SUC-CORE is a 20 000 word subset of the Stockholm-Umeå Corpus 2.0 and Swedish Treebank, annotated with noun phrase coreference. While most coreference annotated corpora consist of texts of similar types within related domains, SUC-CORE consists of both informative and imaginative prose and covers a wide range of literary genres and domains.

Abstract
SUC-CORE is a subset of Stockholm Umeå Corpus 2.0 and Swedish Treebank, annotated with noun phrase coreference. While most coreference annotated corpora consist of texts of similar types within related domains, SUC-CORE consists of both informative and imaginative prose and covers a wide range of literary genres and domains.

1. Introduction
SUC-CORE is a 20 000 word subset of the Stockholm-Umeå Corpus (SUC 2.0) annotated with coreference relations between noun phrases. This subset consists of the same documents as the evaluation set of the Swedish Treebank. Thus, the coreference annotation of SUC-CORE can be combined with the part-of-speech tagging, morpho-syntactic analysis and named entity annotation of SUC 2.0 (Gustafson-ˇCapkov´ a and Hartmann, Eds., 2006), and the syntactic analysis of the Swedish Treebank (Nivre et al., 2008).

Through SUC-CORE, we offer annotated data for development and evaluation of coreference resolution for Swedish. To our knowledge, this is the only Swedish corpus with coreference annotation available for research.

2. Data
SUC 2.0 is a balanced corpus, covering various text types and stylistic levels. It is modeled on the Brown Corpus (Bonelli and Sinclair, 2006) and similar sample corpora with two main categories of texts, informative prose and imaginative prose. The first category consists of e.g., news text, feature articles, and scientific papers, and the second category of different genres of fiction. SUC 2.0 follows the general layout of Brown with 500 samples of text with a length of about 2,000 words each. These text samples are composed of excerpts from longer texts or a selection of short texts (Gustafson-ˇCapkov´ a and Hartmann, Eds., 2006).

SUC-CORE includes both informative and imaginative text of different genres and domains. The informative prose category consists of six files with foreign and domestic news texts and editorials from national and regional morning dailies, magazine articles on interior design, a textbook excerpt on biology, and an academic essay. The imaginative prose section includes excerpts from four novels of different genres (see table 1). Thus, SUC-CORE can be used for development and evaluation against different types of text.

Most comparable data sets annotated with coreference consist of texts of similar types within the same (or related) domains, e.g., newswire text in MUC-6 and MUC-7 (Hirschman and Chinchor, 1997), and newswire, broadcast transcripts, and blogs in ACE (Doddington et al., 2004).
Table 1: Overview of SUC-CORE: file, genre, source, no. of tokens, mentions, and coreference between mentions. Files marked with (*) consist of selections of texts.

<table>
<thead>
<tr>
<th>File</th>
<th>Genre</th>
<th>Source</th>
<th>Tokens</th>
<th>Mentions</th>
<th>Relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>I: Informative prose</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aa05</td>
<td>Press; political reportage* (foreign)</td>
<td>National daily</td>
<td>2056</td>
<td>713</td>
<td>235</td>
</tr>
<tr>
<td>aa09</td>
<td>Press; political reportage* (foreign, domestic)</td>
<td>Regional daily</td>
<td>2073</td>
<td>658</td>
<td>167</td>
</tr>
<tr>
<td>ba07</td>
<td>Press; editorials*</td>
<td>Regional</td>
<td>2100</td>
<td>670</td>
<td>118</td>
</tr>
<tr>
<td>ea10</td>
<td>Skills, trades and hobbies (interior design)*</td>
<td>Periodical</td>
<td>2194</td>
<td>734</td>
<td>154</td>
</tr>
<tr>
<td>ea12</td>
<td>Skills, trades and hobbies (biology)</td>
<td>Textbook</td>
<td>2017</td>
<td>715</td>
<td>178</td>
</tr>
<tr>
<td>ja06</td>
<td>Learned and scientific writing (humanities)</td>
<td>Textbook</td>
<td>2123</td>
<td>755</td>
<td>132</td>
</tr>
<tr>
<td></td>
<td>Total I:</td>
<td></td>
<td>12563</td>
<td>4245</td>
<td>984</td>
</tr>
<tr>
<td>II: Imaginative prose</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kk14</td>
<td>Fiction (Tunström, G. “Det sanna livet”)</td>
<td>Novel</td>
<td>2067</td>
<td>743</td>
<td>343</td>
</tr>
<tr>
<td>kk44</td>
<td>Fiction (Thorvall, K. “När man skjuter arbetare”)</td>
<td>Novel</td>
<td>2016</td>
<td>632</td>
<td>253</td>
</tr>
<tr>
<td>kl07</td>
<td>Crime (Nesser, H. “Det grovmaskiga nätet”)</td>
<td>Novel</td>
<td>2008</td>
<td>664</td>
<td>327</td>
</tr>
<tr>
<td>kn08</td>
<td>Romance (Dagsås, J. “Riddaren i mina drömmar”)</td>
<td>Novel</td>
<td>2004</td>
<td>617</td>
<td>330</td>
</tr>
<tr>
<td></td>
<td>Total II:</td>
<td></td>
<td>8095</td>
<td>2656</td>
<td>1253</td>
</tr>
<tr>
<td>TOTAL I + II:</td>
<td></td>
<td></td>
<td>20658</td>
<td>6901</td>
<td>2237</td>
</tr>
</tbody>
</table>

3.3 Annotation format
Each mention is annotated with a mention type (NAM, NOM, PRO) and connected to a specific span of text through character offsets. In the example below (from file aa05), the following mentions are identified: Robert Mugabe, hans Zanuparti (lit. ‘his Zanu party’), hans (‘his’), partiets stängda dörrar (lit. ‘the party’s closed doors’), partiets (‘the party’).

... officiellt jublade Robert Mugabe och hans Zanuparti på sändagen. Men bakom partiets stängda dörrar måste ...

During annotation, the mention Robert Mugabe with index T5 is marked as mention type NAM and connected to the span 154 to 167 in the source text.

... T5 NAM 154 167 Robert Mugabe
T6 NAM 172 186 hans Zanuparti
T7 NOM 212 235 partiets stängda dörrar
T8 PRO 172 176 hans
T9 NOM 212 220 partiets ...

The coreference annotation is listed as pairwise relations between mentions, e.g., T5 (Robert Mugabe) and T8 (hans, ‘his’). This directed relation can link a mention to a previous or to a subsequent mention.

R1 Coref Anaphora:T8 Antecedent:T5
R2 Coref Anaphora:T9 Antecedent:T6

... This format is similar to the BioNLP Shared Task standoff format.3 We refer the reader to the documentation of SUC-CORE and the BRAT website4 for further details.

4. Distribution
SUC-CORE is distributed by the Linguistics Department at Stockholm University.5 SUC 2.0 and Swedish Treebank are distributed by Språkbanken at Gothenburg University.6

5. Concluding remarks
SUC-CORE covers a wide range of genres and domains. This allows for exploration of coreference across different text types, but it also means that there are limited amounts of data within each type. Thus, future work on coreference resolution for Swedish should include making more annotated data available for the research community.

6. References

4BRAT. http://brat.nlplab.org
5DALI, SU. http://www.ling.su.se/english/nlp/resources