SWE-CLARIN partner presentation: Natural Language Processing Resources from the Department of Linguistics, Stockholm University

Kristina Nilsson Björkenstam, Thomas Björkstrand, Gintarė Grigonytė, Sofia Gustafson-Capkova, Johanna Mesch, Robert Östling, Krister Schönnström, Lars Wallin, Mats Wirén

(1) Section for Computational Linguistics, www.ling.su.se/nlp
(2) Section for Sign Language, www.ling.su.se/sign-language-section
(3) Section for Swedish as a second language for the deaf, http://www.ling.su.se/Sw2LD

Introduction

The aim of the CLARIN Research Infrastructure and SWE-CLARIN is to facilitate for scholars in the humanities and social sciences to access primary data in the form of natural language, and to provide tools for exploring, annotating and analysing these data.

This paper gives an overview of the resources and tools developed at the Department of Linguistics at Stockholm University planned to be made available within the SWE-CLARIN project. The paper also outlines our collaborations with neighbouring areas in the humanities and social sciences where these resources and tools will be put to use.

The Department of Linguistics is part of the Faculty of Humanities and consists of five sections: General Linguistics, Phonetics, Computational Linguistics, Sign Language, and Swedish as a Second Language for the Deaf. While each of these sections has its own research agenda and methodology, they also engage in the major research themes at the department:

• linguistic typology and language documentation
• language acquisition
• sign language

The major contributions to SWE-CLARIN will come from the sections for Computational Linguistics and Sign Language.

Planned collaborations

The section for Computational Linguistics has been collaborating on an informal basis with neighbouring disciplines that depend on rational processing of natural language data. SWE-CLARIN gives us an opportunity to turn these collaborations into cooperations. Past collaborations have involved the following topics:

• Together with the Department of Economics at SU, we have been collaborating on a data set consisting of high-school essays with rich metadata. Previous research has shown that the grading system suffers from discrimination based on social class, gender, ethnicity and age. We developed a system for automatic grading which could improve efficiency in a setting where one seeks to identify candidates for incorrectly graded essays (Östling et al., 2013). Part of the work continues as a collaboration with the project Alfa kaw skruva, funded by Vinnova, the Swedish Pour and Telecom Authority (PTS), and the Swedish Academy.
• Together with the Clinical Text Mining Group at the Department of Computer and Systems Sciences at SU, we are collaborating within the domain of health records. This domain is challenging because the writers are medical professionals and not non-professional writers, which leads to a telegraphic, domain-dependent and specialised language. Following a governmental initiative, health records are being made available on-line. Thus, there is a demand for improved readability of the records, which in turn requires simplification and normalisation of the records, without loss of information (Grigonytė et al., 2014).
• We are collaborating with the National Edition of August Strindberg’s Collected Works and the Department of Literature and History of Ideas at SU on building a linguistically annotated corpus of Strindberg’s literary fiction. This project will expand on earlier work on the Stockholm University Strindberg Corpus (Björkenstam, Gustafson-Capkova & Wirén, 2014).
• We are collaborating with the Institute for Language and Folklore (Institutet för språk och folklivnen) in its support for sign language as a minority language with special status.

Corpora and lexicon

Stockholm-Umeå Corpus (SUC)

Stockholm-Umeå Corpus (SUC) is a balanced corpus of Swedish texts of one million words, annotated with part-of-speech and morpho-syntactic information (Ejerhed et al., 1992) and Named Entities (Källgren, 2006). SUC-CORE is a 20 000 word subset of SUC annotated with coreference relations between NPs (Björkenstam, 2013). SUC-CORE consists of the same documents as the evaluation set of the Swedish Treetbank (Nivre et al., 2008).

Swedish Blog Sentences (SBS)

SBS is a collection of sentences from Swedish blog posts. The texts have been annotated for part-of-speech and Named Entities using Stagger (Östling, 2013). SBS consists of about 2.7 billion tokens, in over 600 000 different blogs (Östling & Wirén, 2013).

Stockholm Internet Corpus (SIC)

SIC is a freely available corpus of Swedish Internet texts. The annotation using Stagger (Östling, 2013) has been manually corrected. The corpus is intended for researchers developing and testing natural language processing tools for Internet texts.

Stockholm University Strindberg Corpus (SUSC)

The Stockholm University Strindberg Corpus (SUSC) consists of seven autobiographical novels by August Strindberg. The texts have been annotated using Stagger (Östling, 2013). The corpus is primarily intended for corpus stylistics (Björkenstam, Gustafson-Capkova & Wirén, 2014).

LONG-MINGLE

LONG-MINGLE is a longitudinal corpus of child-directed speech. It consists of 37 orthographic transcripts of audio and video recordings of free play sessions from longitudinal parent-child dyads with 13 children between 2 and 33 months of age. A subset of this corpus, called MINGLE-3, has been multi-modally annotated with gestures and object-related actions (Björkenstam & Wirén, 2014).

Swedish Sign Language Corpus (SSLC)

The Swedish Sign Language Corpus consists of video recordings of pairs of signers filmed during semi-spontaneous dialogues, narratives and elicitation tasks. There are about 24 hours of video from 42 signers from different regions in Sweden (Mesch, 2012). The corpus is annotated with sgn glosses and utterance-level translations into Swedish, and published with media files and annotation files (Wallin & Mesch, 2014). The SSLC was constructed for the purpose of documenting sign language discourse and to facilitate dictionary-making.

Swedish Sign Language Lexicon

The Swedish Sign Language Dictionary consists of about 10,000 sign entries. Each sign is described in terms of its hand shape, orientation, location, movement, Swedish translations, and sign transcription. Ongoing work includes combining the dictionary with sentence look-up in the SSLC corpus (Mesch, Wallin & Björkstrand, 2012).

Ongoing Sign Language corpora projects

• A corpus of tactic sign language of deaf-blind signers, consisting of video recordings of 8 different signers (male and female) between 38 and 77 years from different regions in Sweden (Mesch, forthcoming).
• A learner corpus of 18 hearing students at SU who are learning Swedish Sign Language as a second language (Schönström & Mesch, 2014, Mesch & Schönström, forthcoming).

Tools

Stagger --- The Stockholm Tagger

Stagger is a Swedish part-of-speech tagger with a pre-trained accuracy of 96.6 percent, which compares favourably to other published results for Swedish (Östling, 2013).

Spacos --- Stockholm Parallel Corpus System

Spacos is a system for finding word alignments in parallel corpora. It is based on Bayesian versions of the IBM alignment models, and uses Gibbs sampling for inference. Spacos is capable of jointly learning word alignments and transferring phrase-of-ignorant-tags, which is useful when trying to align one well-resourced language with a language that lacks basic NLP tools (Östling, 2014).

Automated Essay Scoring

We have developed a system for automatic grading of student essays, using data collected by the Department of Economics at SU. Code and pre-trained models are available (Östling et al., 2013).

Swedish Python Routines (SPyro)

SPyro is a Python library for Swedish NLP, including an interface to the SALDO lexicon and a comond splitter (Östling & Wirén, 2013).

Dialect maps

A tool for investigating regional differences in Swedish using data from bloggers from different parts of Sweden.

Svek: Stockholms Vektormodeller

Svek is a package for finding several different models that represent words using real-valued vectors: CoLLobor & Weston (2008) embeddings and (Reflective) Random Indexing (Cohen et al., 2010).

Stockholm Language Model with Entropy (SLME)

SLME is a N-gram language model, inspired by the Stanford Backoff Model (Brants et al., 2007). It has been designed to provide fast calculation of the entropy of the distribution of words following a given context, \( H(p(\text{word}|\text{context})). \)

Selected references