"The first million is hardest to get": Building a Large Tagged Corpus as Automatically as Possible

Gunnel Källgren
University of Stockholm
Department of
Computational Linguistics
S-106 91 Stockholm
Sweden
gunnel@com.qz.se

Summary: The paper describes a recently started project in Sweden. The goal of the project is to produce a corpus of (at least) one million words of running text from different genres, where all words are classified for word class and for a set of morpho- syntactic properties. A set of methods and tools for automating the process are being developed and will be presented, and problems and some solutions in connection with e.g. homography disambiguation will be discussed.

Key words: corpus work, tagging, parsing, probabilistic methods

0. This paper basically consists of three parts: 1. a brief sketch of a newly started corpus project, 2. a discussion of the problems that this and similar projects will run into as well as of the expected results and possible further developments, and 3. a presentation and demonstration of implemented and running programs that are used on the corpus material.

An important aspect of presenting a project of the size of this one at an early stage is our need for feedback. We realize clearly that we are heading straight for some grandiose mistakes that will cost us time, effort, and headaches, but the Coling participants, by sharing their experience with us, might save us from at least some of the mistakes.

1. The project to be described in this paper started in the autumn of 1989. It is carried out in cooperation between the departments of Linguistics at the universities of Stockholm (Dr. Gunnel Källgren) and Umeå (Professor Eva Ejerhed), and it is supported by the Swedish Research Council for the Humanities and the Swedish National Board for Technical Development.

As a substantial part of the project, we will build up a large corpus of written Swedish. By 'large', we mean at least 1 million words for a start, with an explicit aim of collecting considerably more. The corpus will, as far as possible, be composed of texts from various genres in a way that will match the principles of the Brown and LOB corpora (cf. Francis & Kuc'era 1982, Garside, Leech & Sampson 1987). We will however make one important change of those principles; rather than cutting the text samples at the first sentence boundary after 2,000 words, we will strive for texts or subparts of texts that form a coherent whole.

The construction of a large corpus is, however, not a goal in itself. The corpus is meant to function as a test-bench and a basis for comparison in the development and testing of various models for analysis. In order to have this function, the corpus must be tagged with at least the word-class, the flectional form, and the lemma form of each word. This kind of tagging is, with some exceptions, rather theory-neutral and uncontroversial, but it has to be done correctly and unambiguously. To manage that, without an overwhelming amount of manual labor, we have to develop different kinds of methods and tools, and also to find and use methods and tools developed by other researchers. Once such a million-word corpus exists, proof-read and cross-checked for consistency, it will form an invaluable basis for many kinds of linguistic investigations, but the methods developed and refined in building the corpus will be quite as important as an output, as they can be used for building the even larger corpora that will be necessary for certain kinds of large-scale linguistic analysis. We hope to be able to take a considerable step towards a fully automatized tagging of unrestricted text, but - as noted in the heading and as many multi-millionaires have also noted - the first million is the hardest to get.

Along with the methods for building the corpus, we will also develop a set of simple tools for using it: programs for search on different levels, for excerption and building of concordances, for sorting according to different criteria, and, not least important, for the addition of the user's own tags in a standardized and compatible format. The resulting corpus and the 'tool-kit' for using it will be made available to other researchers in a format suitable for different kinds of personal computers. This will, hopefully, facilitate and thereby increase the research on modern written Swedish.

We must however admit that we do not do this out of an unselfish concern for others. On the contrary our original impetus was a very selfish need to be able to test and develop our own models and ideas on a large scale. This is where the fun really starts, but as most of this is so far not implemented, or only to a small degree, I will not say anything more about it here, but will return briefly to some of it in the section about expected problems and expected results.

2. To build a large corpus in a short time, we will have to rely almost entirely on material that is computer-readable from the beginning, i.e. mostly material that is typeset on computers. This will bring us into a jungle of non-linguistic but time-consuming problems: getting access to data tapes, loading them to our own computers, converting between different formats and different character conventions, deciding which typographic features can be discarded and which contain information that must be kept, deciding how to treat pictures, figures, and diagrams, etc.

On top of this, we have the questions of coverage and representativity. If we could just take any kind of computer readable text until we get a large enough corpus everything would be so much easier, but now we have to find texts from many different genres and, consequently, from many different sources. This will multiply problems of the type mentioned in the preceding paragraph, but it will also force us to cope with copyright restrictions. Our wish to cover different kinds of text genres, including fiction, in combination with our wish to have texts that are coherent wholes and to make all the tagged texts generally available for research purposes, will here bring us in conflict with copyright regulations. If necessary, we will change the proportions between different genres rather than have parts of the corpus not generally available.

The problems sketched in the last two paragraphs are certainly of importance but I will not discuss them here. Rather, I will describe some of the truly linguistic matters we have to deal with and, in the last section, proceed to show possible solutions to some of them.

The best basis for the kind of tagging we want to do is a computerized lexicon that covers as much as possible of the vocabulary of unrestricted text, and that gives as much as possible of the morpho-syntactic information we want to represent. In this respect, we are extremely lucky in that we can have access to three different computerized lexica designed for analysis of Swedish word forms. By the kind permission of the respective lexicon builders, we can test their models and pick the one that suits our special purposes best. The three lexica are the TWOL-lexicon from the University of Helsinki (Karlsson forthcoming), the lexicon from Chalmers University of Technology, Gothenburg, that was originally designed for speech synthesis (Hedelin et al. 1989), and the morphological analyzer developed within the LPS-project at the university of Gothenburg (Sågvall 1989).

This possibility of lexicon look up brings the project a great leap forward, but, alas, with much left to be done. According to statistics (Allén 1970, p. xxv), almost 65% of the word tokens in Swedish texts are ambiguous. (The corresponding figures for English and Finnish are 45% and 15%, respectively. Cf. DeRose 1988, Karlsson forthcoming.) The large figure for Swedish may seem astonishing, but a careful manual check of the output from the look up of 2,000 running words in the Helsinki TWOL-lexicon showed that at least 55% of all words were ambiguous in any way, with an average of 2.6 readings of each ambiguous word.

Ambiguities can be between lemmas (word types) from different word classes, different lemmas within the same word class, and different inflectional forms within the same lemma. A typical example would be the word 'glada' that can either be a noun, the name of a bird, or an adjective, meaning 'happy'. As an adjective, the word is manyways ambiguous between a singular definite reading that can be either neuter or common gender, and a plural reading that can be definite or indefinite and in either case belong to either gender. Ambiguity between different lemmas within the same word class is a less common type. It can be seen in a word (token) like 'svans' that can either mean 'tail' or be the genitive form of 'swan'. We have not counted as ambiguities polysemous words with identical inflectional pattern, like 'krona', which is either 'a crown' or the Swedish currency unit. All these ambiguities have to be sorted out in the disambiguation process.

In this disambiguation, we will mainly use robust methods that for every ambiguous situation will come up with a best possible solution. (Cf. Källgren 1984a,b,1990, Brodda 1983.) This will partly be based on another important step in the process, namely the construction of constituents, in particular noun phrases and prepositional phrases (Church 1988, Källgren 1984c), and partly on a more general algorithm that for pairs or longer sequences of tags calculates the relative probability of alternative tag assignments. The principles behind such algorithms are known, but they have never been tried on Swedish material (DeRose 1988, Marshall 1987, Eeg-Olofsson 1985).

An indispensable step in the disambiguation process is the assignment of clause boundaries, which presupposes established constituents at the same time as it forms an important basis for disambiguating chains of tags. Methods for this are being tested out on Swedish material (Ejerhed 1989). Given this, it might be possible to check the valency structure of predicates, to decide subject and direct object and, more difficult, to decide the role of prepositional phrases in relation to the finite verb.

In all the above steps, we will use robust methods that can give a straightforward, 'flat' analysis of the surface sentences. The final output will be carefully proofread and can then function as a corpus for empirical research, a test-bench for theoretical linguists' models, and a training material in the development of stochastic methods of analysis (cf. Källgren 1988).

3. Several of the programs needed in the project already exist, at least as running prototypes, and can be demonstrated. Among those are a system for converting from the explicit tags of the TWOL- lexicon to our more condensed and sometimes different tags, as well as from our condensed tags to an explicit transcription of them. (Our tags sometimes have a finer subclassification than is at present the case with the TWOL-tags.) In connection with this, we are willing to discuss our set of tags, which, by necessity, is a compromise between what is wanted and what can be achieved with a reasonable amount of effort. Our technique of using temporary, ambiguous tags to postpone decisions in non-deterministic situations will also be discussed. Below are the suggested tags of the word 'hoppa' ('to jump' or 'jump!') given as an example.

Output from TWOL-lexicon:
hoppa "V IMP/INF"
Condensed temporary tag:
Vl1a < hoppa >
where: V = finite verb, l = lexical verb
(i.e. not copula, modal, or auxiliary),
1 = imperative or infinitive, a = active, belonging to the lemma hoppa

Data driven disambiguation procedures can then be applied. The disambiguation will be triggered and governed by the '1', in this case directed to look for, e.g., a preceding auxiliary verb or infinitive marker signalling infinitive as opposed to the possible syntactic environments of imperatives. Assuming that the word appears in a context where it functions as an infinitive, the output will be 'Vlia<hoppa> 'else 'Vlma<hoppa>', but even before this decision is reached, the information that the word is not in any of the other tenses can be used by other disambiguation procedures.

For the disambiguation, we have started on a first prototype of a 'learning' program, i.e. the program can be trained to make a best possible choice in different situations, where the situations are sequences of ambiguous tags (Karlgren 1989). It is a Prolog implementation of principles presented in Källgren (1984b).

For further analysis of the corpus we have a program that identifies subject and direct and indirect object in simple and complex sentences. It is based on an algorithm that has been tested manually (Källgren 1987) with good results, and has now been implemented as an expert system with a set of if...then-rules (Magnberg 1990). The program presupposes that word class disambiguation, constituent construction, and clause boundary identification has been carried out. It will be demonstrated at Coling.

To facilitate the use of the corpus also for non-computational linguists, we plan to supply the completed corpus with a packet of tools. As an example of such tools, a version of the Beta system that is especially designed for making excerptions and concordances on personal computers will be demonstrated (Brodda 1990a, b).

REFERENCES

Allén, S. 1970. Nusvensk frekvensordbok baserad på tidningstext 1. Almqvist & Wiksell, Stockholm.

Brodda, B. 1983. An experiment with heuristic parsing of Swedish. *Proceedings of the First Conference of the European Chapter of the ACL*, Pisa.

Brodda, B. 1990a. *Corpus Work with PC Beta*: a Presentation. In this volume.

Brodda, B. 1990b. *Corpus Work with PC Beta*. Institute of Linguistics, University of Stockholm.

Church, K.W. 1988. A stochastic parts program and noun phrase parser for unrestricted text. *Proceedings of the Second Conference on Applied Natural Language Processing, ACL*, Austin, Texas.

DeRose, S.J. 1988. Grammatical category disambiguation by statistical optimization. *Computational Linguistics* Vol. 14:1.

Eeg-Olofsson, M. 1985. A probability model for computer aided word class determination. *ALLC Journal* 5:1&2.

Ejerhed, E. 1989. A Swedish clause grammar and its implementation. In: Rögnvaldsson, E. & Pind, J. (eds.), *Papers from the Seventh Scandinavian Conference of Computational Linguistics*. Reykjavik.

Francis, W.N. & H. Kuc'era. 1982. *Frequency analysis of English usage: lexicon and grammar*. Houghton Mifflin.

Garside, R., G. Leech & G. Sampson (eds.). 1987. *The Computational Analysis of English*. Longman.

Hedelin, P., A. Jonsson & P. Lindblad. 1989. Svenskt uttalslexikon, Del I och II. *Teknisk rapport* nr. 4, Institutionen för Informationsteori, Chalmers University of Technology, Gothenburg.

Källgren, G. 1984a. HP-systemet som genväg vid syntaktisk märkning av texter. *Svenskans beskrivning* 14, Lund.

Källgren, G. 1984b. HP - A heuristic finite state parser based on morphology. *De nordiska datalingvistikdagarna* 1983, Uppsala.

Källgren, G. 1984c. *Automatisk excerpering av substantiv ur löpande text. Ett möjligt hjälpmedel vid datoriserad indexering*? Institutet för Rättsinformatik, Stockholm.

Källgren, G. 1987. What Good is Syntactic Information in the Lexicon of a Syntactic Parser? In *Nordiske Datalingvistikdage 1987*, Lambda no. 7, Copenhagen University 1988.

Källgren, G. 1987. What Good is Syntactic Information in the Lexicon of a Syntactic Parser? In *Nordiske Datalingvistikdage 1987*, **Lambda no.** 7, Copenhagen University 1988.

Källgren, G. 1988. Linguistic Theory and Large-Scale Natural Language Processing. In: *ELS Conference on Natural Language Applications*, IBM Norway, Oslo 1988.

Källgren, G. 1990. *Making maximal use of morphology in large-scale parsing*. Institute of Linguistics, Stockholm University. Submitted for publication.

Karlgren, J. 1989. *Nagelfar - Statistically Based Grammatical Category Disambiguation*. Institute of Linguistics, Stockholm University.

Karlsson, F. (forthcoming). *A Comprehensive Morphological Analyzer for Swedish*. Manuscript, Department of General Linguistics, University of Helsinki.

Magnberg, S. 1990. *A Rule-Based System for Identifying Sentence Subjects in Swedish*. Project Report. Institute of Linguistics, Stockholm University.

Marshall, I. 1987. Tag selection using probabilistic methods. In: Garside, R., G. Leech & G. Sampson (eds.), 1987.

Sågvall Hein, A. 1989. Lemmatizing the definitions of Svensk Ordbok by morphological and syntactic analysis. A pilot study. In: Rögnvaldsson, E. & Pind, J. (eds.), *Papers from the Seventh Scandinavian Conference of Computational Linguistics*. Reykjavik.