Parsing Swedish

Atro Voutilainen
Conexor oy
atro.voutilainen@conexor.fi

This paper presents two new systems for analysing Swedish texts: a light parser and a functional dependency grammar parser. Their design follows two Helsinki-based frameworks: Constraint Grammar CG (Karlsson & al 1995) and Functional Dependency Grammar FDG (Tapanainen and Järvinen 1997).

CG and FDG

CG is a reductionistic constraint rule formalism whose input is lexically analysed ambiguous text and whose output is disambiguated text. Disambiguation is carried out by constraints on lemmas and tags that discard alternative analyses on the basis of contextual information, typically coded by a linguist. The ENGCG morphosyntactic tagger was introduced in 1992 (Voutilainen & al.) and compared with a state-of-the-art statistical tagger in 1997 (Samuelsson & Voutilainen).

CG was successful in word-class tagging but not adequate for full-scale parsing. A considerable effort on finite-state parsing was made by Koskenniemi, Tapanainen and Voutilainen (see their articles in Roche & Schabes, eds., 1997). A more successful effort was made by Tapanainen and Järvinen, who extended CG into a functional dependency grammar formalism and interpreter/compiler capable of introducing explicit functional dependencies and of applying large grammars efficiently.

Earlier work on Swedish tagging and parsing

As discussed in Voutilainen (forthcoming 2001), most efforts at Swedish tagging and parsing have focused on wordclass tagging, mostly in the statistical paradigm. A somewhat more informative analysis is given by Lingsoft's SWECG (morphology + function tags) and shallow finite-state Abney-style parsers (Kokkinakis & Johansson 1999). The Swedish Core Language Engine (Gambäck 1997) produces full syntactic parses, but, as argued by Gambäck, it seems to work only for texts from very restricted domains.

Swedish Light Syntax

In design, SweLite follows Conexor's EngLite (see demo at www.conexor.fi). The first major component is the morphological analyser, based on a recent extension of Koskenniemi's Two-Level formalism. The analyser contains a large lexicon, morphology and guesser for unknown words. The morphological analyser produces analyses, many of them ambiguous. The parser uses mapping statements to introduce light syntactic ambiguity, so before any disambiguation is done, an ambiguous analysis looks like this:

```
"<tvingas>"
        "tvinga" <Pass> V INF &MV
        "tvinga" <Pass> V PRES &MV
        "tving#as" <Neu> <Indef> N SG/PL NOM &>N &NH
        "tv#in|gas" <Utr> <Indef> N SG NOM &>N &NH
```

Morphological alternatives are given on lines of their own; syntactic ambiguity is shown as the occurrence of several syntactic tags (here: *&MV* = main verb; *&NH* = nominal head; *&>N* = premodifier). Disambiguation is carried out with hand-coded contextual constraints. Here is a sample parse in tabular format:

Man	man	&NH PRON SG NOM
ställer	ställa	&MV V PRES
upp	upp	&AH ADV
verkligt	verkligt	&>A ADV
höga	hög	&>N A NOM
mål	mål	&NH N NOM
,	,	
som	som	&NH PRON NOM
tränarna	tränare	&NH N PL NOM
och	och	&CC CC
skidåkarna	skid#åkare	&NH N PL NOM
tvingas	tvinga	&MV V PRES
leva	leva	&MV V INF
med	med	&AH ADV &AH PREP
.	.	

On the basis of light syntactic tags and morphology, identification of basic linguistic entities, e.g. nominal phrases, is possible. Identifying relations between the entities requires more information.

Swedish FDG

Tapanainen and Järvinen (1997) give examples of indexing rules whereby functional dependencies between words can be introduced. In the best case, a successful grammar gives a complete dependency structure; in practice many sentences receive only partial dependencies (e.g. due to gaps in the grammar or structural peculiarities in the sentence).

A dependency grammar was written for Swedish. The goal of the present grammar is to show the main nominal arguments as well as relations between clauses. The functional description of adverb phrases and prepositional phrases (e.g. agent, source, goal, benefactive, time) remains to be described in a future version.

Here is a sample parse for a newspaper sentence (`One puts up really high goals that trainers and skiiers are forced to live with.'):

1	Man	man	subj:>2	PRON SG NOM &NH
2	ställer	ställa	main:>0	V PRES &MV
3	upp	upp	advl:>2	ADV &AH
4	verkligt	verkligt	ad:>5	ADV &>A
5	höga	hög	attr:>6	A NOM &>N
6	mål	mål	obj:>2	N NOM &NH
7	,	,		
8	som	som	pcomp:>14	PRON NOM &NH
9	tränarna	tränare	subj:>12	N PL NOM &NH
10	och	och	cc:>9	CC &CC
11	skidåkarna	skid#åkare	cc:>9	N PL NOM &NH
12	tvingas	tvinga	mod:>6	V PRES &MV
13	leva	leva	obj:>12	V INF &MV

14	med	med	advl:>13	PREP &AH
15	.	.		

Column 5 contains morphology and light syntax; column 4 shows functional dependencies. For instance, the pronoun *Man* acts as subject for word number 2, *ställer*. Functional dependencies are not given for every word; partial dependencies result in the analysis of sentences where the grammar or lexicon is incomplete or mispredictive. In practice, around 70% of `normal' nonfiction utterances get a complete dependency analysis (i.e. every word in the sentence gets a regent); the remaining sentences get partial dependencies to enable at least some level of usefulness for further processing e.g. in knowledge-intensive applications.

Let us look at a few other sentences. First, an example of coordination (`The old dog and the young cat eat breakfast and then sleep on the mat.'):

0				
1	Den	den	det:>3	DET SG NOM &>N
2	gamla	gammal	attr:>3	A NOM &>N
3	hunden	hund	subj:>8	N SG NOM &NH
4	och	och	cc:>3	CC &CC
5	den	den	det:>7	DET SG NOM &>N
6	unga	ung	attr:>7	A NOM &>N
7	katten	katt	cc:>3	N SG NOM &NH
8	äter	äta	main:>0	V PRES &MV
9	frukost	frukost	obj:>8	N SG NOM &NH
10	och	och	cc:>8	CC &CC
11	sedan	sedan	advl:>12	ADV &AH
12	sover	sova	cc:>8	V PRES &MV
13	på	på	advl:>12	PREP &AH
14	mattan	matta	pcomp:>13	N SG NOM &NH
15	.	.		

Two coordinations are involced: the subjects *Den gamla hunden och den unga katten* and the clauses (verbs) *äter frukost och sedan sover*. In both cases, the first coordinated element is treated as a regent: it bears the function of the coordination and enters a dependency relation with another word. So, for example, *hunden* is labelled as a dependent (subject) of word number 8, *äter*. The coordinating conjunction and non-initial coordinates are regarded as dependents (coordination) of the first coordinate, so for instance words number 4 and 7 are regarded as dependents of *hunden*. Likewise, word number 10 *och* and 12 *sover* are shown as dependents of word number 8 *äter*.

Next, let us look at a complex sentence (from the Finnish *Hufvudstadsbladet* newspaper), *En fransk regeringstalesman meddelade att rapporten nu skickas till tretton EU-länders regeringar, som diskuterar vad man skall göra beträffande sanktionerna*. (`A French government speaker told that the report will now be sent to governments of thirteen EU countries that will discuss what one will do concerning the sanctions.')

0				
1	En	en	det:>3	DET SG NOM &>N
2	fransk	fransk	attr:>3	A SG NOM &>N
3	regeringstalesman	regerings#talesman	subj:>4	N SG NOM &NH
4	meddelade	meddela	main:>0	V PAST &MV
5	att	att	pm:>8	CS &CS
6	rapporten	rapport	subj:>8	N SG NOM &NH

7	nu	nu	advl:>8	ADV &AH
8	skickas	skicka	obj:>4	V PRES &MV
9	till	till	advl:>8	PREP &AH
10	tretton	tretton	attr:>11	<Card> NUM PL NOM &>N
11	EU-länders	EU-#land	attr:>12	N PL GEN &>N
12	regeringar	regering	pcomp:>9	N PL NOM &NH
13	,	,		
14	som	som	subj:>15	PRON NOM &NH
15	diskuterar	diskutera	mod:>12	V PRES &MV
16	vad	vad	obj:>19	PRON NOM &NH
17	man	man	subj:>18	PRON SG NOM &NH
18	skall	skall	v-ch:>19	V PRES &AUX
19	göra	göra	obj:>15	V INF &MV
20	beträffande	beträffande	advl:>19	PREP &AH
21	sanktionerna	sanktion	pcomp:>20	N PL NOM &NH
22	.	.		

This sentence starts with a main clause *En fransk regeringstalesman meddelade* whose nucleus *meddelade* is linked to the axiom. One of the dependents of *meddelade* is the object clause *att rapporten nu skickas till tretton EU-länders regeringar..*; this object status is shown by the functional dependency of *skickas*. This object clause itself is a complex one: a modifier clause *som diskuterar..* is revealed as such by *mod:>12* for *diskuterar*. This verb governs the nominal clause *vad man skall göra beträffande sanktionerna*; this nominal clause is given the status of object, as shown by *obj:>15* for *göra*.

Note in passing that the topicalised object *vad* is described correctly as such; as shown by Tapanainen and Järvinen, the present formalism is suitable for the description of non-projective phenomena as well, which are problematic for several other frameworks (e.g. Link Grammar).

An informal evaluation

The Swedish FDG parser (a development version from 30. 4. 2001) was tested against newspaper articles (6149 words, 406 sentences, 2.-3. 5. 2001) from *Hufvudstadsbladet* and *Dagens Nyheter*. To allow some degree of comparison to another system (Tapanainen and Järvinen 1997), the parser's ability to identify the heads of subjects (S), objects (O) and subject complements (SC) and to link them to their proper regents (main verbs) was measured in terms of precision (the ratio **obtained desired analyses / all obtained analyses * 100**) and recall (the ratio **obtained desired analyses / all desired analyses * 100**).

This evaluation was carried out by examining the (visual) output of the parser, so some correct and incorrect or partial analyses may have remained undetected. I believe even this evaluation gives a reasonably realistic picture of the system's ability to identify these categories. Here are the results (Tapanainen and Järvinen's corresponding results are given in parentheses):

	precision	**recall**
S	98% (95%)	92% (83%)
O	95% (94%)	90% (88%)
SC	97% (92%)	95% (96%)

The Swedish parser seems to be less `prudent' than the English one; there are more complete dependencies (i.e. every word gets a regent in the sentence) in the analyses. This probably shows in

the higher recall of the Swedish system. Maybe contrary to expectations, also the precision of the Swedish system appears to be higher, though with a smaller margin (with the exception of SC).

Of course, one should interpret this comparison with a grain of salt: the evaluation methodology was somewhat informal in both experiments; the functional categories may contain some slight differences; and the texts and languages are not quite identical, either.

Another way of looking at the parser's ability to analyse these main nominal FDs is to compare the number of sentences with S/O/SC to sentences with completely correct S/O/SC analyses. In this data, there were 371 sentences with at least one S or O or SC; of these sentences, 291 (78%) received a faultless analysis as far as functional dependency analysis of S/O/SC goes; each of the remaining 80 sentences contained at least one incomplete or incorrect S/O/SC analysis.

Technical information

The Swedish FDG parser is fast: on a fast PC with Linux, it analyses well over a thousand words per second. The Swedish Lite parser is about 2.5 times faster than FDG.

Both systems, like Conexor's analysers for other languages (English, French, German, Spanish, Finnish) are available for Linux, Sun Solaris, WIN/NT (COM) and Java.

The Swedish parsers become testable on-line at http://www.conexor.fi during the summer of 2001.

References

Björn Gambäck. 1997. *Processing Swedish Sentences: A Unification-Based Grammar and some Applications*. The Royal Institute of Technology and Stockholm University.

Fred Karlsson, Atro Voutilainen, Juha Heikkilä and Atro Anttila (Eds.) 1995. *Constraint Grammar. A language-independent system for parsing unrestricted text*. Mouton de Gruyter.

Kokkinakis D. and Johansson Kokkinakis S. (1999), A Cascaded Finite-State Parser for Syntactic Analysis of Swedish, *EACL'99*.

Three articles (Koskenniemi, Tapanainen, Voutilainen) in Emmanuel Roche and Yves Schabes (Eds.), *Finite State Language Processing*. MIT Press.

Christer Samuelsson and Atro Voutilainen 1997. Comparing a Linguistic and a Stochastic Tagger. *EACL-ACL'97*.

Pasi Tapanainen and Timo Järvinen 1997. A non-projective dependency parser. *ANLP'97*.

Atro Voutilainen, Juha Heikkilä and Arto Anttila 1992. *English Constraint Grammar*. Dept. of General Linguistics, University of Helsinki. Publications 21.