# Computational Linguistics in Sweden

Bengt SIGURD

**Résumé.**

L'informatique, à l'université de Göteborg, est utilisée principalement pour des travaux de lexicologie sous la direction de Sture Allén; le département de linguistique informatique à Umeå, dirigé par Eva Ejerhed, s'est spécialisé dans l'étude syntaxique, tandis qu'à l'université d'Uppsala, l'équipe d'Anna Sågwall Hein s'attache notamment à l'analyse morphologique et à la traduction automatique. Benny Brodda, qui dirige les recherches d'informatique linguistique à l'université de Stockholm avec Gunnel Källgren, a mis au point le langage *Beta* spécifique aux problèmes linguistiques. Le département est aussi connu pour ses systèmes de recherche automatique de l'information. L'université de Lund a développé sous la direction de Bengt Sigurd un générateur de textes écrits et prononcés, *Commentator*, et travaille à la traduction automatique. Enfin, à l'université de Linköping le laboratoire de Lars Ahrenberg analyse le discours pour créer des interfaces en langage naturel. En ce qui concerne les grands projets de corpus informatiques, il faut noter le grand dictionnaire historique de l'Académie suédoise dont l'enregistrement est en cours à Göteborg, les fichiers de plus de 10 millions de mots suédois réalisés à Umeå et Stockholm, et enfin le *London-Lund Corpus of Spoken English*.

Keywords: computational linguistics, Swedish, computer-aided lexicology, morphological analysis, syntactic analysis, machine translation, text generating system, automatic information retrieval, natural linguage interfaces.

Mots-clés : linguistique computationnelle, suédois, lexicologie assistée par ordinateur, analyse morphologique, analyse de la syntaxe, traduction automatique, système de génération automatique de textes, recherche automatique de l'information, interfaces en langage naturel.

It is fairly easy to give a survey of computational linguistics in Sweden today, as Swedish linguistics has just been the object on an international evaluation and peer review.[1] The book includes a separate chapter called Computational Linguistics and this is one of the sources of this article.

[1] Published in ENKVIST (N.-E.), FERGUSSON (Ch.A.), HAJICOVA (E.) & LADEFOGED (P.): 1992, *Linguistic Research in Sweden* (Uppsala: Swedish Science Press).

✉ Dept of Linguistics; University of Lund; Helgonabacken 12; S-22362 Lund (Sweden).
Fax: + 46 46 10 84 40 E-mail: bengt.sigurd@lings.lu.se

The evaluators' appreciation of Swedish computational linguistics is clear from the wordings of the first of the recommendations presented at the end of the book (p. 83): "1. Research in phonetics and computational linguistics should continue to be well supported, enabling Sweden to keep its advanced position in these areas."

Swedish computational linguistics is mostly to be found in the linguistics departments, not under the heading Natural Language Processing (NLP) in departments of computer science or communications as in some countries. This housing of computational linguistics makes research and teaching of computational linguistics in Sweden oriented towards linguistics rather than computers as will also be obvious from this survey.

One of the pillars of Swedish computational linguistics (Swedish: *datalingvistik, datorlingvistik or språkvetenskaplig databehandling*) is the department of computational linguistics (*Institutionen für språkvetenskaplig databehandling, Språkdata*, Humanisten) at Göteborg, headed by professor Sture Allén since 1968 (also permanent secretary of the Swedish Academy). The research there has focussed on text processing with the purpose to build dictionaries of different types. The department has published several frequency lexica based on 1 million words from Swedish newspapers from the year 1965. The texts are derived from different areas such as editorials, sports, business and it is possible to find the frequencies of different word forms, words (lemmata) or word combinations (fixed phrases) and make interesting statistical studies[2]. These lexica are all derived and constructed by computers and there are several very useful lexical data bases for different practical and theoretical purposes available at the department at Göteborg. The latest version of the authoritative word list of the Swedish Academy (*Svenska Akademiens Ordlista, SAOL*) was produced at Göteborg as well as a number of other lexical products. Using high-tech scanning techniques, the department at Göteborg is also preparing a machine-readable version of the great historical dictionary of the Swedish Academy (*Svenska Akademiens Ordbok, SAOB*), some 30 large volumes including the letter *s* published so far; according to predictions, the project will not be finished until *ca* 2040). The machine-readable version of the very comprehensive dictionary of the Swedish Academy will make up a data base which makes all the information of the dictionary available much easier and faster. The department at Göteborg has turned out various studies of computer-aided lexicology[3].

[2] ALLÉN (S.): 1970–, "Nusvensk Frekvensordbok" 1–4, *Data Linguistica* (Stockholm).

[3] Cf. GELLERSTAM (M.) [ed.]: 1988, "Studies in computer-aided lexicology", *Data Linguistica* (Stockholm).

Beside the first chair in computational linguistics at Göteborg, there are now chairs at Uppsala and Stockholm, but there is research in and teaching of computational linguistics also at the universities of Umeå, Linköping and Lund. Umeå is characterized by a syntactic profile, where different syntactic models, *e.g.* finite state and extended context free grammars are tested[4]. Umeå has also started cooperation with the Stockholm department of computational linguistics in order to prepare large Swedish text corpora including tens of millions of words. The Swedish researchers cooperate with international groups within the *Text Encoding Initiative (TEI)*. These tagged corpora are to be used in testing different computerized grammatical and text linguistic models. It is currently felt that grammatical models have to be tested empirically on large corpora in order to be of value—a reaction against previous "arm chair linguistics". The head researcher at Umeå is professor Eva Ejerhed.

Research in computational linguistics at Uppsala has focused on morphological analysis (including 2-level morphology) and parsing techniques (represented *e.g.* by the *Uppsala Chart Parser*) and developed programs which have been used *e.g.* in automatic identification of Russian word forms and syntactic analysis of Swedish sentences using lexical information extensively[5]. The Uppsala department has also tried to develop tools and aids for machine translation. The head researcher at Uppsala is professor Anna Sågwall Hein who has furthermore worked on computerized information retrieval in medical reports.

The research in computational linguistics at Stockholm University is headed by professor Benny Brodda and Gunnel Källgren. Brodda has developed a high level programming language *Beta* which is very convenient for linguists. It includes commands for typical grammatical, morphological and phonological processing as well as facilities to test parsers, produce statistics and build analyses and lexicons of different types[6]. The department at Stockholm is known for its heuristic parsing and tagging techniques and various projects in automatic information retrieval *e.g.* from legal texts. Some of the research has had commercial support, *e.g.* the programs used for trade mark recognition and generation and the program developed to suggest synonyms in word processing[7].

[4] EJERHED, (E.) "A finite state parser for Swedish with morphological analyzer and semantics", *Proceedings SAIS-86 Workshop* (Linköping Dept. Computer Science).

[5] SÅGWALL HEIN (A.): 1987, "Parsing by means of Uppsala Chart Parser (UCP)" in BOLC (L.) ed., *Natural language parsing systems* (Heidelberg: Springer-Verlag).

[6] BRODDA (B.): 1991, "Doing corpus work with PC Beta" in JOHANSSON (S.) & STENSTRÖM (A.B.) eds, *English computer corpora* (Berlin).

[7] KÄLLGREN (G.) & MAGNBERG (S.): 1990, *AutoSyn. Two reports on a system of automatic suppletion of synonyms* (Papers from the Institute of Linguistics University of Stockholm (PILUS) 59).

As at Umeå, Uppsala and Stockholm computational linguistics at Lund coexists with general linguistics and phonetics. A department of phonetics was, in fact, the embryo of the present-day Department of Linguistics at Lund[8]. The computer resources at Lund are shared by phoneticians and computational linguists and some of the research projects at Lund have included both grammatical and phonetic processing. This is the case *e.g.* for *Commentator*, a computer system (developed by professor Bengt Sigurd) which can comment on the events at a screen either in writing or in speech (using speech synthesis). The events on the screen could be *e.g.* ships moving inside and outside a harbour or airplanes flying over a certain area. The comments make up coherent texts including fairly complex sentences telling *e.g.* whether certain objects were approaching or disappearing. The system is also able to decide which events are worthy of comment, *e.g.* preferring commenting on the appearance of an unidentified submarine to commenting on a local fishing boat. The Commentator is a text generating systems and as all such systems it forces the designer to take stand on a number of general issues of semantics, discourse, syntax, lexicon, morphology and phonology[9].

Lund is also known for its machine translation project *Swetra* (Swedish Computer Translation Research; head Bengt Sigurd), presented at several international conferences (*e.g.* COLING). This project is based on a kind of computerized grammar called Referent Grammar (RG) which includes both features of Lexical Functional Grammar (LFG) and Generalized Phrase Structure Grammar (GPSG). The RG functional representations of different languages are (or are made) very similar and can thus be used as interlingua for automatic translation between languages if the word meanings are also standardized. Swetra is a multilanguage MT-system where grammatical and lexical modules of different languages are interconnected and used bidirectionally, *i.e.* both for analysis and generation. Swetra is able to translate ordinary sentences between English, German, Russian and Swedish at a reasonable speed[10].

One version of Swetra is called *Weathra* as it is designed specifically to translate weather bulletins. Another version is called *Stocktra* as it is designed to translate stock market reports. Some versions of the system resort to word-forword (or phrase-to-phrase, as many of the entries of the lexicon are multi-word

[8] Helgonabacken 12, S–22362 Lund.

[9] SIGURD (B.): 1982, "Commentator: A computer model of verbal production", *Linguistics* 20, pp. 611–632.

[10] SIGURD (B.) & GAWRONSKA (B.): 1988 "The potential of SWETRA a multilanguage MT-system", *Computers and Translation* 3, pp. 237–250.

items) translation if the advanced grammatical approach takes too long time or fails and it has been noted that word-for-word translations are often as good as the products of the more sophisticated methods. The Swetra project has resulted in a number of theoretical papers on syntax, morphology and lexicology beside the computer programs which have been used in demonstrations and teaching[11]. There is also a machine translation project (headed by Wolfgang Koch) at the German department at Lund[12], where cooking recipees are to be translated using sophisticated semantic representations and processes.

Computational linguistics (head Lars Ahrenberg) at the University of Linköping is included in the Computer and Information Science department (in the *Natural Language Processing Laboratory, NLPLAB*). The research group at Linköping has taken a special interest in computerized dialogue analysis and discourse representation in order to build sophisticated natural language interfaces to expert systems and data bases (FALIN). But the Linköping group has also implemented a Swedish LFG and tried to develop a pilot MT-system using a conceptual interlingua. Linköping is a center of computer science in Sweden and computational linguistics is only one of the lines at the big department[13].

It is a characteristic of Swedish linguistics that the need for machine readable large empirical texts was acknowledged early. Beside the projects at Göteborg and Umå-Stockholm mentioned above the best known corpus project is the *London-Lund Corpus of Spoken English (LLC)* headed by professor Jan Svartvik at the Dept of English Lund[14]. This corpus represents different types of dialogue or discourse tagged and analyzed in terms of tone units, pauses and grammatical categories[15]. A number of computer programs (written by Mats Eeg-Olofsson) can be used in the processing of the texts for various purposes.

There is a great deal of cooperation and contacts between the Nordic countries and there are regular Nordic meetings of computational linguistics (*Nordiska Datalingvistdagarna*). As is natural, Lund and Copenhagen have very close contacts — travel only takes a few hours by boat and train. The

[11] SIGURD (B.), EEG-OLOFSSON (M.), WILLNERS (C.) & JOHANSSON (Chr.): 1992, "Automatic translation in specific domains", *Praktisk Lingvistik* 15 (Lund: Dept of Linguistics).

[12] Helgonabacken 14, S–22362 Lund.

[13] AHRENBERG (Lars): 1989, "On the integration of linguistic knowledge and world knowledge in natural language understanding" in DAHL (Ö) & FRAURUD (K.) eds, *Papers from the first Nordic Conference on Text Comprehension in Man and Machine, Sigtuna, Oct. 27–28, 1988* (Department of Linguistics, University of Stockholm).

[14] Helgonabacken 14, S–22362 Lund.

[15] SVARTVIK, (J.) ed.: 1990, *The London-Lund corpus of spoken English. Description and research* (Lund: Lund University Press, "Lund Studies in English", 82).

other important contact exists between Stockholm and Helsingfors, but Swedish computational linguists are common participants at international conferences and connected to many other research groups through international or European networks and the number of cooperative efforts within ESPRIT and the like are increasing.