, to realise something is similar both to knowing-as-achievement and to thinking or remembering, and it involves the part of you responsible for intellect.

To know how to do something is similar both to knowing that something is the case and being able to do something. It is related to doing something habitually (though this connection, it seems, is through ability) and it is connected to cognitive success (wisdom, doing things right etc.) more broadly.

It is clear from what has been presented in this and previous chapters that there are recurrent connections between knowledge and other concepts. These connections are explored more closely in the next chapter. In the following chapter (11) the distribution of the forms inventoried here are discussed from an onomasiological point of view.
10. Knowledge and other concepts

10.1 Introduction

The connections between knowledge and the other concepts included in the verse sample were noted already in chapter 5. In the previous chapters (8 and 9) it was seen that these connections are not constant across the domain; different parts differ in their likelihood to be externally motivated and by which concepts. In this chapter, the linguistic connections between knowledge and the other concepts investigated will be further considered, thereby addressing research question B3.

B3. How do the forms for [...] meanings [of the knowledge domain] relate to forms for domain-external meanings?

The chapter is structured as follows. First, a brief review of previous research on the topic is given (section 10.2). This is followed by a presentation of the connections found in the data of this investigation (section 10.3), which is then given an explanation in terms of metonymical shift within frames (10.4). The methodology of the chapter is summarised in table 10.1.

<table>
<thead>
<tr>
<th>Research Question</th>
<th>B3. How do the forms for [...] meanings [of the knowledge domain] relate to forms for domain-external meanings?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operationalisation</td>
<td>How similar are the different contexts of the knowledge domain to the non-knowledge identified in 5?</td>
</tr>
<tr>
<td>Method(s)</td>
<td>Clusters, lexically weighted similarity measure</td>
</tr>
<tr>
<td>Language sample</td>
<td>Core</td>
</tr>
<tr>
<td>Verse sample</td>
<td>Full</td>
</tr>
</tbody>
</table>

Table 10.1: Methodological summary of chapter 10
10.2 Previous literature on the inter-domain connections of ‘know’

In the literature, two etymologies for ‘know’ forms dominate the discussion—from forms for ‘see’ or forms for ‘hear’. The debate has, in part, concerned what might determine which etymology is found in a language. In this section, I give a brief summary of some of the arguments put forth and propose an alternative approach to the question, based on the view of knowledge as a domain with many parts.

One of the two main sources for ‘know’ forms in Indo-European is ‘see’. (the other being a root with no earlier reconstructable meaning than ‘know’). The source is reconstructed to be the perfect *uōid-luid- of *uēid- ‘see’ (Rix & Kümmel 1998: 666, Pokorny 1959: 1125–7). Sweetser (1990: 45) suggests that the connection between seeing and knowing, both diachronic and synchronic, might be universal. This is argued for not primarily through the etymologies of ‘know’ forms cross-linguistically, but through the pervasiveness of seeing-metaphors in speaking about knowing in English—such as the ubiquitous I see. On Sweetser’s view, the connection is due to a metaphorical mapping between the two domains. Since knowing and seeing are similar in for example allowing for distant objects and allowing for focusing of attention, a metaphorical mapping between them is invited.

This claim to universality has been challenged, for instance by Fortescue (2001) who in his sample of 73 languages (biased toward Indo-European as well as large and written languages) found that vision as a source domain was found only in Indo-European languages and by Evans & Wilkins (2000) who argue that Australian languages more commonly get their know-words from hearing. Most commonly, they argue, this change in meaning takes the path through ‘understand’ and a meaning shift to ‘know’ more often involves overt derivational morphology than a shift to ‘understand’.

The large-scale study in Georgakopoulos et al. (2022) finds links from both sight and hearing to knowledge. Interestingly, in their data, the connection between sight and hearing is mediated by cognition. They also note a stronger connection of sight to knowledge and hearing to what they call ‘mental manipulation’ (e.g. ‘understand’)—though it should be noted that their data set is biased toward Eurasia.

The link between hearing and understanding is recognised by Sweetser, who contrasts it with knowing (connected with seeing); “[b]ut hearing is connected with the specifically communicative aspects of understanding, rather than with intellection at large” (1990: 43). In Sweetser’s view, the ‘understand’-‘hear’ connection is tightly connected with the also often observed link between hearing and obeying: “hearing is also considered to represent the kind
of internal receptiveness to the speaker’s intentions which might subsequently lead to compliance with the speaker’s requests - i.e., to heedfulness and obedience” (1990: 43). Fortescue (2001: 21) notes the existence of the ‘hear’ > ‘understand’ shift in several families beyond Indo-European, e.g. in Siberian Yupik Eskimo and Afro-Asiatic.

In short, there is evidence that ‘know’ may be connected with both ‘see’ and ‘hear’, depending on language. As for the question of why this should be so, Evans & Wilkins (2000) raise the suggestion that words translated as ‘know’ in different languages actually refer to different concepts: in an Australian language where ‘know’ derives from ‘hear’, the form’s meaning might be paraphrased along the lines of “because of what I have heard, I say: X; because I heard it from the right people, I can say: X is true” (Evans & Wilkins 2000: 581). In a language where ‘know’ comes from ‘see’, the paraphrase would be appropriately changed (cf. the work of the anthropology of the sense, e.g. Howes 1991). As Evans and Wilkins point out, forms for ‘know’ having different senses depending on source is an untested hypothesis, and they also point to other factors, for instance the role of cultural scripts that put ‘knowing the way’ and ‘knowing the country’ in terms of learning stories, songs or lore. These scripts give rise to frequent bridging contexts of the kind ‘hearing (the names of) the way’ > ‘knowing the way’.

Evans & Wilkins (2000) also suggest that a factor in whether ‘see’ or ‘hear’ is the source for ‘know’ is a difference in the conceptualisation of the domains involved. As mentioned above, Sweetser (1990: 33) argues that one of the main motivations for the vision > cognition metaphorical mapping is the attention directing function of sight, which, in a Western view of the sense modalities, is not shared by hearing. In an Australian context, however, Evans and Wilkins argue, hearing is conceptualised as at least as sensitive to attention-directing as sight. This, then, would invite the hearing > cognition mapping.

Another factor mentioned by Evans and Wilkins is literacy. They suggest that in literate cultures, the connection between knowledge and sight is strengthened, whereas in cultures more focused on oracy, the role of spoken transmission of knowledge is greater, strengthening the connection between knowledge and hearing. Vanhove (2008) however, argues that this is not the case, as a number of pre-literate cultures have the vision-knowledge connection in her sample.

A factor which has not, however, to my knowledge, been considered in the literature is whether ‘hear’ and ‘see’ are connected to different senses of ‘know’—that is, to different parts of the knowledge domain. As mentioned, the connection between ‘understand’ and ‘hear’ has been pointed out in the literature, but as I have argued, also the narrow knowledge domain has internal divisions. In the following I will I will argue that seeing and hearing are
consistently connected with different parts of the knowledge domain, offering a piece in the puzzle of why both etymologies are found. In addition, also the other concepts considered in this investigation show connections with particular parts of the knowledge domain.

10.3 Connections to the extended knowledge domain

In chapter 5, I described how the similarity measure applied in this thesis can be averaged by cluster for a measure of how similarly two clusters are expressed. The same results is displayed in figure 10.1 with a focus on connections between the extended knowledge domain and the other clusters. Values range from 1 (identical expression of all verses in the clusters in all languages) to 0 (no shared expression in any language in any verse-pair)—this is represented in figure 10.1 and subsequent figures in this chapter by the width of the connecting lines. As discussed in chapter 5, intermediate values can come about in different ways. It may be that some contexts in some languages are similar, while others are dissimilar or it may the that all verses are partially similar, i.e. sharing some markers. For example, a value of 0.1 could indicate either that one tenth of the verses are identical or that all verses share one tenth of markers. In practice both things tend to factor in. The size of the nodes indicates the number of verses. As in chapter 5, the lexically weighted similarity measure is used on the core language sample.

In line with results of previous studies—and as will be explored in the following sections—the connections between perception and knowledge are the strongest, the connection with hearing (section 10.3.1) slightly stronger than with sight (section 10.3.2) (in line with the results of Georgakopoulos et al. 2022). There are also relatively strong connections with memory and ability (sections 10.3.3). The connection with thinking is weaker, and the connections with obedience and belief are entirely mediated by hearing and thinking respectively.

10.3.1 Connections with HEAR

Beginning with the strongest connection in the data—the one to ‘hear’—in figure 10.2, the average cross-language similarity of the individual verses in

Two verses are omitted as outliers, John 4:19 ‘The woman said to him, “Sir, I see that you are a prophet” in the SEE cluster and 1 Cor 1:16 (“Now I also baptized the household of Stephanus. Beyond that I do not know if I baptized anyone else” in the REMEMBER cluster. The reason is that their high values come from ‘know’ forms being used for these particular verses, which is interesting but does not tell us anything about how knowing is expressed in terms of other concepts.
Figure 10.1: Average similarity between KNOW and other clusters across the language sample

<table>
<thead>
<tr>
<th>Verse (English—Lexham)</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mark 15:45 And when he learned of it from the centurion [—]</td>
<td>0.51</td>
</tr>
<tr>
<td>Luke 7:37 [—] when she learned that he was dining in the Pharisee’s house [–]</td>
<td>0.48</td>
</tr>
<tr>
<td>John 12:9 Now the large crowd of Jews found out that he was there [—]</td>
<td>0.48</td>
</tr>
</tbody>
</table>

Table 10.2: The three verses most similar to HEAR

the extended knowledge domain to the HEAR-cluster are shown. The verses are represented by their expression in Swedish (chosen as the majority of the relevant distinctions can be seen in the Swedish forms), and rough labels in English have also been added.

Three groups of connections can be distinguished: to the ‘understand’ contexts (top), to the ‘find out’ contexts (top right) and to a single ‘recognise’ context (bottom left). By far the greatest is the similarity to ‘find out’ contexts, a connection discussed already in sections 7.4 and 9.3. In the contexts with the highest similarity, it is around 0.5, interpretable as these verses being equally likely to be expressed with a ‘hear’ as with a ‘know’ form. The verses are given in table 10.2.

A single verse expressed in Swedish with kännä igen ‘recognise’ also shows a strong connection to HEAR, namely Acts 12:14 with a value of 0.27. The explanation for this is that the verse in question concerns recognising a voice (“And recognizing Peter’s voice”)—it seems that some translations focus on the modality in rendering this verse. An interesting example is from
Naro, where, as discussed above (page 190) q’âa- ‘know’ combines with both bóò- ‘see’ and kóm- ‘hear’. The former gives a general know-as-achievement meaning, i.e. ‘find out’, ‘recognise’—this includes contexts of auditory modality (repeated as 90a)—while the latter gives the meaning ‘understand’ (example 63b). The recognition of a voice in Acts 12:14, however, uses kóm- ‘hear’ rather than bóò- ‘see’—apparently an orientation toward the nature of the object.

(90) Naro (Khoe-Kwadi, Botswana)


\[
A \ ba \ a \ ŋem \ ko \ bóò-a \ q’âa
\]

and 3SG.M DEM when PTCP see-CONN know

‘And when he found out’

¹Note that the context clearly determines that this is a case of recognising as identifying—the verses continues with the woman who hears/recognises the voice “announc[ing] that Peter was standing at the gate”.

Figure 10.2: Average similarity of coding of verses in the extended knowledge domain to HEAR
b. Acts 12:14

\[ E\text{-}e\text{-}s \quad k\text{o} \quad P\text{etere}-m \quad d\text{i}-m \quad d\text{öm} \quad b\text{a} \]

\[ \text{when-3SG.F} \quad \text{PTCP} \quad \text{PR.NAME-M} \quad \text{POSS-M} \quad \text{voice} \quad 3\text{SG.M} \]

\[ k\text{óm}-a \quad q\text{'\text{á}a} \]

\[ \text{hear-CONN} \quad \text{know} \]

‘And recognizing Peter’s voice’

The last group distinguishable in figure 10.2 is a connection to the ‘understand’-verses. These connections become clearer if we disregard the rather extreme connections between HEAR and ‘find out’, as shown in figure 10.3. Several examples of this kind of connection were seen in the previous chapter (9.3.3). In addition to the strong connection to ‘understand’, we see in figure 10.3 also that there is an asymmetry in the connection to hearing between ‘know that’ contexts and ‘know person’ and ‘recognise’ contexts where the former are more strongly connected than the latter (excepting the above mentioned Acts 12:14 and Luke 10:4 where the know object is also a voice). In the next section, it will be seen that this asymmetry is paralleled, though reversed, for seeing.

In summary then, HEAR is mainly connected to two groups of contexts in the knowledge domain: ‘find out’ and ‘understand’. There are also connections to individual verses, in particular Acts 12:14, but to some extent also Luke 10:4, which appears to reflect an orientation toward the object of
knowledge being a voice. Finally, a certain asymmetry in the strength of the connections between ‘know that’ contexts and ‘know person’ and ‘recognise’ contexts can be observed.

10.3.2 Connections with SEE

The similarity of the verses in the extended knowledge domain to the SEE-cluster is shown in figure 10.4. The strongest connection is to a verse expressed in Swedish with *upptäcka* ‘discover’—Acts 19:34, “But when they recognized that he was a Jew [—]”. An example of a translation where this context is expressed with ‘see’ is Zarma, as shown in example 91. This is the only context in the knowledge domain expressed by a perception form in Zarma.

(91) Zarma (Songhay, Niger)

a. Acts 19:34

\[\text{Amma waato=kaŋ i \ di kaŋ Yahudance no}\]
\[\text{but \ REL.TEMP 3PL see REL Jew FOC}\]

‘But when they [saw] that he was a Jew’

It is not clear why this verse should so often be expressed by ‘see’ forms. The original Greek does not contain one, nor do the major translations in English, Spanish, French or Portuguese. The context is that Alexander, who is
a Jew, is attempting to address the assembly in Ephesus during a riot but is interrupted by their shouts. Note that the realisation that the object of the complement clause is Jewish is clearly obtained via sight since Alexander has not yet spoken, but this does not explain why this verse is so much more likely to be expressed with ‘see’ than other recognise-verses where the sense is also clearly sight. A potential explanation is that forms for ‘recognise’ are syntactically limited to taking nominals as objects and are therefore not available while keeping to the original form. This is the case in Swedish, where känna igen ‘recognise’ cannot take a clausal object.

SEE is also strongly connected to in particular ‘recognise’ verses and also other ‘know person’ verses. The two strongest connection on the right side of the figure are in fact also examples of this, though the Swedish expression does not reflect this. The verses in question are Acts 7:18 (“who did not know Joseph”, Swedish veta om) and below that Mark 1:24 (“I know who you are”, Swedish veta vem).

As was discussed in chapter 9, many forms for ‘know person’ and ‘recognise’ are connected to sight—I suggested that this might be explained by the centrality of the face, which is highly visual, to inter-personal knowledge. In the case of ‘recognise’ there is also the fact that many of the ‘recognise’-verses in the sample can only be interpreted as recognition through sight, for example Mark 6:54 (see p. 210 for an example of expression with ‘see’) where the recognition is from “the people” “immediately” as the recognised (Jesus) gets out of the boat.

As an example of the extension more divorced from physical sight, consider Culina below (92) where the form for ‘see’ qqui- is also used for knowing people—note that the object of knowledge here is God, which rules out a purely physical interpretation. The speaker is Jesus, who is underlining his special relationship with God as he “[is] from him”.

(92) Culina (Arawan, Brazil)

a. John 7:29

\begin{verbatim}
Narraha ohua-pa qqui o-na-ja najari
\end{verbatim}

but 1SG-FOC see 1-AUX-? DEM

‘But I know him’

In summary, the connections between knowing and seeing appear to be concentrated to the contexts of knowing people, although also many other contexts show a low degree of similarity.

---

1According to one Bible scholar, “The […] details of the story are rather perplexing” (Brinks 2009: 788)
10.3.3 Other connections

There are also connections in the data between the non-perception clusters and the knowledge domain. The connections to THINK and REMEMBER are shown in figure 10.5. The similarity with REMEMBER is higher, but the contexts in which the similarity is found are to a large part shared—to be expected given the close connection between THINK and REMEMBER observed in chapter 5. Two sets of connections can be distinguished: in the upper part of the figure with ‘understand’ and in the lower with ‘know (that)’. The former concerns both THINK and REMEMBER, while the latter is, with some exceptions, more limited to REMEMBER.

Starting with the connection to ‘understand’ (top), this was discussed in the previous chapter, where I suggested that the shared focus on internal processing between understanding and thinking explains this (section 9.3.3). Aymara was discussed as an example (see page 201) where the verb amuya- is used both for ‘think’ and ‘understand’ (though with different complementation strategies).

It might seem puzzling that the similarity to REMEMBER is somewhat higher than to THINK in this group of verses. The reason appears to be that both ‘to remember’ and ‘to understand’ are quite frequently expressed with complex forms including a body-part term (or some general term for the intellect) whereas ‘think’ is expressed by a separate, simple form. As an example, consider Meyah, where a large range of expressions for cognition and emotion include the form odou ‘liver’ (Gravelle 2011: 247), including ‘understand’
(93a) and ‘remember’ (93b). ‘Know’ and ‘think’, on the other hand, do not share this component, but are expressed with the verbs -ejginaga and -osujohu respectively.

(93) Meyah (East Bird’s Head, Indonesia)

a. Mark 9:32

Tina Yesus efen ruforoker rudou en-ebriyi
but PR.NAME 3SG.POSS disciples 3PL.POSS.liver DUR-split
gij mar ongga Ofa anggot insa koma efen
in thing REL 3SG speak.PFV ANAPH DEM 3SG.POSS
ofou guru
meaning NEG

‘But [Jesus’ disciples] did not understand [the meaning of] the statement’

b. Matt 26:75

Petrus odou ah keingga rot mona ongga
PR.NAME liver lie.down ADV PREP time REL

‘And Peter remembered [the time when]’

This observation may be related to the suggestion of Natural Semantic Meta-language that not only KNOW, as discussed, but also THINK is a prime (Wierzbicka 2018b)—this would seem to fit with a tendency for forms for these to be non-compositional while other similar meanings follow a compositional pattern.

One verse in the data is a clear exception to this tendency for a stronger connection to REMEMBER than THINK, namely Luke 1:22(b) “and they realized that he had seen a vision in the temple”. An example of the use of ‘think’ in this verse was given from Doromu-Koki (p. 216). It is not clear why this verse in particular should show such a high degree of ‘think’ forms.

In the lower part of figure 10.5 the similarity to REMEMBER dominates. This is partly due to the observed connection between ‘not know’ and ‘not remember’ (see section 9.6) and partly due to a number of languages generally colexifying cognitive meanings, see the discussion in chapter 6 where several examples of ‘know’ and ‘remember’ sharing expression were given.

Finally, the connections to ability will be considered. This is shown in figure 10.6. It is clear that the similarity is by far the highest to the ‘know how’ verses, a connection already noted in the previous chapter (see section 9.7). It can also be seen that there are weaker connections spread more or less evenly across the domain—this reflects languages where ‘know’ as a whole is
10.3.4 Interim summary

In summary, then, the strongest connection between ‘know’ and other concepts is via ‘find out’ contexts which are in some cases equally likely to be expressed with a ‘hear’ form. In addition, ‘hear’ is connected to a number of verses with a clearly auditory object and to ‘understand’. Sight is mainly connected to contexts of recognising and knowing a person. ‘Think’ and ‘remember’ are connected to ‘understand’ and ‘realise’ verses in addition to general, weaker connections across the domain. Ability, finally, is strongly connected to ‘know how’, though also here we see weaker connections across the knowledge domain.

10.4 Explaining the connections

In chapter 9 I pointed to semantic characteristics of the different forms that might lie behind the connections outside the knowledge domain. For instance,
I argued that the similarity between ‘understand’ and ‘hear’ is partially based on shared characteristic auditory objects and the one between ‘see’ and ‘know person’ on the centrality of the clearly visual face to social interaction. But how, more precisely, does a shared characteristic lead to full or partial co-expression? In the following sections, I outline a frame-based description of metonymical meaning change for the connections pointed out here. First, section 10.4.1 gives some theoretical background. This is followed by an account of the frames and metonymical processes I suggest underlie some of the extensions of other concepts into the knowledge domain. Section 10.5 summarises and concludes the chapter.

10.4.1 Frames, profiles and metonymical change

Polysemy in a language is often thought of as being the result of historical semantic change. The process of semantic change in a language can be represented as in 10.3, following Koch (2016: 24–27), where $T_0$ is a point in time in a language variety’s development, $T_1$...$T_i$ are subsequent points in time, $M_i$ are meanings and $F_0$...$F_i$ is a form in the language, with or without changes. Thus, a semantic shift consists in a form $F$ with the meaning $M_1$ (step 1) taking on the meaning $M_2$ (step 2), first only in specific discourse contexts but with increased use $M_2$ becomes first a bi-meaning of $F$ (step 3) and then a fully parallel meaning (step 4). The process so far (steps 1-4) is sometimes known as innovative meaning change. The continuation of the process (steps 4-7) consists in the loss of the old meaning and is sometimes known as reductive meaning change. It is between these two processes (i.e. at steps 3-5) that a situation of polysemy is found in a language, a situation which can continue in a language apparently indefinitely without the new meaning taking over. In other words, the process described is not a deterministic diachronic development; two meanings may co-exist in a stable state as (equal or non-equal) meanings of a form without the next step in the process taking place. Here, we will be concerned mainly with innovative meaning changes.

On this model, the basis for semantic change lies in discourse innovations—ad hoc innovations by speakers in communicative situations (Koch 2016). A distinction is often made here between speaker- and hearer-driven changes. In speaker-driven changes, a speaker makes use of some trope to innovatively use $F$ for $M_2$, expecting the hearer to infer that this is the intended meaning. In hearer-driven change, the speaker is using $F$ to mean $M_1$ but the hearer interprets them to mean $M_2$—i.e. reanalysis (Koch 2016, 2012b, see also e.g. Detges & Waltereit 2002). I will not here be concerned with the distinction between speaker- and hearer-driven language change, as the explanation for the expression of the knowledge domain I propose is applicable with both.
The steps in the process of semantic change are not actually clearly discrete stages but a continuous change. We can thus alternatively represent the process as in Figure 10.7 below. A given co-expression in a language at a given point in time can be placed on this continuum. So far in this thesis, I have spoken mainly in terms of meanings being colexified—this implies a position on the far right of the scale. For a language to colexify something, according to the classical definition of François (2008: 170) means that “given two functionally distinct senses [—] it [the language] can associate them with the same lexical form”—we are dealing here with co-expression incorporated into the language system. I will argue, however, that many co-expressions in the knowledge domain are more fruitfully viewed from the point of view of the left-hand side of the continuum, i.e. as situational discourse innovations. Applying this lens to the knowledge domain allows us to explain the majority of co-expressions within the domain with a single model.

An important part of the semantic change process has thus far been glossed over, namely what I above termed “use of some trope” in moving from one meaning to another; how can the step from an old meaning to a new meaning be made while maintaining understandability? Two processes have been seen as central to this question: metonymy and metaphor. Metonymy refers to extension based on contiguity, whereas metaphor refers to extension based on similarity. As examples, consider the shift from Latin coxa ‘hip’ to Italian coscia ‘thigh’ (change based on contiguity, i.e. metonymy) or Old English bælg ‘bag, purse’ to Middle English bely ‘body part between the breasts and thighs’

---

**Table 10.3:** Schematic representation of a semantic shift (Koch 2016)

<table>
<thead>
<tr>
<th>Step</th>
<th>Time</th>
<th>Form</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$T_0$</td>
<td>$F_0$</td>
<td>$M_1$</td>
</tr>
<tr>
<td>2</td>
<td>$T_1$</td>
<td>$F_1$</td>
<td>$M_1(M_2)$</td>
</tr>
<tr>
<td>3</td>
<td>$T_2$</td>
<td>$F_2$</td>
<td>$M_1 &gt; M_2$</td>
</tr>
<tr>
<td>4</td>
<td>$T_3$</td>
<td>$F_3$</td>
<td>$M_1 = M_2$</td>
</tr>
<tr>
<td>5</td>
<td>$T_4$</td>
<td>$F_4$</td>
<td>$M_1 &lt; M_2$</td>
</tr>
<tr>
<td>6</td>
<td>$T_5$</td>
<td>$F_5$</td>
<td>$(M_1)M_2$</td>
</tr>
<tr>
<td>7</td>
<td>$T_6$</td>
<td>$F_6$</td>
<td>$M_2$</td>
</tr>
</tbody>
</table>

---

**Figure 10.7:** Semantic change at different levels of integration in a language

$F_1$ with meaning $M_1$ is occasionally extended by speakers to $M_2$

$F_1$ with meaning $M_1$ is regularly used for $M_2$

$F_1$ has both meanings $M_1$ and $M_2$
(change based on similarity, i.e. metaphor)—both examples are from [Koch (2016)] For further discussions of metaphor and metonymy see e.g. [Barcelona (2002)](and others in the same volume) or [Kövecses (2010): chapter 12).

In section 3.1 the view of meaning as conceptualisation and involving scenes and frames was introduced. Frames are subject to construal, i.e. different ways of viewing the same frame. The construal process profiling was specifically described—some part of the frame is put into focus as the profile against the rest of the frame which is the background. What is important about profiling for the purposes of this chapter is that it allows for meaning changes within a frame—by a shift in profile, the meaning of a word can change ([Koch 2016: 40]). Since the changes are contiguity-based, this constitutes metonymical change. The process is described by [Koch (2016: 40)] as follows “[a]n element $E_1$ of a given frame (= source concept SC) shifts from figure [=profile] to ground [=base], while a contiguous element $E_2$ (= target concept TC) shifts from ground to figure” (Koch also consider the case where the ground as a whole becomes the figure, but this is of less relevance to us).

In summary: the meaning of a constructions can be understood as a profiled part of a frame which, crucially, as a whole is presupposed by the construction. By metonymy—extension based on contiguity—what part of the frame is profiled can shift. Through this, a speaker may use a construction originally meaning one part of the frame for another part. Similarly, a hearer can understand the use of a construction as referring to a novel part of the frame. If this discourse-based shift of meaning is frequent and salient enough, the result is semantic change in the language.

To illustrate the process with an example, consider the shift of Medieval Latin *collātio* from meaning ‘reading of the *Collationes*’ to ‘light meal after reading of the *collatio*’, as discussed by [Koch (2016) 38–40]. The first meaning is understood against a frame of monastic life, including the various actors as well as the place of the readings in relation to other parts of monastic life, including the light meal following the readings. This frame can then be construed with not the readings but the meal profiled (by the speaker or the hearer—this is the difference between speaker- and hearer-induced change) and the meaning of *collātio* can shift.

An important feature of this model is that it allows for what has generally been regarded as inter-domain shifts, e.g. between perception and cognition ([Sweetser 1990]). The reason for this is that many domains appear to share frames, and whether a given expression of it ends up in one domain or another depends on construal. [Rakhilina, Ryzhova & Badryzlova (2022: 255)] discuss this as “overlap between contiguous [semantic] fields”. These frames which may belong to more and than one domain can be termed bridging frames and their presence results in contiguous domains (cf. [Rakhilina, Ryzhova & Badryzlova (2022: 255)]).
In the following, I will argue that some frames of the knowledge domain can also be construed as frames belonging to other domains through profiling, explaining how forms from these domains come to be used for knowing.

10.4.2 Bridging frames of the knowledge domain

Beginning with the by far strongest connection in the data, between ‘hear’ and the ‘find out’ contexts, it was noted that some of these contexts are in fact equally likely to be expressed by a ‘hear’ form as a ‘know’ form. The underlying frame being expressed in these verses is one of information transfer through spoken language which, following well-established models of communication going back as far as Shannon (1948), can be described as consisting of someone who receives the knowledge (=A), someone who utters the piece of information (=B), the act of utterance, the auditory perception of it and the incorporation of it in the mind of the receiver (clearly, there are countless more possible elements, but these are are the ones which are evidenced by cross-linguistic expression). The frame is depicted in figure 10.8.

Expression with a ‘know’ form is explained by a construal where the part marked i is profiled—the (metaphorical) entry into the mind of the hearer or, in other words, the resultant knowledge state. This, however, is only one possible profiling of the frame. We find evidence for two additional construals in the data—a ‘hear’-profile where the information being auditorily perceived by A is profiled (ii) and a ‘say’-profile where the utterance of the information by B is profiled (iii). These correspond to the linguistic expression of ‘find out’ events by ‘know’, ‘hear’ and ‘say’ respectively.
Figure 10.9: A coming-to-know event with different possible construals: i) knowledge of a person B enters the mind of person A, ii) person B is seen by person A

The model of knowledge transfer expressed by this frame may remind the reader of the conduit metaphor. In a famous paper, Reddy (1979) argued that our ordinary way of speaking (English in particular, but the argument can be extended to many other languages as well) shows that we think of communication as being analogous to taking a message—which is like a thing—putting it into words (note that this expression is evidence of the metaphor), sending it via speech or writing to a receiver who unpacks it and digests the message. One of Reddy’s points is that this is a mistaken view of how communication works. I will not here, however, go into his arguments for why this is the case, as the main point here is that, as argued for by Reddy, everyday language reflects this model of information transfer—thus, the same model I have suggested based on the cross-linguistic variation in expression is also supported by arguments from English usage.

The frames of recognition (of a person) are, I argue, quite similar—as we saw, this is where the connection to sight is the strongest. As in the previous frame, two participants are involved—A and B—though the utterance-action of participant B is not included. Crucially, this means that hearing is replaced by sight—sight being the main way in which we identify each other (although, as was seen above, this dominance of sight can be challenged by context-specific factors such as recognising a voice). The frame is shown in figure 10.9.

In theory, it seems, we might also find frames where what is seen and thus known is not a person (or an object) but an event, and an extension of ‘see’ into ‘know’ by this route. The evidence in the data for this is weak. There is the mentioned verse to which the connection to sight is strongest—Acts 19:34, “But when they recognized that he was a Jew”—where we find ‘see
Figure 10.10: A coming-to-know event with different possible construals: i) knowledge enters the mind of person A, ii) some information is processed in the mind of person A

that’ for ‘came to know that’. Note, however, that what is known here—while syntactically often a clause—is semantically person-like.

This lack of extension from seeing events to knowing them might be partially a result of the verse sample and, in turn, the contexts of the New Testament. The possibility of this kind of extension can not be ruled out. Nonetheless, the recurrent connection cross-linguistically between ‘see’ and ‘recognise’ and ‘know person’ suggests that frames containing a person as the object of knowledge are the most salient frames connecting knowledge and sight. That is, it is people who are prototypically known through sight.

Moving on to how ‘think’ extends into the knowledge domain, I argued that this is through meanings with a clear component of internal processing. The frame I suggest as the basis for the extensions therefore does not contain a component external to knower; instead, the cause of the knowledge is internal processing. Profiling this component of the frame leads to expression in terms of ‘think’. The frame is shown in figure [10.10]

Metonymy based on shared frames is also a possible explanation for the extension of ‘be able’ to ‘know how’. It should be clear that a frame which contains knowledge-how usually also contains ability and as I discuss in [2.2.1] the line between them is not clear. An extension from the ability to do something to the knowledge-how to do something should perhaps therefore not even be seen as metonymical shift, but as a result of this blurred boundary.

The explanations offered here cover mainly full colexification—where two senses share the exact same form—but as I have shown, there are also many cases of partial colexification in the data. For example, ‘understand’ is frequently expressed by some complex form containing the form for ‘hear’. Can the model presented here account also for cases of partial colexification—where the forms for senses are partially the same—and can it be used to explain
why some extensions seem more likely to be realised as partial colexification than others?

Drawing on the discussion of isomorphism in chapter [8] I suggest that part of the explanation is that whereas full colexification is explained by profile shifts within a frame, partial colexification is the result of shifts between distinct but similar frames. Take for example understanding and hearing. ‘Understand’ is conceptualised as containing components which are not shared by ‘hear’—e.g. the depth discussed in section 2.2.4. However, much of the conceptualisation is the same as for ‘hear’—perhaps due to the shared characteristic objects, as I argue in 9.3.3. This opens the possibility of expressing the ‘understand’ frame by means of the ‘hear’-frame. But if the difference between ‘understanding’ and ‘hearing’ is also deemed communicatively important, full colexification will be less appropriate and ‘understand’ will be expressed as being like hearing but with something more. In accordance to the principle of isomorphism, this something more semantically is reflected as something more formally.

10.5 Summary and conclusions

To summarise, there are recurrent cross-linguistic connections between the knowledge domain and other concepts. I have shown in this chapter that these connections are not distributed equally over the different parts of the knowledge domain and that there are in fact clear patterns in what parts are connected to which concepts. Hearing is primarily connected with ‘find out’ contexts—these are in fact, arguably, straddling the boundary between knowledge and perception—and also with ‘understand’. Sight is primarily connected to recognising and knowing people, whereas the closely intertwined ‘thinking’ and ‘remembering’ are connected primarily to ‘understanding’ and ‘realising’. Finally, there is a strong connection between ‘be able’ and ‘know how’.

A greater attention to this internal diversity may offer insights into the debate outlined above of whether it is ‘to hear’ or ‘to see’ which is the primary source for ‘know’-words. I suggest that it may be both, depending on which part of the knowledge domain one looks at.

Based on cognitive linguistic models of semantic change, I then sketched a process whereby forms enter the knowledge domain by metonymical shift within frames. Two points about this may be emphasised here. The first is that I am not arguing that the processes described here are the only ways in which forms enter the knowledge domain—in the previous chapters I discuss examples where metaphor may play a role (such as the motion based form discussed in 9.3.3), and there are surely others.

The second point is that although the model of semantic change applied
here is based on discourse innovations, the kind of data used in this thesis does not allow us to draw firm conclusions about the extensions’ integration into the language systems—that is, if they are emergent innovations or conventionalisations. Thus, we do not know if when in a translation, for example, ‘hear’ is used for ‘find out’ this is an extension based on metonymy within frames or a result of ‘hear’ and ‘find out’ conventionally being colexified in the language. Also in the latter case, however, the account given here offers a possible explanation for the historical basis of the underlying semantic shift.
11. The dimensions of knowledge: Typology

11.1 Introduction

Linguistic typology, in a definition by Comrie (2001: 25), can be characterised as “the systematic study of cross-linguistic variation”. The typologist thus investigates questions concerning what is universal in language, what is rare or common and what the limits of variation are. This chapter concerns these kinds of question in relation to knowledge predication, addressing research question B4:

B4. How are [the] distinctions [in the knowledge domain] expressed in the languages of the world?

Linguistic typology began with attempts to classify languages as wholes into types—isolating, synthetic etc. (see e.g. Ramat 2010). Quickly, however, the difficulties of doing this in an empirically adequate way were recognised and typology moved on to considering parts of languages—specific, usually grammatical, phenomena such as word order or case marking. Largely, the approach still consisted in identifying discrete types and assigning languages to these. Illustrative examples are the word order typologies of Greenberg (e.g. 1966) and then Dryer (e.g. 1992) or, more lexically oriented, the satellite vs. verb framing typology of Talmy (1985). Recent works, however, have questioned whether it is empirically adequate to classify even subparts of languages into discrete types. Types are inevitably generalisations, and letting them be the basis of analysis can obscure relevant facts. As an example, consider the results of Levshina (2016) suggesting a north–south continuum for the distribution of verbs for ‘letting’ in the languages of Europe—a result which would be impossible with traditional discrete types. Instead, Levshina (2016) and others taking this approach consider the linguistic behaviour of concrete tokens or exemplars.

Wälchli & Cysouw (2012: 703) call this a “typology without types”, describing it as an approach where “generalizations can be made without there being any need to reduce the attested diversity of categorization patterns to discrete types”.

245
Research Question | B4. How are [the] distinctions [in the knowledge domain] expressed in the languages of the world?
---|---
Operationalisation | What does the distribution of the markers for different dimensions look like—in the verse sample and in the language sample?
Method(s) | CMDS coordinates, Point-Biserial Correlation
Language sample | Full
Verse sample | Knowledge

Table 11.1: Methodological summary of chapter

It is intuitively clear that a division into discrete types is even more difficult when it comes to semantic typology. Whereas grammar is by and large a closed system in a language and the possibilities can often be exhaustively listed (e.g. SVO, VSO etc.) the lexicon is infinitely expandable. Thus, while it is possible to determine trends and patterns in the data of this investigation, establishing discrete types is less useful.

In the following, therefore, the range of variation in the expression of the six dimensions described in the preceding chapters will be presented with an aim of minimal data reduction. I will not attempt to fit the data into types, but consider the concrete distribution of forms in the sample, quantified, when useful, in various ways. This allows for pointing out general patterns and tendencies and, at time, the identification of proto-types—ways of doing things that languages appear to cluster around. The full language sample is considered, though as in chapter 9 the dimensions used are those of chapter 7, that is, based on the core language sample. Section 11.2 considers the marking of know-as-achievement vs. know-as-state, 11.3 considers the distribution of forms for ‘understand’ and 11.4 deals with how ‘recognise’ and ‘know person’ meanings are expressed. In section 11.5 the ‘realise’ markers found in the sample are considered, in 11.6 the typology of lexical negation of knowledge is considered and 11.7 deals with knowledge-how. Section 11.8 summarises and concludes. Table 11.1 summarises the methodological details of the chapter.

11.2 Achievement and state

In section 9.3.2 the fundamental division of the knowledge domain into state and achievement was characterised from a semantic point of view. In this section, I consider the expression of this distinction in the languages of the sample. The focus will be on languages where the expressions of know-as-
state and know-as-achievement share some lexical element. Logically, there are four possibilities regarding the marking of state and achievement in these cases:

1. Neither state nor achievement is marked
2. Achievement is marked relative to state
3. State is marked relative to achievement
4. Both achievement and state are marked

Examples of languages employing all four possibilities are found in the sample, though languages of type 2 are the most common.

As a preliminary example of the first type—type 1—we can consider English, which, as mentioned in section 2.2.7, has been analysed as allowing both state and achievement meanings of know. The same appears to be true of Middle English, which is included in the sample. As another example of a language showing this pattern, consider Moro. The form -liyet is used without any overt aspectual marking both for the stative sense (94a) and the achievement sense (94b).

(94) Moro (Heibanic, Sudan)
   a. Acts 19:15
      \[i-g-a-liyet-o\] yesu
      1SG-CL-FIN-know-?too PR.NAME
      ‘I know Jesus’ (AS)
   b. Acts 19:34
      orn ndo leda \[la-liyet-o\] ma gënëmu gafo yawud
      but when people PL-know-3SG.O that he be Jew
      ‘But when they recognized that he was a Jew’

An example of type 2 is Trinitario-Javeriano-Loretano, already discussed above (example 58 on p. 185), and repeated below. The verb -echo- is used for both state and achievement, but as an achievement it is marked with the Perfective -po.
(95) Trinitario-Javeriano-Loretano (Arawakan, Bolivia)

a. Mark 12:14

Tata, \textit{v-echo} te 'to \textit{p-echji-s'o} to father 1PL-know PREP COMP 2SG-speak-APPL.ACT ART

\textit{je'chugne}

truth

‘Teacher, we know that you speak the truth’ (AS)

b. Acts 19:34

\textit{Etotsero} te \textit{na-echo-po} to \textit{jurion-ira-ri’i ma} but PREP 2PL-know-PFV COMP jew-NMLZ-? ART

\textit{Alejandro}

PR.NAME

‘But when they recognized that he was a Jew’

As an example of type \textit{3}, consider Mapudungung, also discussed above (see example 65)—the example is repeated below. Here, \textit{kim-} is used for both state and achievement, but it is the state meaning that is marked, by \textit{-nie}, the Present progressive.

(96) Mapudungun (Araucanian, Chile)

a. \textbf{Smeets (2008: 194)}

\textit{kim-nie-ke-n} kuyfì ĕi nge-nu-n know-PRES.PROG-CF-1SG.IND formerly 3POSS be-NEG-NMLZ

\textit{traktor}

tractor

‘I know that in the old days there were no tractors’

b. Luke 9:11

\textit{Welu ti pu che kim-lu} Jesus ĕi

but DEF COLL person know-TEMP PR.NAME 3POSS

\textit{amu-n} kañpüle go-VERBAL.NOUN other.side

‘But when they learnt that Jesus was going to there’

The last type \textit{4} is the most difficult to find clear-cut examples of (perhaps due to economy constraints) but a possible example is Naro, where—as we
saw in section 9.3.2—the achievement sense is expressed by a serial construction with the verb ‘see’ or ‘hear’. Also the state sense, however, generally appears overtly marked for aspectual construction in the sample, with the Imperfective *hāa* as well as the Perfective form of the main verb, which with some verbs, such as ‘know’ and ‘not know’, has a present tense interpretation (Visser 2013: 194).

(97) Naro (Khoe-Kwadi, Botswana)


\[
\begin{align*}
A \ ba \ a \ ẽem \ ko \ bōò-a \ q’āa \\
\text{and 3SG.M DEM when PTCP see-CONN know}
\end{align*}
\]

‘When he found out’

b. Mark 14:68

\[
\begin{align*}
C’úùa-r \ hāa \ [-]
\end{align*}
\]

not.know-PFV IPFV

‘I neither know...’

The strategies for marking described above can be quantified using the directionality measure introduced in section 8.2.1. There, the measure was used to consider the directionality of formal motivation between the clusters KNOW, UNDERSTAND and FIND OUT, but it can also be applied to the relationship between state and achievement (defined by Constructional Dimension 1). Unlike in chapter 8, the values are here considered language by language, and the direction by whether the number is positive (state to achievement) or negative (achievement to state).

To represent how marked the achievement and state sense are in an absolute manner (i.e. not relative to each other) the mean of the directionality measure between groups are used. This is used in preference to a general count of the number of markers, since the coding schema differed from language to language and did not have as a general purpose to reflect the complexity of forms (cf. section 3.6.2). A relative measure is therefore most appropriate.

One further important part in understanding the numbers presented below is how productive the phenomena underlying the patterns are. The numbers used don’t tell us anything about whether, for example, knowledge achievements across the board are expressed in terms of knowledge state or if this is limited to one or a couple of contexts. To take also this into consideration, we count the proportion of contexts in each group which share at least one lexical element with the other group.
The results are shown in figure 11.1 below, where the x-axis shows the directionality and the y-axis the total amount of marking for state or achievement. Languages are represented by their ISO code (see Appendix A) and their color represent the macro-area to which they belong. The size of the labels, finally, reflects the proportion of verses between state and achievement which share lexical elements—the bigger, the higher the proportion. It appears from the results that the observation that the combination of no directionality and much marking—i.e. marking of both senses—is relatively rare, and that there is a clear relationship between an increase in directionality and amount of marking. Note that this connection is in part mandated by the method—as the average amount of marking approaches zero, so does necessarily the directionality.

Naro (nhr) has extreme values on both dimensions. A closer analysis reveals that this is due to the almost consistent expression of stative knowledge in the construction exemplified above with both the lexical verb inflected and an auxiliary (97). In addition, also the achievement sense is marked—this contributes to the high value of average amount of marking. This, as well as the apparent default status of the mentioned construction, raises the question that it ought to have been analysed as a whole in the coding, but the application of the principles described in section 3.6.2 lead to the decision to code the markers separately.

Naro is the most extreme when it comes to average amount of marking, but
Carib (car) and Teso (teo) are almost as extreme, though on the other end of the directionality parameter. This is due to a combination of two factors. The first is that in both languages, there is fairly consistent marking of achievement contexts relative to state contexts, i.e. the majority of contexts show this pattern. The second is that both languages have two markers which are used to mark the achievement sense and sometimes co-occur. This drives up the value.

Zooming in on where the majority of languages are located, it can be noted that there is an overweight of languages on the right—i.e. languages in which achievements are marked relative to states, see figure 11.2.

It is difficult to discern any areal trends in the data, other than a mild tendency for African and Eurasian languages to be found on the right-hand side and languages from Oceania to be found on the left. The same lack of clear areal trends can be seen when the values are plotted on a map, as in figure 11.3, where colors represent the directionality values and level of transparency the proportion of contexts across groups which share a lexical element.

11.3 Understanding

Next, the expression of ‘understand’ in the sample is explored. The relationship between knowing and understanding has in previous research generally been framed in terms of colexification, i.e. if the same form is (or can be) used...
for both concepts, or if they are expressed separately. The analogous approach using the data of this investigation would be to ask if the clusters that we have identified as ‘understand’ and ‘know’ are expressed by the same forms or not. Answering this question simply in terms of yes or no, however, would be a considerable simplification of the way the data actually looks. It is true that there are languages where we find no distinction between ‘know’ and ‘understand’ as well as languages where a separate form is used consistently in the ‘understand’ cluster. As commonly, however, there are forms which have a partial extension in the ‘understand’ cluster, and which may or may not have uses outside of it as well. Rather than speaking of ‘understand’ and ‘know’ being colexified, therefore, it is more useful to speak of them as being more or less similarly expressed.

Using the same methods introduced in chapter 5, this similarity—the degree to which a language co-expresses two clusters—can be quantified. The similarity measures discussed there are based on pairwise comparison of the coding of verses. Once a clustering algorithm has been applied to these data, the similarities can be averaged across these clusters, giving a measure for how similar two clusters are in any given language. In map 11.4, the similarity between the clusters UNDERSTAND and KNOW in the languages of the sample

**Figure 11.3:** Map showing the directionality values between state and achievement in the languages of the sample.

![Directionality of state-achievement relationship](image-url)
Figure 11.4: The similarity between the UNDERSTAND and KNOW clusters in the languages of the sample are displayed. 1 (a fully filled pie chart) represents identical coding across all verses in both clusters, and 0 (a pie chart with no filling at all) represents no similarity between any of the verses.

As can be seen, there are relatively few languages which are positioned at the end poles of the continuum total colexification–total differentiation. There are also clear areal patterns—note the dominance of very low similarity across North America and Eurasia, and the groups of high similarity in South America, Africa and New Guinea.

In general, intermediate values can come about in two ways. Firstly, by some verses in the clusters being expressed similarly and some differently. In other words, there is in the language a form with extension in the ‘understand’ cluster, but a limited extension. Secondly ‘understand’ and ‘know’ may share a lexical root but be distinguished by constructional markers. Due to the way similarity is calculated, this results in an intermediate value, even if all the verses in a cluster are marked.

The latter case is discussed in section 9.3.3—here the focus will be on the extension of constructions within the UNDERSTAND cluster. In chapter 9 the Point-Biserial correlation was introduced to determine what markers correlate with what dimension, and thereby a number of markers which on their own
or in combination with other markers mean something like ‘understand’ were identified. Their form and its implication for semantics was discussed, but nothing was said of these markers’ distribution in the sample. This is what we will turn to now.

One way to approach this issue would be to simply consider the markers’ distribution one by one. This, however—in addition to giving results difficult to overview—would ignore the important fact that the distribution of markers sometimes overlaps. That is, some markers occur together as constructions. To represent this, the distribution of markers is represented language-wise. If there in a language is only one marker which correlates with the relevant dimension—in this case Dimension 1 of a lexical only analysis—all verses where this marker appears are marked with 1, and all others with 0. If there are several markers, we need to determine if they co-occur. If they do not, the matter is simple and the most frequent markers is coded with 1, the second most frequent with 2 and so on. If there is overlap, however, it is assumed that the markers make up a construction and they are coded as such. As an example, imagine that a language has two markers A and B, which co-occur. We then consider all the possible combinations of A and B (A, B and AB) and rank them in frequency. The most frequent is assigned the label 1a (or more accurately $n_a$ where $n$ is the number of the set of markers under consideration) the second most frequent 1b and so on. This captures the intuition that constructions which partly involve the same markers are somehow related.

The distribution of the identified forms is shown in figure 11.5. Each column is a language and each point is a verse. Forms which have been identified as ‘understand’ forms are in color, the most frequent form in purple and the second most frequent in orange—in addition, the information of connections to other domains from 9.3.3 is incorporated. Forms related to ‘hear’ are blue, forms related to ‘know’ in green and forms which are based in motion metaphors in yellow (see the legend in figure 11.5). Constructional variants are indicate by shape. All other forms are coded here as zero, indicated by black, smaller points for ease of viewing—the figure should not be read as indicating that the expression outside of the coded forms is uniform. The languages are sorted according to how many occurrences there are of non-zero form, i.e. how many verses use one of the forms correlated with the dimension. The positioning of the verses along the y-axis reflects the first dimension in the CMDS analysis.

These figures confirm the areal patterns identified above. In Eurasia and North America, the forms identified dominate the ‘understand’ contexts (which is clearly identifiable as a cluster) whereas in Oceania and South America, many languages lack ‘understand’ forms altogether. Africa represents a somewhat intermediate case where some languages share the Eurasian pattern but
some have a form with only a handful of occurrences. Note also that the pattern of deriving ‘understand’ from ‘known’ seems particularly prevalent in Africa.

We can furthermore observe a difference in the frequency of the identified forms outside of the UNDERSTAND cluster (uppermost in figure 11.5). In the Eurasian languages, these are in general one or a couple of contexts per languages, whereas in particular in the South American languages, but also in North America and to some degree Oceania and Africa, the forms have a relatively wider distribution outside of ‘understand’. This tells us that that at least extensionally, these forms do not match completely in meaning—Eurasian type ‘understand’ refers to a smaller range of contexts than e.g. South American ‘understand’ types. This suggests that a semantic characterisation simply in terms of translation equivalence to ‘understand’ is inadequate, and that further steps to investigate the semantics of these forms—such a those in chapter 9—are needed.

11.4 Knowing people

We now move on to considering the expression of Dimension 3 of the CMDS analysis, i.e. meanings like ‘recognise’ and ‘know person’. In figure 11.6 the distribution of forms which express something like ‘know person’ or ‘recognise’ is shown. The figure is constructed in the same way as figure 11.5 but is

Figure 11.5: The distribution of forms correlated with Lexical Dimension 1 across the verses in the extended knowledge domain by area
based instead on forms correlating with Dimension 3. Also, as we have seen, it is the connection with ‘see’ rather than ‘hear’ which recurs—thus, connection with ‘see’ is marked in orange. In figure 11.6, only those languages where there is at least one marker correlated with Dimension 3 are shown; see figure 11.8 below for a view also of languages where the distinction is not found.

As can be seen, the distribution of these forms among the verses is not random. We can observe a cline from left to right, where in languages with few occurrences of the relevant forms these are found in the verses with the highest value in Dimension 3, i.e. those with a ‘recognise’ meaning (see section 9.4) and as the number of ‘know person’ forms increases, they extend gradually downward along the dimension. As can be seen in figure 11.6, the contexts fall into four more or less clearly separated clusters along the dimension. These are marked in the figure with horizontal line.

In section 9.4, I argued that ‘recognise’—corresponding to the uppermost cluster of Dimension 3—can be semantically characterised as differing from general ‘know people’ by being an achievement and furthermore by having a notion of repetition and identification involved. Can the two groups of ‘know person’ also be interpreted semantically?

A consideration of the contexts reveals that the distinction between the second and third cluster is something like the difference between knowing a
John 10:4
[—] and the sheep follow him because they know his voice.

Acts 19:34
But when they recognized that he was a Jew [—]

Mark 14:07
[—] “I do not know this man whom you are talking about!”

John 1:4
[—] “From where do you know me”? [—]

John 7:29
I know him, because I am from him and he sent me.

Acts 19:15(a)
“Jesus I know [—]”

Mark 1:34
[—] he did not permit the demons to speak, because they knew him.

Acts 7:1
until another king arose over Egypt who did not know Joseph.

Acts 19:15(b)
[—] and Paul I am acquainted with [—]”

Mark 1:24
I know who you are — the Holy One of God!”

<table>
<thead>
<tr>
<th>Verse</th>
<th>English (Lexham)</th>
</tr>
</thead>
<tbody>
<tr>
<td>John 10:4</td>
<td>[—] and the sheep follow him because they know his voice.</td>
</tr>
<tr>
<td>Acts 19:34</td>
<td>But when they recognized that he was a Jew [—]</td>
</tr>
<tr>
<td>Mark 14:07</td>
<td>[—] “I do not know this man whom you are talking about!”</td>
</tr>
<tr>
<td>John 1:4</td>
<td>[—] “From where do you know me”? [—]</td>
</tr>
<tr>
<td>John 7:29</td>
<td>I know him, because I am from him and he sent me.</td>
</tr>
<tr>
<td>Acts 19:15(a)</td>
<td>“Jesus I know [—]”</td>
</tr>
<tr>
<td>Mark 1:34</td>
<td>[—] he did not permit the demons to speak, because they knew him.</td>
</tr>
<tr>
<td>Acts 7:1</td>
<td>until another king arose over Egypt who did not know Joseph.</td>
</tr>
<tr>
<td>Acts 19:15(b)</td>
<td>[—] and Paul I am acquainted with [—]”</td>
</tr>
<tr>
<td>Mark 1:24</td>
<td>I know who you are — the Holy One of God!”</td>
</tr>
</tbody>
</table>

Table 11.2: Verses with high values in dimension 3, from high to low. The horizontal lines in the table correspond to the lines in figure 11.6

person directly and knowing about a person. Table 11.2 gives the relevant verses.

Among these, Acts 19:34 sticks out somewhat. This can be seen also in how the Lexham version translates the verse—with recognise, which, as discussed, is more characteristic of the first cluster. The reason for this context’s placement here seems to be twofold. On the one hand, it concerns identification. It is also clearly an achievement (“when the people”). This makes it similar in semantics to other ‘recognise’ verses. On the other hand, the fact that the verse in most translations has a clausal object (“that he was a Jew”) makes it more similar to contexts further down on Dimension 3, i.e. ‘know that’. These factors—in addition to a tendency for both this verse and, as discussed in the previous chapter, other ‘recognise’ verses to be expressed with forms for ‘see’ (see section 10.3.2)—gives the verse this intermediate position.

Also the verse with the highest value in the cluster—John 10:4—seems to be somewhat anomalous, and can explained by similar factors. We have the achievement and identification components making it similar to ‘recognise’, but it differs in taking a voice rather than a person as object—as will be seen below, non-person nominals vary considerably in their expression from language to language. As was already observed in section 10.3.1 a voice as object is also sometimes reflected by an expression including ‘hear’, which makes the verse more similar to the low part of the dimension.

With these preliminary remarks, the general characteristics of the verses
in table 11.2 can be considered. The verses in the second group all involve an element of direct contact—either simply via perception or via what was in section 2.2.2 referred to as reciprocal causal contact. These contexts will therefore be referred to as direct personal knowledge. The verses in the third cluster, on the other hand, express knowledge which does not imply direct contact. In some cases, direct contact is actually ruled out by the context, e.g. in Acts 7:10 where the object of knowledge has been dead for generations. This group will be referred to as indirect personal knowledge. In short, the following scale describing the extension of forms can be posited:

Recognise (a person) < Know a person (directly) < Know a person (indirectly)

If a form distinct from the general word for ‘know’ is used for some meaning on the scale, it will also be used for meanings to the left of it, barring the use of some other form distinct from general ‘know’.

With these meanings identified, we can then characterise the expression of Dimension 3 across languages by identifying four prototypes. As stressed in the introduction to this chapter, actual linguistic facts at the token level are inevitably messy and as can be seen in figure 11.6 there is considerable variation at the level of individual languages and verses. The types described below should therefore not be understood as discrete classes into which languages fall but rather as idealised prototypes around which languages appear to cluster. There will be no attempt to classify the languages into these types and the cross-linguistic patterns of distribution will instead be approached by considered individual languages which are close to the prototypes.

That said, the following prototypes can be identified:

1. The Recognise type—only the ‘recognise’ meaning is distinguished
2. The Direct personal knowledge type—‘recognise’ and direct personal knowledge are distinguished
3. The Indirect personal knowledge type—‘recognise’, direct personal knowledge and indirect personal knowledge are distinguished
4. The Extended personal knowledge type—‘recognise’, direct personal knowledge and indirect personal knowledge in addition to some other contexts are distinguished

Among the last three types, there is furthermore a distinction between those languages where ‘recognise’ is distinct from the other ‘know person’ meanings and those where it is not. Among the first, we can distinguish between those languages where ‘recognise’ is related to ‘know person’ and those where it
is not. The issue of co-expression with ‘see’ is something that could also be addressed from an onomasiological viewpoint, but this will not be pursued here.

As an example of the Extended personal knowledge type, consider Modern Standard Arabic (henceforth simply Arabic) where we find the two verbs ‘alima and ‘arafa for ‘know’. The latter is the one used for ‘recognise’ and personal knowledge (example 98a) but the use is much wider than that, as example 98b illustrates. In fact, it seems more reasonable to describe the other verb, ‘alima, as having the restriction of only taking objects with a propositional meaning, and consider ‘arafa the, so to speak, semantically unmarked of the pair. Examples 98c-98d illustrate the use of ‘alima; the most common object is clausal (as in 98c) but we also find nominal objects, but only when object is nonetheless clearly interpretable as having propositional content. Thus, in 98d ‘thoughts of their heart’ clearly has propositional content (cf. “he knew what they where thinking” or “that they though like this”).

(98) Modern Standard Arabic[1](Afro-Asiatic, Saudi Arabia)

a. Mark 14:71

‘inni lā ‘a’rifu hadhā il-rajul-a alladī
PRT NEG know.1SG.IMPF this DEF-man-ACC REL
tatāhaddathīna
speak.2PL.M.IMPF

‘I do not know this man whom you are talking about!’

b. Matt 25:26

‘arafa ‘ann-ī ‘ahṣudu min ḥaytu lam
know.2SG.M.PF that-1SG harvest.1SG.IMPF from where NEG
‘azra’wa‘ajmadī min ḥaytu lam ‘abdūr
plant.1SG.IMPF.JUSS and-gather.1SG.IMPF from where NEG

‘You knew that I reap where I did not sow and gather from where I did not scatter seed.’

c. Mark 13:28

ta’lamūna ‘anna aṣ-ṣayf-a qarīb-un
know.2PL.M.IMPF COMP DEF-summer-ACC near-NOM

‘you know that summer is near’

---

1Transliteration of the Arabic follows Versteegh (2014: xiv–xv)
d. Luke 6:8

\[
\text{wa-}\text{'id }\text{'alima }\text{yasū'u niyyāt-i} \\
\text{and-then } \text{know.3SG.M.PF } \text{PR.NAME.NOM thought.PL-ACC} \\
\text{qulūb-i-him} \\
\text{hearts-GEN-3PL.M}
\]

‘But [Jesus] knew [the] thoughts [of their hearts]’

In support of the view of ‘arafa as the unmarked member of the pair, it can be noted that in at least two Arabic varieties—Maltese and Shuwa Arabic—the ‘know that’ form ‘alima form is no longer found (not in the NT and not in other materials available to me) and the form related to ‘arafa is the one used across the board.

Is the distinction made in this type of language appropriately described in terms of an opposition between knowing facts and knowing people? As discussed in section 2.2.3, language-specific studies of some European languages (e.g. Vatrican 2004), suggest a broader characterisation of the opposition in these languages (we return to this issue below in connection to non-personal nominal objects). I do not argue against these language-specific analyses, but only note that from a cross-linguistic perspective, it does appear that knowing people plays a special role which should be acknowledged in the analysis.

As an example of the Indirect personal knowledge type, consider Yuracaré, where the form for ‘know person’ -iyepe—contrasting with -ile- for ‘know that’—is clearly used for both direct and indirect knowledge of people, as shown in 99a,99b. There is no example in the verse sample of -iyepe- being used with a clausal object—cf. Tlachichilco Tepehua as shown in example 53c in chapter 7—but it is clear from the contexts that the meaning of -iyepe- cannot be limited to direct personal knowledge of someone. For example in 99a the speaker is a demon possessing someone addressing Jesus who has just entered a temple and the words are in reference to Jesus’ identity as the son of God. The English also indicates this meaning: “I know who you are” rather than “I know you”.

\footnote{With the allomorph \text{nnepe} in 99a}
Yuracaré (Isolate, Bolivia)

a. Mark 1:24

Së-ja mi-nnep-i mëë, naa Dios a-Bonto
1SG-TOP 2SG.OBJ-know-1SG.SBJ 2SG DEM god 3SG.POSS-son

yij-yitaj-bëshë-m-ti-la
REDUP-good-thing-2SG.SBJ-DS-INS

'I know who you are—the Holy One of God!'

b. John 7:29

Së-ja emejmesh iyepe-i-la ati naa lati-ja
1-TOP well know-1SG-INS 3SG DEM 3SG-TOP

ti-bacheti
1SG.OBJ-send

'I know him, because I am from him and he sent me'

c. Luke 8:53

Mammui ca-l-aimimë-u Jesús
totality 3SG.O-AFF.O-laugh-3PL.SBJ PR.NAME

ma-ile-ti-la naa tuhui-ti-la
3PL.OBJ-know-DS-INS DEM dead-DS-INS

'And they began laughing at him, because they knew that she was dead'

Yuracaré is also an example of a language where the forms for ‘recognise’ and ‘know person’ are the same—*iyepe-* is used for both.

As an example of the Direct personal knowledge type, there is Swedish. Consider the examples in 100 below; *känna* on its own is used only in cases of direct personal acquaintance (100a). For ‘recognise’ meanings we find the particle verb *känna igen* lit. ‘know again’ (example 100b), thus also providing an example of ‘recognise’ being distinct from but related to ‘know person’.

In any non-personal acquaintance cases of knowing a person, either some construction with a particle is used or instead a construction with *veta* ‘know that’, which generally takes clausal objects (100c).
(100) Swedish (Indo-European>Germanic, Sweden)
a. John 7:29

\[ \text{jag} \ \text{känner} \ \text{honom} \ \text{därför att} \ \text{jag} \ \text{kommer} \ \text{från} \]

\[ 1\text{SG} \ \text{know.PRS} \ 3\text{SG.M.O} \ \text{because} \ \text{COMP} \ 1\text{SG} \ \text{come.PRS} \ \text{from} \]

\[ \text{honom} \]

\[ 3\text{SG.M.O} \]

‘I know him because I am from him’

b. Mark 6:54

\[ \text{så} \ \text{snart} \ \text{de} \ \text{steg} \ \text{ur} \ \text{båt-en} \ \text{kände} \ \text{folk-et} \]

\[ \text{as} \ \text{soon} \ \text{3PL} \ \text{step.PST} \ \text{out.of} \ \text{boat-DEF} \ \text{know.PST} \ \text{people-DEF} \]

\[ \text{igen} \ \text{honom} \]

\[ \text{again} \ 3\text{SG.M.O} \]

‘as soon as they stepped out of the boat, the people recognised him’ (AS)

c. Acts 19:15

\[ \text{Jesus} \ \text{känner} \ \text{jag} \ \text{till}, \ \text{och} \ \text{vem} \ \text{Paulus} \ \text{är} \]

\[ \text{PR.NAME} \ \text{know.PRS} \ 1\text{SG} \ \text{to} \ \text{and} \ \text{who} \ \text{PR.NAME} \ \text{be.PRS} \]

\[ \text{vet} \ \text{jag} \]

\[ \text{know.PRS} \ 1\text{SG} \]

‘Jesus I know, and Paul I am acquainted with’

Finally, we have the Recognise type with Central Alaskan Yupik as an example. The form elitag- is used in positive ‘recognise’ contexts in the verse sample (example 101a), whereas in negative contexts, the form nallu- ‘not know’ is used (example 101b), which is also the form which is used, negated, as a general ‘know’ form, including knowing people (101c.

(101) Central Alaskan Yupik (Eskimo-Aleut, USA)
a. Mark 6:54

\[ \text{tamaa} \ \text{yuut} \ \text{elitag-aat} \]

\[ \text{immediately} \ \text{people} \ \text{recognise-3PL}>\text{3SG.IND} \]

‘people recognized him immediately’
b. Matt 17:12

\[\text{ta}’\text{ggaam nallu-at}\]

but not.know-?

‘[Elijah has already come,] and they did not recognize him’

c. John 7:29

\[\text{wiingga nallu-nrit-aqa}\]

1SG not.know-NEG-1SG>3SG.IND

‘I know him’

As for languages where ‘recognise’ and ‘know person’ are entirely distinct from each other, in addition to ‘know that’, this appears to be rare, but an example is Thai, where the form for ‘know person’ is \(\text{riu}-\text{càk}\) (the combination of two roots meaning ‘know’, with the first being used on its own for ‘know that’, see p. 212) and the form for ‘recognise’ is \(\text{cam dày}\) ‘hold GET’. Examples are given in (102).

(102) Thai (Tai-Kadai, Thailand)

a. John 7:29

\[\text{tēe raw riu}-\text{càk phraâ??oy}\]

but 1PL know-know 2/3.GOD

‘[But] I (lit. we) know him’

b. Mark 6:54

\[\text{prá?chaachon kő cam phrá? yeesuu dày}\]

people then hold HON PR.NAME GET

‘people recognized [Jesus]’

As has already been stressed, the cline from ‘recognise’ to ‘know person directly’ to ‘know person indirectly’ is a tendency with plenty of “exceptions” in individual languages and verses. Whether these exceptions are best explained by differing construals of particular verses or by details of the semantics of forms in particular languages is an open question.

One interesting case of exceptions which deserves mention, however, is Mi’kmaq. In Mi’kmaq we find two forms: \(\text{geitu-geji}\) ‘know (animate/inanimate object)’ and \(\text{nenu-}\) ‘recognise, know person’. The first verb is the general ‘know’ word, used for ‘know that’. The alternation depending on whether the object is animate or inanimate is characteristic of the entire Mi’kmaq verbal
system. The precise forms taken are unpredictable and lexically determined, but there is a great deal of regularity in the system, see Fidelholtz (1968) for details. Note that even with clausal objects, the verb geitu-/geji- sometimes takes the animate form, agreeing with the subject of the complement rather than 3rd person singular inanimate. The verb nenu- is used for ‘recognise’ (103a) and indirect personal knowledge (103b) but it is not, contrary to expectation, used in cases of clear direct personal knowledge (103c), where instead geji- is used. An hypothesis is that this unusual pattern relates to the fact that Mi’kmaq distinguishes animate and inanimate objects throughout the verbal system, thus in some sense preempting the ‘know that’/‘know person’ division.

(103) Mi’kmaq (Algic, Canada)

(a) Luke 24:31

\[ \text{na tujiw nenu-a’titl} \]

PRT then know-3PL>3SG.AN

‘and they recognised him’

(b) Acts 7:18

\[ \text{ta’n mu nenu-agul So’sep-al} \]

REL NEG know-3SG>3SG.AN.NEG PR.NAME-?

‘who did not know Joseph’

(c) Luke 7:29

\[ \text{gatu ni’n geji-’g} \]

but 1SG know-1SG>3AN

‘But I know him’

The distribution of forms in the languages discussed can be seen in figure 11.7. The figure shows the same data as figure 11.6 but with only those languages which have been discussed here shown.

There are clear areal trends in whether languages make a distinction along Dimension 3 or not, as can be seen in figure 11.8 where the same data as in figure 11.6 is displayed, but area-wise and with also languages without the distinction included. We can note that the distinction is rarely found in Africa or Oceania and that in the latter, when the distinction does appear, the connection with sight (indicated by orange) is very strong. In North and Meso-America, the distinction is present in all languages of the sample, although it should be noted here that the coverage of North America is quite poor. In South America,
making the distinction is about as common as not making it, and the connection to sight is also very clear. In Eurasia, finally, the distinction dominates completely and the synchronic connection to sight is absent.

The same data are shown in map 11.9. The size of the symbols corresponds to the number of verses where a ‘know person’ form is used, and the color to the proportion of these where the form is related to ‘see’, from dark blue (all of them), via blue (fewer) to light blue (none). It appears from the map that the distinction in Eurasia is centered in Europe, though there are also several languages in Asia where the distinction is found. We see also that the synchronic connection to ‘see’ is found mainly in New Guinea and Meso-America.

So far, the discussion has focused on verses where the object of knowledge is a person. But as we saw in section 2.2.3 and has been mentioned here and there throughout the section, forms which are used for knowledge of people may in some cases also be used for other nominals. As was already mentioned in connection with the extension of ‘recognise’ into non-person objects, the materials for investigating this in the NT are somewhat scarce. Nonetheless, the verse sample includes a number of contexts where the object of knowledge is a non-person nominal. Two initial observations can be made regarding these. The first is that there is no language in the sample where knowing non-person nominals is distinct both from knowing facts and knowing people. The second is that there is no language in the sample where only non-person nominals are distinguished, without knowing people also being so. This may seem

\footnote{The possible exception is in Dogrib (Athabaskan, Canada) where we find the form hok’èezò specifically for ‘know the way, a place’—this is discussed more below.}
Figure 11.8: The distribution of forms correlated with Dimension 3 in all languages by area

Figure 11.9: Map showing the languages in the sample and the frequency of ‘know person’ forms in them
trivial, but it is not given a priori that it is knowing people specifically which should be linguistically privileged and it suggests that, from a cross-linguistic perspective, it is the distinction between ‘know that’ and ‘know person’ which is fundamental.

The verses including non-person nominals are given in table 11.3. Due to what is available in the NT, the nominals included are mainly what I call proposition-like, to a greater or lesser degree.

In figure 11.10, the expression of these verses in the language sample is shown. Note that the majority of the verses—the two with knowing/recognising a voice excepted; see also the discussion of these above—are located in the lower part of Dimension 3. This means that generally, they are expressed in the same way as ‘know that’ contexts, rather than ‘know person’ contexts. It is possible that this distribution would be different if more concrete, less propositional objects could have been included, but this remains a question for future research.

It can be noted, however, that the extension of ‘know person’ forms also to non-personal nominal objects is not an exclusively European phenomenon. For instance, in Tlachichico Tepehua the ‘know person’ form -mispa- is used with the majority of nominals. As an example, consider John 14:4 in 104a.

The exception is Luke 6:8 (“he knew their thoughts”) where we have instead -k’atsa- ‘know’. Note, however, that in this case the object appears to be a relative clause (“what they were thinking”) rather than a noun properly, see 104b.
(104) Tlachichilco Tepehua (Totonacan, Mexico)

a. John 14:4

\[\textit{vachu’ misp’a-ya-t’ik ni lakatin}\]
also know.2-IPFV-2PL ART road

‘And you know the way’

b. Luke 6:8

\[\textit{Pero ni Jesús ix-k’atsa-y-cha yu ix-tat’ajun}\]
but(Sp.) ART PR.NAME PST-know-IPFV-1AM REL PST-be

\[\textit{pastak-nan}\]
think-ANTIP

‘But [Jesus] knew their thoughts’

The distinction between -\textit{mispa}- and -\textit{k’atsa}- is, briefly, analysed in Watters’ (1988) grammar of Tepehua. He writes that “the two verbs translated ‘know’ parallel (but do not match) the difference between Spanish conocer and saber” (Watters 1988: 70) and describes their meaning as reproduced in figure 11.11. From this brief explanation, it appears that the notion of experiential knowl-

\[\textsuperscript{1}\]

\[\textit{It is not clear to me if the variation between fully drawn and dashed lines is a result of typographical limitations or is intended to signify something; I have therefore reproduced it in figure 11.11.}\]
edge, discussed in section 2.2.3, plays a role in the choice between -mispa- and -k’atsa-. This is also the case, it appears, in Trinitario (Francoise Rose p.c.), where we also see a certain extension to non-person nominals of the ‘know person’ forms.

Finally, there is one type of object that deserves special mention, namely roads or paths. We have already encountered an example of the special linguistic status of this above; in Dogrib there is a special form for ‘know the way, a place’ hok’èezo, compared with the general form (used for both facts, people and other nominals) yek’èezo. Note that the two are formally quite similar and likely related. The two forms’ use is exemplified in 105. The object in 105a appears to in this case be a clause (“where I am going”) but 105c is an example from a Dogrib dictionary with the form without an explicit object.

(105) Dogrib (Athabaskan-Eyak-Tlingit, Canada)
   a. John 14:4
      
      edì  ts‘ò qht’ı  sił  hok‘èahso  høt’e
      where go.1SG FOC know.way.2PL FOC
   ‘And you know the way where I am going’

   b. Mark 10:19
      Moses we-nàowo=deè wek’èizø ne
      PR.NAME 3SG-law=big know.2PL be
   ‘You know the commandments (Moses’s law)’
This is the only example in the sample of a language with a special form for ‘know the way, a place’, but there is an etymological connection between ‘know’ and ‘way, path’ noted in many Eskimo languages Fortescue (2001), Estonian teadma from tee ‘road’ (Bernhard Wälchli, p.c.) and Mandarin zhi¹dao ‘know’, literally ‘know (the) way, method, direction’. These occasional but geographically widespread connections between ‘know’ and ‘road’ show that directions or places has a special place among things which can be known, which gives an interesting insight into what is considered important to know across cultures.

To summarise, Dimension 3 of the CMDS analysis is sensitive to the distinction between knowing facts and knowing people. The highest values of the dimension correspond to ‘recognise’ meaning, followed by direct knowledge of people to indirect knowledge of people. Languages in which a form dedicated to the higher values of Dimension 3 is found differ in how widely this form used, where in languages with a sparsely used form, it is used only for ‘recognise’ and as the number of occurrences grows, the form extends into ‘know person (directly)’ and ‘know person (indirectly)’. In languages with a relatively wide extension, it is not uncommon for these forms to extend also to non-person nominals. This seems to suggests—as was discussed specifically for Arabic as well as for Tepehua and Trinitario—that in some languages, the distinction is semantically characterised by something other than the ‘know that’/‘know person’ distinction, such as was discussed in section 2.2.3 regarding experiential knowledge. This, however, is a difficult hypothesis to test given the current materials, and remains a question for future research. It can also, however, be noted that the distinction between ‘know that’ and ‘know person’ has a particular role cross-linguistically, and that it thus seems that it is here a description of the knowledge domain cross-linguistically should start.

It is interesting to compare what has been found here with the philosophical theories introduced in section 2.2.2. As we saw, Benton (2017) argues that three grades of personal involvement in knowing people can be distinguished, namely:

- First grade: second-hand known facts about people, e.g. from reading about a person
- Second grade: knowledge that comes from “perceptual access”, e.g.
from seeing a person

- Third grade: interpersonal knowledge, comes from the knower’s treatment of the known as a subject and vice versa, e.g. from talking with someone

The first two grades can be reduced to propositional knowledge, in Benton’s view. The third, however, Benton views as special and terms interpersonal knowledge. In this third grade there is a “reciprocal causal contact” and the knowers treat each other as subjects.

How does the Benton’s view compare to the results presented here? At first sight, it might seem that a straightforward one-to-one mapping to the scale in [?] is possible; in this case ‘recognise’ would be the second grade, ‘know a person (directly)’ would be the third grade and ‘know a person (indirectly)’ would be the first grade. However, as I have argued, ‘recognise’ has at least two components: a person as typical object and achievement meaning. Related to both of these is a strong connection to sight. The ‘recognise’ meaning therefore does not map onto Benton’s categories and its place at the end of the scale given above can be explained by it being a salient and in several ways distinct frame.

There is no clear evidence in the data of a further distinction between Benton’s second and third grades—contexts of both these kinds fall under ‘know a person (directly)’. Note, however, that languages display considerable variation in the expression of these ‘know person (directly)’ contexts. Given the relatively small number of languages where the distinction between ‘know that’ and ‘know person’ is made at all and the contexts representing these meanings, it may be that future more in-depth investigations will reveal that distinctions of this sort are relevant in languages. For instance, I have argued that Swedish känna is clearly oriented toward this distinction.

An interesting further property of interpersonal knowledge which was noted in section [2.2.2] is that this kind of knowledge is usually symmetrical. If A has interpersonal knowledge of B, B has interpersonal knowledge of A. I suggested in section [2.2.2] that this makes interpersonal knowledge a candidate for what is known as naturally reciprocal events—events that by their nature always or typically contain reciprocity. In some languages, a class of predicates expressing such events are treated in a special manner when it comes to marking of reciprocity, and I raised the question of whether ‘know (interpersonally)’ might be part of such classes. In the languages examined here, no examples were found of ‘know person (interpersonally)’ being a member of a class of naturally reciprocal predicates. This is natural if the origin of forms for ‘know person’ is in ‘recognise’ context and not in the ‘know person (interpersonally)’ directly—‘recognise’ is not semantically reciprocal.
A final complication regarding Dimension 3 may be addressed here. It is clear from the data and previous literature that syntactic constraints play a role in choosing a ‘know’ word for an object, not only semantic ones—generally speaking, in a language making the relevant distinction one form takes (primarily) clausal objects and the other takes (primarily) nominal objects. Now, the syntactic form of an object will to a high degree correlate with its semantic features (i.e. a proposition will be clause and a human will be a nominal) which makes distinguishing what are syntactic and what are semantic constraints difficult. This too is something which should be considered by future research.

### 11.5 Realising

In figure 11.12, the distribution of forms correlating with Dimension 4 is shown. It is immediately obvious that most languages appear not to be oriented toward this dimension in their expression of knowledge predication—an expected result as we move higher up in the dimensions of the CMDS analysis. Markers associated with the Known (typically complementiser) are shown in orange and forms connected with ‘thinking’ or ‘remembering’ are in green.

It can also be noted that we find a few languages with very frequent forms correlating with Dimension 4. This is interesting, given that the semantics of
the dimension were analysed in chapter 7 as being relevant mainly to a small number of ‘realise’ verses with high values in the dimension. This distribution might have several explanations. The most typologically interesting one is that the semantics relevant to the distinction of this group are also relevant more widely.

Some of these forms are what appear to be the languages’ most general terms for ‘know’. These are Teso, where the verb -jen- ‘know’ occurs 53 times out of a total of 55 verses (as seen above, sometimes with directional affixes to modify the meaning). The correlation with Dimension 4 has its base in this very wide use of the form. The two verses which do not use the form, use instead -pup- ‘listen’ and are an ‘understand’ verse and a ‘find out’ verse—these two come together in Dimension 4 in their distinction from ‘realise’, which explains the correlation. The other language is North Tanna, where the form itun correlates with Dimension 4. The reason here is the three-way distinction in said language between -itun ‘know’, -ėruru ‘not know’ and -őto ‘hear’. Once again, the meanings which the form is not used with are found close together in Dimension 4. Note that -itun ‘know’ is also highly (and significantly with p < 0.001) negatively correlated with Dimension 2, i.e. with non-external knowledge.

More interesting is the case of Dimasa, where we find two forms for ‘know’ which appear to be roughly equally common—mithi and tharmai. The former goes back to the proto-Tibeto-Burman *syey (Matisoff n.d.). The second form, tharmai, most likely consists of thar ‘fully, completely’ and mai ‘find, receive’.

This form, tharmai, is correlated with Dimension 4. Its counterpart, however, mithi, is more strongly negatively correlated with Dimension 1 and not significantly with Dimension 4 (p > 0.1)—in other words, it is most clearly connected to the ‘know-as-state’ and ‘know’ as opposed to the ‘understand’ semantics of Dimension 1. The distribution of the forms is shown in figure 11.13. Based on this, a possible characterisation of the situation in the language would be that mithi is the original basic ‘know’-word which has bee joined by, and will potentially be replaced by, tharmai which originally had ‘realise’-semantics (as is supported by the etymology proposed) but is now extending.

The distribution of languages with markers correlated with Dimension 4 is shown in figure 11.14.

There does not appear to be any clear areal patterns in how or whether the ‘realise’ meaning is expressed or not.
Figure 11.13: Distribution of *mithi* and *tharmai* in Dimasa, Dimensions 1 and 4

Figure 11.14: Map of markers correlated with Dimension 4 - 'realise'
11.6 Not knowing

In section 7.7, we saw that two main groups can be distinguished among the languages which have lexical negation of knowledge—those where lexical negation is limited to one context and really represents the lexical negation of ability and those where lexical negation has a wider extension.

It is the latter group which are examples of lexical negation of knowledge proper. The typology of lexical negation of knowledge can be understood as varying along two scales. First, in how widely the lexically negated form is used, and second, to what extent the positive form for ‘know’ occurs with standard negation.

Moro (Heibanic, Sudan) represents the extreme on both scales. In Moro, as we saw in figure 7.12 above, the extended knowledge domain is entirely dominated by two forms—-liyet ‘know’ and -iâba ‘not know’, the latter of which is used in 14 out of in total 15 negative contexts in the sample. In one negative context in the verse sample do we find -liyet negated. Considering the entire NT, there are 21 occurrences of -ero -liyet-, the standard negation, compared with 208 occurrences of -iâba. It is at this point unclear if there is a difference in meaning between the two forms of expression, and if so, what it is. Thus Moro both extends the lexical negation form to basically the entire extended knowledge domain and lacks occurrences of a positive ‘know’ form with standard negation.

A more common pattern is that the lexically negated form is used mainly within the narrow knowledge domain, but does not extend to ‘not understand’. Two types of cases can be distinguished; those where there are distinct form for ‘know’ and ‘understand’ and only the former is substituted by the lexically negated form in negation and those where the same form covers both ‘know’ and ‘understand’ in the positive but the lexically negated form is found only for the former meaning.

As an example of the first kind, consider again the example from Naro discussed in section 9.3.2. In Naro, we find a form for ‘know’ q’âa- and a lexically negated form c’âuâa-. This latter form, however, does not occur when the complex form for ‘understand’ kóma q’âa (lit. ‘hear know’) is used—this form uses standard negation. The example from section 9.3.2 is repeated in

1106

Note that due to a lack of verses in the NT there are no examples ‘not find out’ in the verse sample. The explanation for this is likely that the primary context for the ‘find out’-meaning in the NT is a temporal adverbial clause and the negation of this is for pragmatic reasons less common. Further data is required to investigate this group’s interaction with lexical negation.
As an example of the other pattern—a distinction lacking in the positive but being made by the contrast between standard negation and lexical negation—consider North Tanna. The form covering the entire knowledge domain in the positive is -\textit{itun}, with the exception of some ‘find out’ contexts which are instead expressed with -\textit{ato} ‘hear’. In the negative, both standardly negated -\textit{itun} as well as the lexical negation form -\textit{ruru} is used. Considering the distribution of these forms along Dimensions 1—sensitive to the ‘know’–‘understand’ distinction—together with Dimension 5 reveals that the choice is sensitive to the ‘know’–‘understand’ distinction—see figure 11.15. Illustrative examples are given in 107. In 107a we see the use of the lexical negation form -\textit{ruru}, contrasted with standard negation of the positive ‘know’ form -\textit{itun} in 107b. In the former, what is being expressed is a lack of factual knowledge, whereas in 107b it is a lack of deeper understanding which is expressed; the problem is not the lack of information, but an inability to make the correct connections.
(107) North Tanna (Austronesian, Vanuatu)

a. Mark 13:35

\[
\text{məto \textit{inu \textit{itəmat nə-k-ot-əruru \ niAN \textit{iəvə \rəvə \textit{rəvə because 2PL 2-NPST-PL-not.know when lord POSS} nǐmə \textit{un \textit{otuwa \ lən house DEM FUT-3SG-come ?}}
\]

‘for you do not know when the master of the house is coming’

b. John 8:27

\[
\text{Məto \textit{ilat k-əs-ot-ɪtun-ən məmə \textit{lesu \ in but 3PL 3PL-NEG-PL-know-NEG COMP PR.NAME 3SG t-at-əghat-in Tata UhGIN 3SG-?-speak-TR father god}}
\]

‘But they did not understand that the speaking about God’ (AS)

In many of the languages in the sample, however, standard negation of the general ‘know’ form coexists with lexical negation without any apparent underlying semantic distinction. There are no languages in the sample where lexical negation clearly makes a distinction within the narrow knowledge domain (e.g. only being used for ‘know that’ and not ‘not know person’).

In figure [11.16] the distribution of lexical negation in the verse sample is shown. Languages where lexical negation is found are in blue, those where it is not are black. The size of the marker corresponds to the number of times the lexical negation occurs (in the case of more than one lexical negation form, the sum of these is considered). It can be noted that there appears to be a certain areality here, with Africa and in particular Oceania over-represented in the lexical negation category.

Of particular interest in map [11.16] is the language in the upper left corner of the map. This is Central Alaskan Yupik, where knowing is generally expressed by the negation of a verb \textit{nallu}-meaning ‘not know’. This is exemplified in (108).

(108) Central Alaskan Yupik (Eskimo-Aleut, USA)

a. (Miyaoka 2012 777)

\[
\text{nallu-nrit-aqa elpe-nun kenk-uci-qa not.know-NEG-IND.1SG>3SG 2SG-ALL love-NMLZ-ABS.2SG>1SG}
\]

‘I know that you love me’
This way of expression knowing—by the negation of ‘not knowing’—as well as the relatively high frequency of lexical negation of knowledge suggests that there is something uncommonly central about not knowing vis-à-vis knowing.

11.7 Knowing how

Among the languages where we find forms expressing ‘know how’, three main types can be distinguished:

1. ‘Know how’ is expressed with a form distinct both from ‘know that’ and
‘be able to’

2. ‘Know how’ is expressed by the same form as ‘be able to’

3. ‘Know how’ is expressed with a distinct construction (but the lexical item is the same as ‘know that’)

There is also a fourth, peripheral, type, in which ‘know how’ is not distinct from ‘know that’ but where this form also extends to ‘be able’.

Beginning with the first type, an example of this is Car Nicobarese where the form huveu ‘know how’ (109a) is distinct from both akahal ‘know that’ (109b) and kô(ôren) ‘be able’ (109c).

(109) Car Nicobarese (Austroasiatic>Nicobaric, India)
   a. Luke 12:56
      
      huveu rô yin veeny inlahen ngih tumlat,  
      know.how word/language 2PL tell manner DEM earth
      ngam haliöngö inrê  
      DEM sky also

      ‘You know how to interpret the earth and sky’
   b. Mark 12:14
      
      mahaköphöti; akaha-lôn in tô Meh tô-kôn  
      teacher; know-mind 1PL COMP 2SG ADJ-be.truthful

      ‘Teacher we know that you are truthful’
   c. Luke 14:6
      
      ngaich chaa nö öt kô nö sâpö u  
      PRT 3PL PRT NEG able PRT answer 3PL

      ‘And they were not able to make a reply to these things’

As an example of a language where ‘know how’ is distinguished from ‘know that’ with the same form as ‘be able’, consider Swedish. As discussed in section 9.7, the form kunna is a modal verb meaning ‘can’ but is also used in the meaning ‘to know how’. This is shown in examples (110) (repeated from section 9.7).

(110) Swedish (Indo-European>Germanic, Sweden)
a. Matt 26:61

\[ \text{Jag kan riva ner Gud-s tempel och bygga upp det igen inom tre dag-ar.} \]

‘I am able to destroy the temple of God and rebuild it within three days’

b. Luke 12:56(a)

\[ \text{Hycklare, jord-en-s och himl-en-s utseende kan ni tyda [—]} \]

‘Hypocrites, you know how to evaluate the appearance of the earth and sky’

c. Acts 21:37

\[ \text{Jaså, du kan grekiska} \]

‘Oh, so you know Greek?’

As a subtype of the second type, the extension of modal verbal morphology into ‘know how’ can be noted.

In the third type of languages, we find the same lexeme used for both ‘know that’ and ‘know how’ but we find a specialised construction for the latter, most typically a non-finite form of the complement verb juxtaposed with ‘know’, as in Basque. The examples from section 9.7 are repeated below in 111.

(111) Basque (Isolate, Spain)

a. Artiagoitia et al. (2003: 646)

\[ \text{Banekien Mikel berandu etorriko zena.} \]

‘I knew that Mikel would arrive late’
We can approximate a view of the distribution of these types cross-linguistically by considering the markers correlated with Dimension 6. This shown in map 11.17.

The most striking feature of map 11.17 is the almost complete lack of forms correlated with the dimension in South and Meso-America. I currently cannot say if this is an artefact of the data or if this reflects some actual typological pattern. Lexical items for ‘know how’ distinct from both ‘know that’ and ‘be able’ are scattered geographically with no particular pattern, though it may be noted that these kinds of forms are missing in Africa and Oceania. Habitual markers (cf. section 9.7) are concentrated to Papua New Guinea, though there are only two examples.

It should be remembered that Dimension 6 is based almost entirely on only three verses—the results shown in map 11.17 are therefore necessarily quite...
tentative. In general, it is clear that there is much left to investigate regarding 'know how' meanings and constructions, and the three types suggested here are best understood as a possible framework for these further investigations.

11.8 Summary and conclusion

In the preceding section, the cross-linguistic expression of the dimensions identified in chapter 7 and further semantically characterised in chapter 9 was described.

The distinction between know-as-state and know-as-achievement is frequently marked by using the same verb in different constructions. There does not appear to be any universal directionality of the formal relationship between the two senses—state may be marked in relation to achievement or the other way around. Additionally, both senses may be unmarked or both marked. However, a direction from state to achievement is most common. There are no clear areal patterns regarding this marking, suggesting that we are dealing with an unstable phenomenon.

When it comes to the expression of 'understand', I argued that the traditional view of regarding this as a question simply of colexification or not is inadequate. The extension of forms for 'understand' is highly variable and intertwined with the extension of 'know' forms. In general, it can be observed that languages of Eurasia and North America most completely distinguish the 'understand' contexts, while many languages of Oceania and South America treat 'understand' as a part of 'know'. Africa falls somewhere in-between, with many languages having forms with only a handful of uses. The considerably variability of forms furthermore supports the point that characterising these form simply as translation equivalents of 'understand' is inadequate.

The expression in the languages of the sample of the 'recognise' and 'know person' meanings of Dimension 3 is also quite variable, but can be explained by a cline from 'recognise' contexts, through 'know person (directly)' to 'know person (indirectly)'. The variation in the extension of 'know person' forms follows this scale. It was noted that while the distinction of 'know person' is the norm in Eurasia, it is also found across the other macro-areas. Further, the role of non-person nominal objects in this distinction was considered, and I argued that the data shows that cross-linguistically, these are secondary to person objects.

The expression of 'realise' shows no clear areal patterns—forms are found across the macro-areas. Perhaps there is a slight under-representation in Oceania and North America (though saying anything certain about the latter is difficult due to the small number of languages). It was also seen that for some languages, the forms which correlate with Dimension 4 have a considerably
wider distribution than to the small group of ‘realise’ contexts mainly sepa-
rated out in the dimension. I suggested the possibility that this is the result of
some semantic parameter of ‘realise’ in fact being more widely relevant to the
verse sample and the case of the two ‘know’ forms of Dimasa was discussed.

Lexical negation of knowledge is geographically concentrated to Oceania
and Africa, with some scattered occurrences in the Americas. It is not found in
any of the Eurasian languages of the sample. Two main types of lexical nega-
tion of knowledge were distinguished—the extension of a negative modal form
to one or two contexts in the knowledge domain and lexical negation of knowl-
edge proper. Among the latter, we can distinguish between those languages
where the lexical negation is restricted to the narrow knowledge domain and
those where it extends across the entire extended knowledge domain. There
is variability also within the former group, but further research is required to
pinpoint the semantic and pragmatic factors determining whether lexical or
standard negation is used.

Finally, the expression of ‘know how’ was considered. Three main types
were distinguished, and their distribution in the language sample was con-
sidered based on the properties of the markers correlated with Dimension 6.
Apart from a somewhat puzzling dearth of forms in South and Meso-America,
no clear areal patterns emerge.
12. Conclusions

This thesis centres around two research questions, repeated below:

A. What is the cross-linguistic status of knowledge predication as a distinct construction?

A1. Is knowledge predication generally distinguished in languages, and if so, what are the boundaries of it?

A2. How universal is a distinct construction for knowledge predication?

B. What linguistic distinctions are made in the knowledge domain—what meanings are distinguished?

B1. How do these meanings relate to one another—are some more basic or central?

B2. How can the distinctions be characterised semantically?

B3. How do the forms for these meanings relate to forms for domain-external meanings?

B4. How are these distinctions expressed in the languages of the world?

With regards to the first question, I have shown that ‘know’ can be distinguished as a **generalised domain**—a domain which is an abstraction from regularities in the categorisations of individual language varieties (see section 3.2). In fact, two overlapping domains are distinguishable—the **extended knowledge domain** containing the meanings ‘know’, ‘understand’ and ‘find out’ and the **narrow knowledge domain**, consisting of ‘know’ (‘know that’, ‘know person’, ‘know an object’ ‘recognise’ and ‘know how’)—see chapter 5. In other words, English is not—as was raised as a possibility in the introduction—an outlier cross-linguistically in subsuming the latter under one form; ‘know’ is cross-linguistically generally not limited to ‘know that’. Somewhat less commonly, but still quite frequently, languages subsume the meanings of the extended knowledge domain under one and the same ‘know’ form. In answer to research question A1 then, we can conclude that knowledge predication is generally distinguished cross-linguistically, and that there are two
main ways of delimiting it, either coinciding with English *know* or broader, including also ‘understand’ and ‘find out’.

In the 83-language variety sample—arrived at through a novel development of the Diversity Value method—there are only two languages which can be said to lack a word corresponding to ‘know’, both from New Guinea—Kalam (Nuclear Trans New Guinea) and Fasu (Isolate). In both languages, we find a form which includes, but is not limited to, ‘know’ meanings. There are some additional languages where the line between ‘know’ and other domains is “blurred” in various ways. There appears to be a geographical concentration of these languages to New Guinea, suggesting a connection to local cultural or communicative practices.

Whatever analysis of Kalam, Fasu and similar languages is settled on, it is nevertheless an important result that *almost* all languages in the sample have a form for the concept ‘know’, a finding which has implications for the philosophical analysis of knowledge. The answer to research question A2 is thus that a distinct construction for knowledge predication is at the least a near universal.

Within the extended knowledge domain, several meaning distinctions are recurrently made in the languages of the sample. In chapters 7-11 these are investigated using various methods, building throughout on the Classical Multi-Dimensional Scaling analysis presented in chapter 7. Following the frame-based approach to meaning (cf. section 3.1), these can be thought of as giving a preliminary inventory of the frames making up the knowledge domain—cross-linguistically salient idealisations of experience. The frames identified are summarised in table 12.1; this gives an answer to research question B of what distinctions are made in the knowledge domain.

Most widespread in the sample are two interacting distinctions—between *know-as-state* and *know-as-achievement* and between ‘understand’ and ‘know’. The distinctions are generally made at different levels of linguistic expression—lexical and constructional—and the following generalisation can be formulated:

**Generalisation I. The lexical-constructional divide:** If a language makes a lexical distinction in the extended knowledge domain, it will generally be between ‘know’ and ‘understand’ but if a language makes a constructional distinction, it will generally be between know-as-state and know-as-achievement.

Generalisation I can perhaps be explained by economical reasons—whereas the difference between know-as-state and know-as-achievement can be understood in terms of general aspectual distinctions expressed by the same constructions, ‘understand’ and ‘know’ are unique to the knowledge domain.
<table>
<thead>
<tr>
<th>Frame &amp; Dim.</th>
<th>Description &amp; Sub-frames</th>
<th>Discussed in</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘know that’</td>
<td>The core of knowledge predication—prototypically knowing a clause</td>
<td>2.1, 8</td>
</tr>
<tr>
<td>know-as-state</td>
<td>One of the two main aspectual construals of knowing—as an ongoing relation</td>
<td>7.4, 9.3.2, 11.2</td>
</tr>
<tr>
<td>know-as-achievement</td>
<td>One of the two main aspectual construals of knowing—as coming to know.</td>
<td>7.4, 9.3.2, 11.2</td>
</tr>
<tr>
<td>‘understand’</td>
<td>Closely related to ‘know’, but distinguished by a cluster of properties including depth, complex objects, internal processing etc.</td>
<td>7.4, 9.3.3, 10.3.1, 10.3.3, 11.3, 9.3.3</td>
</tr>
<tr>
<td>• ‘understand (a) language’—a bridging context to ‘hear’</td>
<td></td>
<td></td>
</tr>
<tr>
<td>‘find out’</td>
<td>Know-as-achievement through hearing</td>
<td>7.4, 9.3.4, 9.3.4</td>
</tr>
<tr>
<td>• ‘find out by hearsay’—specifying the source of knowledge as vague and second-hand</td>
<td></td>
<td></td>
</tr>
<tr>
<td>‘know person’</td>
<td>Knowledge predication with a person (or facts about a person) as the object of knowledge</td>
<td>7.5, 9.4, 10.3.2, 9.4, 11.4, 9.4, 11.4</td>
</tr>
<tr>
<td>• ‘recognise’—identify a person (mainly through sight)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• ‘know person (directly)’—know someone through direct contact, perceptual or social</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• ‘know person (indirectly)’—know who someone is or know facts about them—does not require personal contact</td>
<td></td>
<td></td>
</tr>
<tr>
<td>‘realise’</td>
<td>A kind of know-as-achievement characterised by internal processing</td>
<td>7.6, 9.5, 11.5</td>
</tr>
<tr>
<td>‘not know’</td>
<td>The predication of lack of knowledge</td>
<td>7.7, 9.6, 11.6, 11.6, 11.6, 11.6, 11.6</td>
</tr>
<tr>
<td>• ‘not know how’</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• ‘not understand’</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• ‘not recognise’</td>
<td></td>
<td></td>
</tr>
<tr>
<td>‘know how’</td>
<td>Having the knowledge to do something</td>
<td>7.8, 9.7, 11.7</td>
</tr>
</tbody>
</table>

**Table 12.1:** Summary of the identified cross-linguistic frames of knowledge predication
This finding is interesting considering the argument that understanding is distinct from and irreducible to knowledge and holds a special value. The frequency with which a meaning like ‘understand’ is distinguished in languages is interesting from this point of view. On the other hand, ‘knowing’ and ‘understanding’ are not on equal terms linguistically in that ‘understand’ is frequently expressed in terms of ‘know’ whereas relationships in the other direction are not found. This might be taken as evidence that ‘understand’ is reducible to ‘know’, or at least dependent on it.

It is also interesting that there is considerable overlap in the characterisation derived at based on the cross-linguistic forms given in chapter 9—where also the other dimensions are characterised semantically, answering research question B2—and the explications of ‘understand’ forms discussed in section 2.2.5. This hints that ‘understand’ in all its complexity is a cross-linguistically relevant concept.

In addition to the frequent distinction between ‘know’ and ‘understand’, many languages distinguish externally caused knowledge (‘find out’) which is closely connected to hearing. Fewer languages make the distinctions between ‘know that’ and ‘know person’ or ‘recognise’ and yet fewer distinguish ‘realise’ and ‘know how’. Some language also single out not knowing by means of lexical negation of knowledge predication—distinct lexical stems meaning ‘not know’.

There are systematic differences in the linguistic expression of the different parts of the knowledge domain. These differences can be explained by the knowledge domain having a prototypical core—‘know that’—with the other meanings in the narrow knowledge domain more peripheral and the meanings of the extended knowledge domain being yet more peripheral. The following generalisation can be proposed, answering research question B1 concerning the internal relations of the knowledge domain:

Generalisation II. **The core–periphery structure**: If forms at the core and the periphery of the knowledge domain are partially colexified, the latter may be motivated by the former, but not the other way around.

The parts of the knowledge domain also differ in their connection to other domains; there are clear patterns in what parts of the knowledge domain are connected to which concepts, answering research question B3 concerning domain-external connections. Of particular interest is the observation that ‘see’ is connected to ‘recognise’ and ‘know person’ while ‘hear’ is more strongly connected to ‘find out’, ‘understand’ and ‘know that’. The following generalisation can be made:
Generalisation III. **Personal knowledge and sight:** If in a language a distinction is made between ‘know that’ and ‘know person/recognise’ and there are synchronically evident connections to other domains, ‘know person/recognise’ rather than ‘know that’ will be connected to sight.

Generalisation III offers insights regarding the debate concerning the linguistic relationship of knowledge to the sense modalities; the question asked is (partly) wrong—we shouldn’t be asking to what other domains ‘know’ per se is connected, but what senses of ‘know’ are connected with what domains. The synchronic data can be given a diachronic interpretation: the path from ‘see’ to ‘know’ primarily goes through ‘know person’, while the path from ‘hear’ to ‘know’ primarily goes through ‘find out’ or ‘understand’.

Taking a token-based approach, the thesis makes several observations about the expression of these meanings in the languages of the sample, answering research question B4 of how the distinctions in the knowledge domain are expressed in the languages of the world.

A quantification of the directionality of the formal relationships between know-as-state and know-as-achievement is given. The considerable variability in which meaning is formally unmarked might be taken as support of the view that ‘know’ is neither state nor achievement (cf. section 2.2.7), but has an aspectual structure which can be construed as both; since no meaning is basic, the variation in which is taken as the basis for the other is expected.

The encoding of ‘know person’ and ‘recognise’ shows a number of interesting patterns. I argued that the extension of forms regarding these meanings can be understood via the following tendency scale (see section 11.4):

**The personal knowledge scale**

Recognise (a person) < Know a person (directly) < Know a person (indirectly)

In reference to this scale, the following generalisation can be formulated:

**Generalisation IV. Extension of ‘know person’ forms:** If a form distinct from ‘know that’ is used for a meaning on the personal knowledge scale, it will also be used for meanings to the left of it on the scale, barring another form distinct from ‘know that’.

The role of non-person nominal objects in this distinction appears rather marginal cross-linguistically and the following generalisation can be formulated:

**Generalisation V. Non-person nominals as objects of knowledge:** If in a language a ‘know’ form distinct from ‘know that’ is used
for a non-person nominal, then it will also be used for ‘know person’

That the distinction cross-linguistically is thus primarily between knowing facts and knowing people—with non-person nominal objects being intermediate—suggests that a cross-linguistic characterisation of this distinction should start with ‘knowing people’.

Finally, the following generalisation regarding the expression of ‘know how’ can be stated:

Generalisation VI. **The knowledge–ability scale**: If there is a form in a languages which expresses both ‘know that’ and ‘be able’, this form will also express ‘know how’.

Generalisation VI is interesting given the suggestion of the relationship of ‘know that’, ‘know how’ and ‘be able’ as a continuum in section 2.2.1—I suggested that these meanings are perhaps best thought of as a collection of interrelated and overlapping continua and that languages may vary in how these continua are carved up linguistically.

Territory of Information, intersubjective concerns and epistemic authority have not been discussed in this thesis, though they were raised as potentially relevant aspects in chapter 2. The reason is that it was not present among the first six dimensions selected for consideration. There are, however, effects of these kinds of concerns in the data. For example, there appears to be an unwillingness in some translations to render ‘you know’—from the perspectives introduced in section 2.2.6 this is not surprising.

The frames, connections and generalisations summarised in this chapter may be visualised in the form of a network semantic map. In chapter 7 I introduced the distinction between network semantic maps and proximity maps. I argued that there are analytical problems faced in trying to construct a network map of an unknown domain and that proximity maps offer solutions to these problems. However, the very same problems that motivate the use of proximity maps make network maps particularly useful in discovering what remains unclear, as well as in clearly formulating testable hypotheses about what is thought to be known. In figure 12.1 a network semantic map of the knowledge domain and connected concepts is shown.

The nodes are equivalent to the inventory of frames in table 12.1. As for the edges, two kinds are distinguished in map 12.1 represented by arrows and lines. Lines represent essentially colexification. Arrows, on the other hand, represent hypothesized diachronic relationship. The map may serve as a roadmap for future research by providing clear (and falsifiable) predictions.
Figure 12.1: A semantic map of the knowledge domain

Lying beyond the focus of this thesis but central to the understanding of knowledge predication and thus future research in the area are two issues. The first is the possibility of variation in the boundaries of knowledge—what makes *knowing* something different from just *believing* it or *thinking* it across the cultures and languages of the world? This question requires insights from not only linguistics but philosophy, anthropology and other human sciences. The second is the function of knowledge predication in interaction. Why, how and when do people predicate knowledge cross-linguistically? And—perhaps more interestingly—why, how and when do people predicate different kinds of knowledge? When do we *understand* things rather than just *know* them, when do we *know people* and when do we *not know* things? I have demonstrated in this thesis that knowledge predication is not monolithic but shows considerable internal variation and these insights will be crucial for fully understanding how people use knowledge predications to construct their world.
## A. Appendix: The languages in the study

Format of each entry is as follows:

- **Language name (Family, Country, ISO-639)**
- **Bible version (Vernacular title). Year**
- **Descriptive sources**

Languages in the extended sample are marked with *

In addition to the works cited below, I have when available also made use of internet resources such as Google Translate and Wiktionary.

### North America

A. **Bribri (Chibchan, Costa Rica, bzd)**


- Jara Murillo 2018
- Krohn 2023

B. **Comaltepec Chinantec (Otomanguean > Western Otomanguean, Mexico, cco)**


- Anderson 2018
- Pace 1990

C. **Cherokee (Iroquoian, USA, chr)**

The New Testament in Cherokee. 1860

- Nuttle et al. 2014-2023
- Cherokee Nation 2022
- Wikipedia contributors 2023
- Montgomery-Anderson 2008
- Feeling 2003
- Holmes & Sharp Smith 1977

D. **Chol (Mayan, Mexico, ctu)**


- Vázquez Álvarez 2011
- Scott & Warkentin 1980
- Aulie, Aulie & Scharfe de Stairs 2009

E. **Nopala Chatino (Otomanguean > Eastern Otomanguean, Mexico, cya)**


- Pride & Pride 2010
- McIntosh 2015

F. **Dogrib (Athabaskan-Eyak-Tlingit, Canada, dgr)**


- Saxon & Siemens 1996
- Welch 2015

G. **Central Alaskan Yupik (Eskimo-Aleut, USA, esu)**

Yupik Bible. 2012

- Miyaoka 2012

H. **Hopi (Uto-Aztecan, USA, hop)**


- Albert & Shaul 1985

---

291
I. Tol (Jicaquean, Honduras, jic)
Holt 1999 Dennis & Dennis 2001

J. Mi’kmaq (Algic, Canada, mic)
Haberlin et al. n.d. Fidelholtz 1968

K. Tlahuitoltepec Mixe (Mize-Zoque, Mexico, mxp)
The New Testament in Mixe de Tlahuitoltepec (Ja Øgyajpxy Ja Ømyadya aky Midi Xukpaa j timp — Ja Nitsookaj tin). 1987
Romero-Méndez 2009 Martínez 2015

L. Pennsylvania German* (Indo-European > Germanic, USA, pdc)
Burridge 2006 Miller 2014

M. Western Highland Purepecha (Tarascan, Mexico, pua)
Hernández Domínguez 2015 Lathrop 2007

N. Tlachichilco Tepehua (Totonacan, Mexico, tpt)
The New Testament in Tepehua de Tlachichilco. 2004
Watters 1988

South America

A. Mapudungun (Araucanian, Chile, arn)
Smeets 2008

B. Central Aymara (Aymaran, Bolivia, ayr)
The Bible in Aymara. 2011

C. Galibi Carib (Cariban, Suriname, car)
Courtz 2008

D. Cashibo-Cacataibo (Pano-Tacanan, Peru, cbr)
Zariquey Biondi 2011

E. Cuiba (Guahiboan, Colombia, cui)
Kerr & Berg 2018

F. San Blas Kuna (Chibchan, Panama, cuk)

G. Culina (Arawan, Brazil, cul)
The New Testament in Culina (Culina Mâdija). ND
Dienst 2006

H. Kotiria (Tucanoan, Colombia, gvc)
Stenzel 2004 Waltz 2007
I. Murui Huitoto (Huitotoan, Colombia, huu)

[Woitylak 2017, Burtch 1983]

J. San Andres Creole English* (English lexifier creole, Colombia, icr)
The New Testament in Islander Creole. 2015

[Bartens 2013]

K. Kaiwá (Tupian, Paraguay, kgk)

[Loman de Barros 2014]

L. Southern Nambikuára (Nambiquaran, Brazil, nab)

[Kroeker 2001, Kroeker 1996]

M. Pilagá (Guaiacuruan, Argentina, plg)

[Vidal 2001]

N. Huallaga Huánuco Quechua (Quechuan, Peru, qub)
The Bible in Quechua del Huallaga (Huánuco) Pillcu Quechua (Tayta Diosninchi Isquirbichishan). 2010

[Weber 1989]

O. Rikbaktsa (Nuclear-Marco-Je, Brazil, rkb)


P. Epena (Chocoan, Colombia, sja)


Q. Trinitario-Javeriano-Loretano (Arawakan, Bolivia, trn)

[Rose 2014]

R. Yuracaré (Isolate, Bolivia, yuz)
The Bible in Yuracare. 2000

[van Gijn 2006]

S. Sanumá (Yanomamic, Brazil, xsu)

[Borgman 1990, Perri Ferreira 2017]

Eurasia

A. Standard Arabic (Afro-Asiatic > Semitic, Saudi-Arabia, arb)
The Holy Bible, New Arabic Version (Ketab El Hayat). 1982

[Wehr & Cowan 2012, Schulz, Krall & Reuschel 2000]

B. Bengali (Indo-European > Indo-Aryan, Bangladesh, ben)
The Bible in Bengali Common Language Version. 2000

[David 2015, Ali, Moniruzzaman & Tareque 2007]

C. Car Nicobarese (Austroasiatic > Nicobaric, India, caq)
The Car Nicobarese Bible. 1927/2014


D. Chechen (Nakh-Daghestanian, Russian Federation, che)
The Bible in Chechen. 2012

E. Mandarin Chinese (Sino-Tibetan > Sinitic, China, cmn)
   Scripture Text of Chinese Union Version with New Punctuation. 1996
   Huang & Shi 2016

F. Dimasa (Sino-Tibetan > Brahmaputran, India, dis)
   Brahma 2013, Jacquesson 2008

G. Middle English* (Indo-European > Germanic, United Kingdom, enm)
   Wycliffe’s Bible. 1388
   Lewis & McSparran 2000-2018

H. Basque (Isolate, Spain, eus)
   The Bible in Basque (Elizen Arteko Biblia). 1983
   Hualde & Ortiz de Urbina 2003

I. Hungarian (Uralic, Hungary, hun)
   Rounds 2009

J. Iu Mien (Hmong-Mien, Vietnam, ium)
   The Bible in Iu Mien [new Latin script] (). 2007
   Arisawa 2016

K. Japanese (Japonic, Japan, jpn)
   New Interconfessional Translation Bible (Seisho Shinkyoudoyaku). 1987
   Hasegawa 2015

L. Kannada (Dravidian, India, kan)
   The Holy Bible, Easy Reading Version, in the Kannada language of India. 1992
   Zydenbos 2020, Sridhar 1979

M. Halh Mongolian (Mongolic, Mongolia, khk)
   New Testament Portions in Halh Mongolian. 2004
   Kullmann & Tserenpi 2001

N. Korean (Koreanic, Korea, kor)
   The Bible in Korean: Revised Version. 1995
   Martin 1992

O. Swedish (Indo-European > Germanic, Sweden, swe)
   (Personal knowledge.)

P. Thai (Tai-Kadai, Thailand, tha)
   The Bible in Thai, New Contemporary Version. 2000
   Jenny 2019

Q. Turkish (Turkic, Turkey, tur)
   The Bible in Turkish (Kutsal Kitap Yeni Ceviri). 2009
   Ketrez 2012

Africa

A. Afrikaans* (Indo-European > Germanic, South Africa, afr)
   Donaldson 1993

B. Hamer-Banna (South Omotic, Ethiopia, amf)
   Petrollino 2016

C. Zarma (Songhay, Niger, dje)
   The New Testament in Zarma. 2004
   Bernard & White-Kaba 1994, Seydou Hanafiou 1995
D. Toro So Dogon (Dogon, Mali, dts)
The New Testament in Dogon. 1995
Heath 2014

E. Igbo (Atlantic-Congo > Volta-
Congo > Igboid, Nigeria, ibo)
The New Testament in Igbo. ND
Williamson & Blench 2006 Eme-
nanjo 2015

F. Southern Kisi (Atlantic-Congo >
Mel, Sierra Leone, kss)
Kisi Bible. 2012
Childs 1995

G. Matal (Afro-Asiatic > Chadic,
Cameroon, mfh)
Verdizade 2018

H. Ma’di (Central Sudanic, Uganda,
mihi)
The Bible in Ma’di. 1977
Blackings & Fabb 2003

I. Moro (Heibanic, Sudan, mor)
New Testament in Moro. 1965
Black & Black 1971 Blench 2006
Rose 2013

J. Murle (Surmic, South Sudan, mur)
The New Testament in Murle. 1996
Arensen 1982 Miller 1986
Lokonobei & De Jong 1989

K. Naro (Khoe-Kwadi, Botswana, nhr)
The New Testament in the Naro
language of Botswana (Kabas Qae-
xg’ae sa). 2012
Visser 2013 Visser 2001

L. Tamasheq (Afro-Asiatic > Berber,
Mali, taq)
Heath 2005

M. Teso (Nilotic, Uganda, teo)
The Bible in Teso (Ebaibuli). 1961
Barasa 2017 Dimmendaal 1983

N. Uduk (Koman, Sudan/Ethiopia,
udu)
Killian 2015

O. Wolaytta (Ta-Ne- Omotic, Ethiopia,
wal)
The Bible in Wolaitta. 1997
Wakasa 2008 Lamberti & Sottile
1997

P. Yalunka (Mande, Sierra Leone, yal)
Lüpke 2005 Creissels 2010

Oceania

A. Angor (Senagi, Papua New Guinea,
agg)
The New Testament in Angor
(GodindiHoafi). 2001
Litteral 1980

B. Bauzi (Geelvink Bay, Indonesia,
bvz)
Briley 1997 Clouse 1997

C. Daga (Dagan, Papua New Guinea,
dgz)
The New Testament in Daga. 1974
Murane 1974 Murane & Murane
2007

D. Fasu (Isolate, Papua New Guinea,
faa)
The New Testament in Fasu (Ko-
timo Kawe Someaka Masane Oy-
apo). 2011
Loeweke & May 1980 May &
Loeweke 1981 Loeweke & May
1966
E. Kire (Lower Sepik-Ramu, Papua New Guinea, geb)
Stanhope 1972
F. Kalam (Nuclear Trans New Guinea, Papua New Guinea, kmh)
The New Testament in Kalam (Kalam Baybol Buk Gor minim nuk kisin angayak) 1992
Pawley & Bulmer 2011, Pawley 1966
G. Doromu-Koki (Manubaran, Papua New Guinea, kqc)
The Bible in Doromu-Koki (Sei di Uka Ago Ruaka di Buka Bedakai). 2011
Bradshaw 2012
H. Kuot (Isolate, Papua New Guinea, kto)
Lindström 2002
I. Meyah (East Bird’s Head, Indonesia, mej)
The New Testament in Meyah. ND
Gravelle 2011
J. Wubuy (Gunwinyguan, Australia, nuy)
Heath 1984, Horrack 2018, Heath 1982
K. Olo (Nuclear Torricelli, Papua New Guinea, ong)
Staley 2007, McGregor & McGregor 1982
L. Rotokas (North Bougainville, Papua New Guinea, roo)
Robinson 2011, Firchow & Firchow 2008
M. Kriol (English lexifier creole, Australia, rop)
Malcolm n.d.
N. Tabaru (North Halmahera, Indonesia, tby)
The Bible in Tabaru. 2008
Kotynski 1988, Holton 2008, Angin 2018
O. North Tanna (Austronesian > Malayo-Polynesian, Vanuatu, tnn)
Sverredal 2018, Nehrbass 2012
P. Warlpiri (Pama-Nyungan, Australia, wbp)
Warlpiri Bible
Simpson 1983, Swartz 2012
Q. Yele (Isolate, Papua New Guinea, yle)
B. Appendix: The verse sample of the study

<table>
<thead>
<tr>
<th>Category</th>
<th># of verses</th>
<th>Category</th>
<th># of verses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Know that</td>
<td>7</td>
<td>Hear (clear source)</td>
<td>3</td>
</tr>
<tr>
<td>Not know</td>
<td>5</td>
<td>Hear (unclear source)</td>
<td>4</td>
</tr>
<tr>
<td>Know wh-</td>
<td>3</td>
<td>Realise</td>
<td>3</td>
</tr>
<tr>
<td>Know person</td>
<td>7</td>
<td>Understand (aspectual)</td>
<td>4</td>
</tr>
<tr>
<td>Know other nominal</td>
<td>6</td>
<td>See</td>
<td>5</td>
</tr>
<tr>
<td>Remember</td>
<td>8</td>
<td>See (extension)</td>
<td>3</td>
</tr>
<tr>
<td>Understand (depth)</td>
<td>5</td>
<td>Hear (perception)</td>
<td>4</td>
</tr>
<tr>
<td>Know how</td>
<td>5</td>
<td>Obey</td>
<td>4</td>
</tr>
<tr>
<td>Ability</td>
<td>8</td>
<td>Believe</td>
<td>5</td>
</tr>
<tr>
<td>External ability</td>
<td>3</td>
<td>Think</td>
<td>3</td>
</tr>
<tr>
<td>Recognise</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Total: 100</td>
<td></td>
</tr>
</tbody>
</table>

**Table B.1:** Categorisation of the verses in the sample.

Know that

1. Mark 12:14
   And when they came, they said to him, “Teacher, we know that you are truthful and you do not care what anyone thinks, because you do not regard the opinion of people but teach the way of God in truth. Is it permitted to pay taxes to Caesar or not? Should we pay or should we not pay?”

2. Matt 26:2
   “You know that after two days the Passover takes place, and the Son of Man will be handed over in order to be crucified.”

3. Mark 13:28
   “Now learn the parable from the fig tree: Whenever its branch has already become tender and puts forth its leaves, you know that summer is near.

4. Luke 12:30
   For all the nations of the world seek after these things, and your Father knows that you need these things.
5. John 19:10
   So Pilate said to him, “Will you not speak to me? Do you not know that I have authority to release you, and I have authority to crucify you?”

6. Matt 25:26
   But his master answered and said to him, ‘Evil and lazy slave! You knew that I reap where I did not sow and gather from where I did not scatter seed.

7. Luke 8:53
   And they began laughing at him, because they knew that she was dead.

8. 1 Cor 1:16
   (Now I also baptized the household of Stephanus. Beyond that I do not know if I baptized anyone else.)

9. Mark 13:35
   Therefore be on the alert, for you do not know when the master of the house is coming—whether in the evening, or at midnight, or when the rooster crows, or early in the morning—

10. John 8:27
   (They did not know that he was speaking to them about the Father.)

11. Mark 9:32
   But they did not understand the statement, and they were afraid to ask him.

   But Jesus said, “Father, forgive them, for they do not know what they are doing.”
   And they cast lots to divide his clothes.

13. Mark 1:24
   saying, “Leave us alone, Jesus the Nazarene! Have you come to destroy us? I know who you are—the Holy One of God!”

14. John 15:15
   No longer do I call you slaves, because the slave does not know what his master is doing. But I have called you friends, because everything that I have heard from my Father I have revealed to you.

15. John 6:6
   (Now he said this to test him, because he knew what he was going to do.)

16. Mark 1:34
   And he healed many who were sick with various diseases and expelled many demons. And he did not permit the demons to speak, because they knew him.
17. John 1:48  
Nathanael said to him, “From where do you know me?” Jesus answered and said to him, “Before Philip called you, when you were under the fig tree, I saw you.”

18. Mark 14:71  
And he began to curse and to swear with an oath, “I do not know this man whom you are talking about!”

19. John 7:29  
I know him, because I am from him and he sent me.”

20. Acts 19:15 (a)  
But the evil spirit answered and said to them, “Jesus I know, and Paul I am acquainted with, but who are you?”

21. Acts 19:15 (b)  
But the evil spirit answered and said to them, “Jesus I know, and Paul I am acquainted with, but who are you?”

22. Acts 7:18  
until another king arose over Egypt who did not know Joseph.

Know other nominal

23. John 14:4  
And you know the way where I am going.”

24. John 10:4  
Whenever he sends out all his own, he goes before them, and the sheep follow him because they know his voice.

25. John 4:32  
But he said to them, “I have food to eat that you do not know about.”

26. Mark 10:19  
You know the commandments: ‘Do not murder, do not commit adultery, do not steal, do not give false testimony, do not defraud, honor your father and mother.’”

27. Mark 12:24  
Jesus said to them, “Are you not deceived because of this, because you do not know the scriptures or the power of God?

But he knew their thoughts and said to the man who had the withered hand, “Get up and stand in the middle,” and he got up and stood there.

Remember

29. Matt 26:75  
And Peter remembered the statement Jesus had said, “Before the rooster crows, you will deny me three times,” and he went outside and wept bitterly.
30. Mark 11:21
   And Peter remembered and said to him, “Rabbi, look! The fig tree that you cursed has withered!”

   Remember Lot’s wife!

32. Luke 23:42
   And he said, “Jesus, remember me when you come into your kingdom!”

33. Heb 10:32
   But remember the former days in which, after you were enlightened, you endured a great struggle with sufferings,

34. John 16:21
   A woman, when she gives birth, experiences pain because her hour has come. But when her child is born, she no longer remembers the affliction, on account of the joy that a human being has been born into the world.

35. Luke 16:25
   But Abraham said, ‘Child, remember that you received your good things during your life, and Lazarus likewise bad things. But now he is comforted here, but you are suffering pain.

36. John 2:22
   So when he was raised from the dead, his disciples remembered that he had said this, and they believed the scripture and the saying that Jesus had spoken.

Understand (depth)

37. John 13:17
   If you understand these things, you are blessed if you do them.

38. 2 Cor 1:13
   For we are not writing anything else to you except what you can read or also understand. But I hope that you will understand completely,

39. 2 Cor 1:14
   just as you have also understood us in part, that we are your reason for boasting, just as you are also ours in the day of our Lord Jesus.

40. Acts 8:30
   So Philip ran up to it and heard him reading aloud Isaiah the prophet and said, “So then, do you understand what you are reading?”

41. Mark 14:68
   But he denied it, saying, “I neither know nor understand what you mean!” And he went out into the gateway, and a rooster crowed.

Know how
42. Rom 8:26
And likewise also, the Spirit helps us in our weakness, for we do not know how to pray as one ought, but the Spirit himself intercedes for us with unexpressed groanings.

43. Acts 21:37
And as he was about to be brought into the barracks, Paul said to the military tribune, “Is it permitted for me to say something to you?” And he said, “Do you know Greek?

44. 1 Cor 14:16
For otherwise, if you praise in your spirit, how will the one who fills the place of the outsider say the “amen” at your thanksgiving, because he does not know what you are saying?

45. Luke 12:56 (a)
Hypocrites! You know how to evaluate the appearance of the earth and the sky, but how is it you do not know how to evaluate this present time?

46. Luke 12:56 (b)
Hypocrites! You know how to evaluate the appearance of the earth and the sky, but how is it you do not know how to evaluate this present time?

Ability

47. Luke 14:6
And they were not able to make a reply to these things.

48. Mark 1:40
And a leper came to him, entreating him and kneeling down and saying to him, “If you are willing, you are able to make me clean.”

49. Matt 17:16
And I brought him to your disciples, and they were not able to heal him.”

50. Matt 26:61
and said, “This man said, ‘I am able to destroy the temple of God and rebuild it within three days.’”

51. Matt 26:40
And he came to the disciples and found them sleeping, and he said to Peter, “So, were you not able to stay awake with me one hour?

52. Luke 1:22(a)
And when he came out he was not able to speak to them, and they realized that he had seen a vision in the temple. And he kept making signs to them, and remained unable to speak.

And behold, a woman was there who had a spirit that had disabled her for eighteen years, and she was bent over and not able to straighten herself up completely.
54. Matt 7:18

A good tree is not able to produce bad fruit, nor a bad tree to produce good fruit.

External ability

55. Luke 8:19

Now his mother and brothers came to him, and they were not able to meet with him because of the crowd.

56. Mark 2:4

And when they were not able to bring him to him because of the crowd, they removed the roof where he was. And after digging through, they lowered the stretcher on which the paralytic was lying.

57. Mark 6:5

And he was not able to do any miracle in that place except to lay his hands on a few sick people and heal them.

Recognise

58. Mark 6:54

And as they were getting out of the boat, people recognized him immediately.


And their eyes were opened, and they recognized him, and he became invisible to them.

60. Matt 17:12

But I say to you that Elijah has already come, and they did not recognize him, but did with him whatever they wanted. In the same way also the Son of Man is going to suffer at their hands."

61. Acts 12:14

And recognizing Peter’s voice, because of her joy she did not open the gate, but ran in and announced that Peter was standing at the gate.

62. Matt 7:20

As a result, you will recognize them by their fruits.


And when he found out that he was from the jurisdiction of Herod, he sent him over to Herod, who was also in Jerusalem in those days.

64. Acts 23:34

So after reading the letter and asking what province he was from, and learning that he was from Cilicia,

65. Mark 15:45

And when he learned of it from the centurion, he granted the corpse to Joseph.
66. Acts 9:30  
And when the *brothers* found out, they brought him down to Caesarea and sent him away to Tarsus.

But when the *crowds* found out, they followed him, and welcoming them, he began to speak to them about the kingdom of God, and he cured those who had need of healing.

68. John 12:9  
Now the *large crowd of Jews* found out that he was there, and they came, not only because of Jesus, but so that they could see Lazarus also, whom he raised from the dead.

69. Luke 7:37  
And behold, a woman in the town who was a sinner, when she learned that he was dining in the Pharisee’s house, brought an alabaster flask of perfumed oil,

70. Acts 19:34  
But when they recognized that he was a Jew, they were shouting with one voice from all of them for about two hours, “Great is Artemis of the Ephesians!”

71. Acts 22:29  
Then immediately those who were about to examine him kept away from him, and the military tribune also was afraid when he realized that he was a Roman citizen and that he had tied him up.

72. Luke 1:22(b)  
And when he came out he was not able to speak to them, and they realized that he had seen a vision in the temple. And he kept making signs to them, and remained unable to speak.

Understand (achievement)

73. Acts 25:25*  
But I understood that he had done nothing deserving death himself, and when this man appealed to His Majesty the Emperor, I decided to send him.

*Excluded from analysis due to a high proportion of hapax legomena*

74. Matt 13:51  
“Have you understood all these things?” They said to him, “Yes.”

75. Matt 17:13  
Then the *disciples* understood that he had spoken to them about John the Baptist.

76. Matt 16:12  
Then they understood that he did not say to beware of the leaven of bread, but of the teaching of the Pharisees and Sadducees.
77. Matt 2:10
   Now when they saw the star, they rejoiced with very great joy.

78. Mark 7:2
   And they saw that some of his disciples were eating their bread with unclean—that is, unwashed—hands.

79. Mark 11:13
   And when he saw from a distance a fig tree that had leaves, he went to see if perhaps he would find anything on it. And when he came up to it he found nothing except leaves, because it was not the season for figs.

80. Mark 8:23
   And he took hold of the blind man’s hand and led him outside the village, and after spitting in his eyes, he placed his hands on him and asked him, “Do you see anything?”

81. Mark 13:26
   And then they will see the Son of Man arriving in the clouds with great power and glory.

82. John 5:6
   Jesus, when he saw this one lying there and knew that he had been sick a long time already, said to him, “Do you want to become well?”

83. John 4:19
   The woman said to him, “Sir, I see that you are a prophet.

84. Acts 12:3
   And when he saw that it was pleasing to the Jews, he proceeded to arrest Peter also. (Now this was during the feast of Unleavened Bread.)

85. John 19:8
   So when Pilate heard this statement, he was even more afraid,

86. Luke 15:25
   “Now his older son was in the field, and when he came and approached the house, he heard music and dancing.

87. Matt 13:19
   When anyone hears the word about the kingdom and does not understand it, the evil one comes and snatches away what was sown in his heart. This is what was sown on the side of the path.
88. Matt 13:23
   But what was sown on the good soil — this is the one who hears the word and
   understands it, who indeed bears fruit and produces, this one a hundred times as
   much, and this one sixty, and this one thirty.”

   Obey

89. Mark 1:27
   And they were all amazed, so that they began to discuss with one another, saying,
   “What is this? A new teaching with authority! He even commands the unclean
   spirits and they obey him.”

90. Col. 3:20
   Children, obey your parents in everything, for this is pleasing in the Lord.

91. 1 Pet. 3:6
   like Sarah obeyed Abraham, calling him lord, whose children you have become
   when you do good and are not frightened with respect to any terror.

92. Luke 8:25
   And he said to them, “Where is your faith?” But they were afraid and were aston-
   ished, saying to one another, “Who then is this, that he commands even the winds
   and the water and they obey him?”

   Believe

93. John 4:50
   Jesus said to him, “Go, your son will live.” The man believed the word that Jesus
   spoke to him, and he departed.

94. John 11:27
   She said to him, “Yes, Lord, I have believed that you are the Christ, the Son of
   God, who comes into the world.”

95. John 20:8
   So then the other disciple who had come to the tomb first also went in, and he saw
   and believed.

96. Acts 15:11
   But we believe we will be saved through the grace of the Lord Jesus in the
   same way those also are.”

97. Acts 9:26
   And when he arrived in Jerusalem, he was attempting to associate with the disci-
   ples, and they were all afraid of him, because they did not believe that he was a
   disciple.
Think

98. John 11:13

(Now Jesus had been speaking about his death, but they thought that he was speaking about real sleep.)

99. Matt 20:10

And when the first came, they thought that they would receive more, and they also received a denarius apiece.

100. John 20:15

Jesus said to her, “Woman, why are you weeping? Who are you looking for?” She thought that it was the gardener, and said to him, “Sir, if you have carried him away, tell me where you have put him, and I will take him.”
C. Appendix: Link to Online Supplementary Materials

https://annasjob.github.io/KnowledgePredication/
References


Davies, Mark. 2008-. The corpus of contemporary American English (COCA). Online corpus. \url{https://www.english-corpora.org/coca/}


Zee (eds.), *Motion Encoding in Language and Space*. Oxford: Oxford University Press.


Loeweke, Eunice & Jean May. 1980. General grammar of Fasu (Namo me): Lake Kutubu, Southern Highlands Province. In Don Hutchisson (ed.), *Grammatical stud-


McIntosh, Justin. 2015. *Aspects of phonology and morphology of Teotepec eastern Chatino*. University of Texas at Austin dissertation.


Östling, Robert. 2015. Word order typology through multilingual word alignment. The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. 205–211.


Universität Leipzig, Sächsische Akademie der Wissenschaften & InfAI. 2023. *Wortschatz Leipzig*. Online corpus. [https://wortschatz.uni-leipzig.de](https://wortschatz.uni-leipzig.de)


Vafaean, Ghazaleh. 2018. *Progressives in use and contact: a descriptive, areal and typological study with special focus on selected iranian languages*. Stockholm University, Department of Linguistics dissertation.


The present thesis is an investigation of the cross-linguistic expression of knowledge predication (‘He knows that it’s raining’, ‘she knows the boy’ etc.) The thesis investigates knowledge predication in parallel texts—specifically New Testament translations—in a genealogically and areally stratified variety sample of 83 languages. A selection of verses representing knowledge predication, as well as near-lying concepts, are manually analysed and annotated for lexical and constructional information, and analysed using both qualitative and quantitative methods. Generalisations regarding the cross-linguistic expression of the knowledge domain are made—describing for example restrictions on co-expression patterns—and a semantic map of the domain of knowledge predication is presented.