What kinds of trees grow in Swedish soil?

A Comparison of Four Annotation Schemes for Swedish

Joakim Nivre

Växjö University
School of Mathematics and Systems Engineering
E-mail: Joakim.Nivre@msi.vxu.se

1 Introduction

One of the issues brought up in this workshop concerns the relationship between the syntactic properties of a given language and the choice of linguistic theory for annotation purposes. Our Swedish treebank consortium, consisting of researchers from Växjö University, KTH and Stockholm University, is currently facing a specific instance of this issue in trying to define an annotation standard for a large-scale treebank of Swedish written and spoken language.

In this paper, I will discuss and compare four different annotation schemes that have been proposed for Swedish in terms of their suitability for Swedish syntax as well as their relationship to linguistic theory and annotation schemes proposed for other languages. Other aspects that will be touched upon are the availability of parsers and/or annotated training data for developing parsers, the different requirements for annotation of spoken and written language, and the different needs of different user groups.

By way of background, I will start by reviewing some basic facts about the syntax of Swedish, a Germanic verb second language with moderately fixed word order. In doing this I will also introduce the Scandinavian tradition of descriptive grammar, in particular the influential field model due to Diderichsen [6]. The background section also contains a brief discussion of existing annotation schemes for other languages and their relation to current linguistic theory.

The main part of the paper will be devoted to a discussion and comparison of the following four annotation schemes for Swedish:

- MAMBA (Teleman [29])

- SynTag (Järborg [12])

- SWECG (Birn [2])

- S-CLE (Gambäck [7])

The four schemes fall naturally into two groups, MAMBA and SynTag being standards designed for manual annotation of corpus material, while SWECG and S-CLE are primarily general purpose parsing systems which have corpus annotation as one of their (potential) applications.

2 Background

2.1 Swedish Grammar

Swedish is a Germanic language characterized by a fairly reduced inflectional morphology and a relatively flexible word order. Nouns are inflected for number, definiteness and case (with only two cases remaining except for pronouns); verbs are inflected for tense only (with several tense forms being formed analytically); adjectives agree with their head nouns with respect to gender, number and definiteness. Syntactically, Swedish is a typical verb second language, which means that the finite verb occupies the second position in most main clauses (the notable exception being polar interrogatives).

The structure of Swedish main and subordinate clauses can be conveniently described using a grammatical field model of the kind pioneered by the Danish grammarian Paul Diderichsen [6], which has become firmly established as the principal descriptive framework for Swedish syntax. A modified version of this model is used, for example, in the recently completed Swedish Academy Grammar (Teleman et al. [30]). The schema for Swedish main clauses, consisting of three main fields, is depicted in Figure 1.

Initial field	Middle field			End field		
	V_1	N_1	A_1	V_2	N_2	A_2
	Finite V	Subject	Sentence adverbials	Nonfinite V Particles	Complements	Adverbials (TPM)

Figure 1: Swedish main clause schema

The initial field may be occupied by any major constituent of the clause, the unmarked case being the subject of the clause. The middle field contains three positions, marked V_1, N_1 and A_1.[1] The first of these is the position of the finite verb, the second is the position of the subject (if it is not placed in the initial field), and the third is the position for sentence adverbials (including negation). The end field also contains three positions, marked V_2, N_2 and A_2, the first containing nonfinite verb forms and verb particles, the second containing complements of the main verb, and the last containing adverbial modifiers, notably time, place and manner adverbials, as well as the agent in passive clauses. Figure 2 exemplifies the flexibility of Swedish word order by giving several variants of the same main clause, illustrating their analysis in the field model.[2]

The first example in Figure 2 is a polar interrogative, which is characterized by the initial field being empty and the finite verb therefore occupying the first position. The second example illustrates the unmarked word order in declarative main clauses, with the subject in the initial field. The following three examples show that sentence adverbials, postverbal complements, and end field adverbials can all be placed in the initial field. The final example, which is placed below the English gloss, illustrates the formation of *wh*-interrogatives, which syntactically have the same structure as declaratives except that the *wh*-phrase, in this case *vad* (what), is placed in the initial field. In the interest of saving space, I have chosen examples where all the constituents are realized by single words, but all positions except V_1 may of course be filled by longer phrases. In addition, the end field positions may contain more than one phrase (e.g. two objects in the N_2 position).

The word order in subordinate clauses is much stricter in Swedish and follows a different schema,

---

[1] In general, positions marked V_i hold verb forms (including particles), positions marked N_i hold complements of the verb (including the subject), while positions marked A_i hold adverbials and other optional modifiers.

[2] These examples are based on a similar set of examples in Jörgensen and Svensson [14].

Initial field	Middle field			End field		
	V_1	N_1	A_1	V_2	N_2	A_2
—	har	hon	inte	läst	boken	ännu
hon	har	←—	inte	läst	boken	ännu
inte	har	hon	←—	läst	boken	ännu
boken	har	hon	inte	läst	←—	ännu
ännu	har	hon	inte	läst	boken	←—
	has	*she*	*not*	*read*	*the-book*	*yet*
vad	har	hon	inte	läst	←—	ännu

Figure 2: Swedish main clauses analyzed in the main clause schema

which is exemplified in Figure 3. The main difference is that the initial field is replaced by a complementizer field, and that the order of the positions in the middle field is N_1–A_1–V_1 instead of V_1–N_1–A_1. The first example is a complement clause introduced by the subjunction *att* (that). The two examples below the English gloss illustrate relative clause formation, with the relative pronoun *som* realizing the subject function in the first example (corresponding to *who has not read the book yet*) and the object function in the second example (*which she has not read yet*).

Comp field	Middle field			End field		
	N_1	A_1	V_1	V_2	N_2	A_2
att	hon	inte	har	läst	boken	ännu
that	*she*	*not*	*has*	*read*	*the-book*	*yet*
som	←—	inte	har	läst	boken	ännu
som	hon	inte	har	läst	←—	ännu

Figure 3: Swedish subordinate clause schema with examples

As we will see later on, the field model of grammatical description has been influential not only in the Swedish grammar tradition but has also played a role in previous attempts at grammatical corpus annotation for Swedish. This is hardly surprising since the robustness and familiarity of the model makes it a natural starting point for any analytical scheme aimed at large scale analysis of Swedish written or spoken language. In this context, it should also be mentioned that several modifications to Diderichsen's original field model have been proposed in order to make it more suitable for the analysis of spoken language. Thus, Teleman et al. [30] present an extended main clause schema with two additional fields, occurring before and after the three fields of the original model. This model is exemplified in Figure 4.

Extended clause							
Pre-field	Main clause proper						Post-field
	Initial field	Middle field			End field		
		V_1	N_1	A_1	V_2	N_2	A_2
nej	hon	har	←—	inte	läst	boken	ännu	eller hur?
no	*she*	*has*		*not*	*read*	*the-book*	*yet*	*has she?*

Figure 4: Extended main clause schema

2.2 Treebanks and Linguistic Theory

The number of treebanks available for different languages is growing steadily and with them the number of different annotation schemes. This makes it very difficult to say something general about the relation between annotation schemes and linguistic theory, but broadly speaking I think we may distinguish three main kinds of annotation in current practice:

- Annotation of constituent structure

- Annotation of functional structure

- Theory-specific annotation

This is obviously not a proper taxonomy, since theory-specific annotation may concern both constituent structure and functional structure. Rather, the first two categories are meant to cover more or less theory-neutral annotation schemes, focusing on constituent structure or functional structure, respectively. It should also be pointed out immediately that the annotation found in many if not most of the existing treebanks actually combines two or even all three of these categories. Still, I believe that the categories may be useful in discussing existing annotation schemes and their relation to linguistic theory. I will treat the categories in the order in which they are listed above, which I think roughly corresponds to the historical development of treebank annotation schemes.

The annotation of *constituent structure*, often referred to as *bracketing*, is the main kind of annotation found in pioneering projects such as the Lancaster Parsed Corpus (Garside et al. [9]) and the original Penn Treebank (Marcus et al. [18]). Normally, this kind of annotation consists of part-of-speech tagging for individual word tokens and annotation of major phrase structure categories such as NP, VP, etc. Figure 5 shows a representative example, taken from the IBM Paris Treebank using a variant of the Lancaster annotation scheme.

```
[N Vous_PPSA5MS N]
[V accedez_VINIP5
    [P a_PREPA
        [N cette_DDEMFS session_NCOFS N]
    P]
    [Pv a_PREP31 partir_PREP32 de_PREP33
        [N la_DARDFS fenetre_NCOFS
            [A Gestionnaire_AJQFS
                [P de_PREPD
                    [N taches_NCOFP N]
                P]
            A]
        N]
    Pv]
V]
```

Figure 5: Constituency annotation in the IBM Paris Treebank

Annotation schemes of this kind are usually intended to be theory-neutral and therefore try to use mostly uncontroversial categories that are recognized in all or most syntactic theories that assume

4

some notion of constituent structure. Moreover, the structures produced tend to be rather flat, since intermediate phrase level categories are usually avoided, as well as complex structures such as Chomsky adjunction. The drawback of this is that the number of distinct expansions of the same phrase category can become very high. For example, Charniak [4] was able to extract 10,605 distinct context-free rules from a 300,000 word sample of the Penn Treebank. Of these, only 3943 occurred more than once in the sample.

The status of grammatical functions and their relation to constituent structure has long been a controversial issue in linguistic theory. Thus, whereas the standard view in transformational syntax since Chomsky [5] has been that grammatical functions are derivable from constituent structure, proponents of dependency syntax such as Mel'čuk [21] have argued that functional structure is more fundamental than constituent structure. Other theories, such as LFG, steer a middle course by assuming both notions as primitive.

When it comes to treebank annotation, the annotation of *functional structure* has become increasingly important in recent years. The most radical examples are perhaps the annotation schemes based on dependency syntax, exemplified by the Prague Dependency Treebank of Czech (Hajic [10]) and the METU Treebank of Turkish (Oflazer et al. [22]), where the annotation of dependency structure is added directly on top of the morphological annotation without any layer of constituent structure. Figure 6 shows a simple example of dependency annotation from the Prague Dependency Treebank.
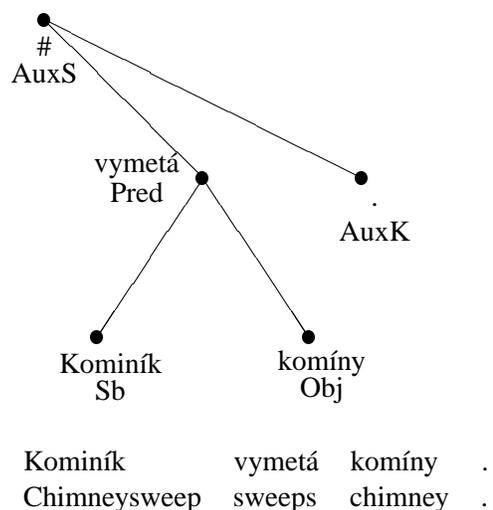

Figure 6: Functional annotation in the Prague Dependency Treebank

The trend towards more functionally oriented annotation schemes is also reflected in the extension of constituency-based schemes with annotation of grammatical functions. Cases in point are SUSANNE (Sampson [23]), which is a development of the Lancaster annotation scheme mentioned above, and Penn Treebank II (Marcus et al. [19]), which adds functional tags to the original phrase structure annotation. One of the most interesting examples in this respect is the annotation scheme adopted in the TIGER Treebank of German (Brants and Hansen [3]), developed from the earlier NEGRA treebank and annotation scheme, which integrates the annotation of constituency and dependency in a graph where node labels represent phrasal categories while edge labels represent syntactic functions.

The third kind of annotation scheme that is found in available treebanks is the kind that adheres to a specific linguistic theory and uses representations from that theory to annotate sentences. Thus, HPSG

has been used as the basis for treebanks of Bulgarian (Simov et al. [27]) and Polish (Marciniak et al. [20]), and the Prague Dependency Treebank mentioned earlier is based on the theory of Functional Generative Description (Sgall et al. [26]). There has also been work done on automatic f-structure annotation in the theoretical framework of LFG (see, e.g., Sadler et al. [24]).

In conclusion, we may perhaps say that there has been a trend towards more functionally oriented annotation schemes in recent years, and that theory-specific annotation schemes have become more common, but that it is probably still true to say that the dominant paradigm in treebank annotation is the kind of theory-neutral annotation of constituent structure with added functional tags represented by schemes such as the Penn Treebank II standard.

3   Annotation Schemes for Swedish

In this section, I will review four different annotation schemes for Swedish:

- MAMBA (Teleman [29])

- SynTag (Järborg [12])

- SWECG (Birn [2])

- S-CLE (Gambäck [7])

As stated in the introduction, MAMBA and SynTag are schemes designed for manual annotation of corpus material, while SWECG and S-CLE are primarily general purpose parsing systems that may be used for treebank annotation. The latter two schemes will therefore be treated more succinctly.

3.1   MAMBA

MAMBA is an annotation scheme for grammatical description of spoken and written Swedish described in Teleman [29]. It was developed at Lund University in the 1970s and used to annotate a corpus of about 250,000 words, containing both spoken and written material, which is still the largest corpus of Swedish annotated above the level of parts-of-speech. Since then it has been used in a number of studies dealing with written and spoken Swedish, and has almost acquired the status of a *de facto* standard, although most of the studies lie within the fields of quantitative stylistics and sociolinguistics, rather than grammatical theory as such.

Annotation of written texts or transcribed speech in MAMBA presupposes that the text has been segmented into *macrosyntagms* according to the principles described in Loman and Jörgensen [17]. The concept of *macrosyntagm*, inspired by Diderichsen's [6] *syntagm* and Hockett's [11] *macrosegment*, is defined as a sequence of words (produced by a single speaker/writer) that is maximal with respect to the syntactic relations (coordination, superordination and subordination) holding between its constituent units. Four types of macrosyntagms are recognized:

- Sentences

- Sentence fragments

- Interjectional macrosyntagms

- Vocative macrosyntagms

Interjectional and vocative macrosyntagms are especially relevant for the analysis of spoken language, where they normally occur in the pre- and post-fields of the extended main clause schema discussed in section 2.1. However, the grammatical annotation scheme of MAMBA only applies to sentential macrosyntagms (sentences and sentence fragment).

The annotation scheme consists of two layers, the first being a *lexical* analysis, consisting of part-of-speech information including morphological features, and the second being a *syntactic* analysis, in terms of grammatical functions. Both layers are flat in the sense that they consist of tags assigned to individual word tokens, but the syntactic layer also gives information about constituent structure, as exemplified in Figure 7, which shows the annotation of the sentence *Genom skattereformen införs individuell beskattning av arbetsinkomster.* (Through the tax reform, individual taxation of work income is introduced.)

```
*GENOM              PR        AAPR
SKATTEREFORMEN      NNDDSS    AA
INFÖRS              VVPSSMPA  FV
INDIVIDUELL         AJ        SSAT
BESKATTNING         VN        SS
AV                  PR        SSETPR
ARBETSINKOMSTER     NN    SS  SSET
.                   IP        IP
```

Figure 7: Two-layered annotation in MAMBA

The first column of annotation is the lexical analysis, while the second column is the syntactic analysis. The grammatical subject of the sentence is the phrase *individuell beskattning av arbetsinkomster* (individual taxation of work income), where the head word *beskattning* (taxation) is assigned the simple tag SS for subject, while the pre-modifying adjective *individuell* (individual) is tagged SS and AT for adjectival modifier; in the post-modifying prepositional phrase, the noun *arbetsinkomster* (work income) is tagged SS and ET for post-modifier, while the preposition *av* (of) is tagged SS, ET and PR for preposition.

According to Teleman [29], the syntactic analysis in MAMBA represents an eclectic approach based on dependency grammar, Diderichsen's field model and immediate constituent analysis. Thus, even though the explicit annotation does not involve any bracketing as described in section 2.2, it should be possible to translate the annotation into a phrase structure analysis. Figure 8 illustrates the mapping from tree structures to flat annotation using an example from Teleman [29].

On the other hand, the phrase structure that is recognized in the analysis is generally very flat and to a large extent modeled after Diderichsen's clause schemas (cf. section 2.1). Thus, the main constituents recognized in the clause are the following:

- Verb (-V)

- Subject (-S)

- Object (-O)

- Predicative (-P)

- Adverbial (-A)

7

Macrosyntagm
Subject (SS)　　　Finite verb (FV)
Determiner (DT)　　Head word (–)
Attribute (AT)　　Head word (–)

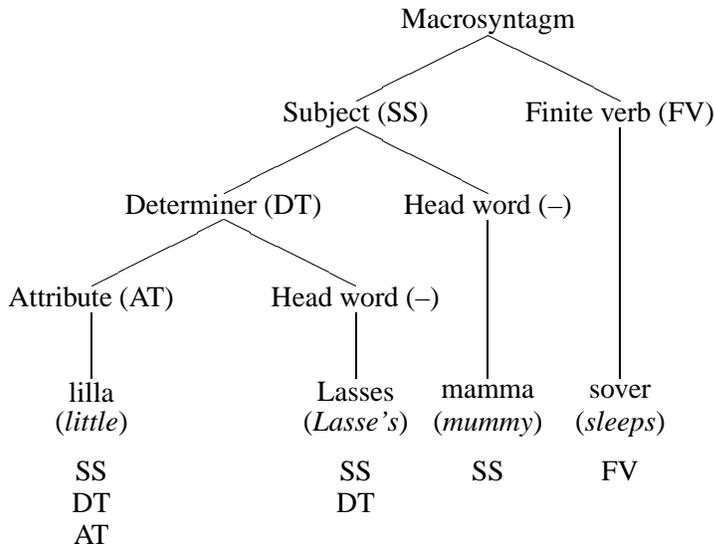| lilla | Lasses | mamma | sover |
(*little*)	(*Lasse's*)	(*mummy*)	(*sleeps*)
SS	SS	SS	FV
DT	DT		
AT			

Figure 8: Constituency and syntactic annotation in MAMBA

A more fine-grained classification of these constituents is obtained by varying the first letter of the two-letter tag. Thus, finite verbs are tagged FV, non-finite verbs IV; logical subjects are tagged ES, formal subjects FS, and other subjects SS, etc. In addition to the constituents recognized in Diderichsen's field model, there is a limited phrase structure analysis of noun phrases, prepositional phrases, adjective phrases, and subordinate clauses. It is worth noting that there is no phrase structure category corresponding to the verb phrase.

3.2 SynTag

SynTag is an annotation scheme developed at Göteborg University in the 1980s and described in Järborg [12]. It has been used to manually annotate a corpus of newspaper text, consisting of about 100,000 words. Like MAMBA, the SynTag scheme consists of two layers, of which the first is a part-of-speech tagging (without morphological information) and the second is an annotation of grammatical functions. The second layer is again hierarchical and encodes grammatical functions at different levels of subordination. Figure 9 shows the SynTag annotation of the sentence *Då bestämde hon sig för att bli skådespelerska i stället.* (Then she decided to become an actress instead.)

The first column contains the part-of-speech tagging, while the remaining columns give the syntactic analysis at different levels of subordination. The first four words — *då* (then), *bestämde* (decided), *hon* (she), *sig* (herself) — all realize major syntactic functions according to the standard field model for Swedish main clauses and are tagged as adverbial (ADV 0), finite verb (FIN 0), subject (SUB 0) and argument (ARG 1), respectively. The number 0 after the first three tags indicates that these functions are only realized once in the clause, while the number 1 after ARG signifies that it is the first of several instantiations of this syntactic function. The rest of the sentence is the phrase *för att bli skådespelerska i stället* (to become an actress instead), which is tagged as the second argument of the verb (ARG 2) in the first column. In the next column, the structure of this phrase is analyzed further with the preposition *för* being tagged as a relator (REL 0), *bli* (become) as an infinitive verb (INF 0), *skådespelerska* (actress) as a predicative (PRED 0) and the prepositional phrase *i stället* (instead) as

```
Då                  ab (ADV 0)
bestämde            vb (FIN 0)
hon                 pn (SUB 0)
sig                 pn (ARG 1)
för                 pp (ARG 2) (REL 0)
att                 ie (ARG 2)
bli                 vb (ARG 2) (INF 0)
skådespelerska      nn (ARG 2) (PRED 0)
i                   pp (ARG 2) (ADV 0) (REL 0)
stället             nn (ARG 2) (ADV 0) (H 0)
```

Figure 9: Two-layered annotation in SynTag

an adverbial (ADV 0). The internal structure of the final prepositional phrase is analyzed in the third column, where *stället* (lit. the place) is tagged as the head (H 0) and *i* (in) as a relator (REL 0).

According to Järborg [13], the syntactic analysis in SynTag is intended, as far as possible, to be theory-neutral and to avoid too far-reaching assumptions about the nature of syntactic structures. At the same time, the analysis clearly favors *relational* categories such as subject and object over *compositional* categories such as noun phrase and verb phrase. It is also surface-oriented in the sense that priority is given to indications from word order, agreement and lexical relation marking (function words, case inflection). The syntactic model recognizes three basic syntactic structures, two clause structures and one phrase structure, which are depicted in Figure 10. In addition to the syntactic functions occurring in the basic structures, the model recognizes special functions such as *relator* (for prepositions and subjunctions) and *conjunction*.

Sentence Structures			
Subject	Verb	Predicative	Adverbial*
Subject	Verb	Argument*	Adverbial*
Phrase Structures			
Descriptor*	Head	Argument*	

Figure 10: Basic syntactic structures in SynTag

On the whole, there are many similiarities between the annotation schemes SynTag and MAMBA, which both give priority to the annotation of functional structure over phrase structure. In both cases, the influence of Diderichsen's field model is also clearly discernible, notably in the absence of a (finite) verb phrase category and the treatment of both finite and nonfinite verb forms as primary constituents of the clause. The main difference between the two schemes is found in the categories used for syntactic classification, where SynTag uses fewer and more abstract categories, presumably in an attempt to be less theory-dependent. Moreover, the choice of category labels such as *argument* and *relator* indicates a semantically oriented view of syntactic functions.

3.3 SWECG

Swedish Constraint Grammar (SWECG) is a system for part-of-speech disambiguation and shallow syntactic analysis of running Swedish text, developed within the Constraint Grammar framework (Karlsson [15]) and described in Birn [2]. The output of the system is again an annotation in two layers, one layer of morphological tags (including part-of-speech information) and one layer of syntactic tags. Figure 11 shows the annotation of the sentence *Dessa entreprenöriella faktorer hade än så länge dämpat explosionen.* (These entrepreneurial factors had so far dampened the explosion.)

```
"<*dessa>"
     "denna" <**c> <DEM> <MD> DET UTR/NEU DEF PL NOM @DN>
"<entreprenöriella>"
     "entreprenöriella" A UTR/NEU DEF/INDEF PL NOM @AN>
"<faktorer>"
     "faktor" N UTR INDEF PL NOM @SUBJ
"<hade>"
     "ha" <AUX> V ACT PAST @+FCV
"<än_så_länge>"
     "än_så_länge" <COLLOCATION> ADV @ADVL
"<dämpat>"
     "dämpa" V ACT SUPINE @-FMV
"<explosionen>"
     "explosion" N UTR DEF SG NOM @OBJ
"<$.>"
      "$." CLB <PUNCT> @
```

Figure 11: Two-layer annotation in SWECG

The syntactic annotation is fairly similar to that of MAMBA and uses traditional grammatical function labels such as @SUBJ, @OBJ and @ADVL. However, it differs from both MAMBA and SynTag in having a completely flat structure with no explicit indication of hierarchical structure.

Tapanainen and Järvinen [28] has recently extended the Constraint Grammar formalism into a new formalism called Functional Dependency Grammar (FDG). There is now an FDG parser for Swedish available, which produces output similar to that of SWECG, but where the syntactic annotation also includes explicit relational information and in fact encodes a complete dependency tree. Figure 12 shows the FDG annotation of the sentence *Genom skattereformen införs individuell beskattning av arbetsinkomster.* (Through the tax reform, individual taxation of work income is introduced.)[3] The syntactic dependency annotation is found in the third column, with the second column containing a lemmatization and the fourth column being the morphological analysis.

3.4 S-CLE

The Swedish Core Language Engine (S-CLE) is a Swedish version of the Core Language Engine (Alshawi [1]) and is described in Gambäck and Rayner [7]. The system analyses Swedish sentences and converts them into a language-independent logical-form like representation using a unification-

---

[3]This is the same sentence that is annotated according to the MAMBA scheme in Figure 7.

```
Genom              genom              advl:>3    %AH PREP
skattereformen     skatte#reform      pcomp:>1   %NH N SG NOM
införs             införa             main:>0    %MV V PRES
individuell        individuell        attr:>5    %>N A SG NOM
beskattning        beskattning        subj:>3    %NH N SG NOM
av                 av                 mod:>5     %N< PREP
arbetsinkomster    arbets#inkomst     pcomp:>6   %NH N PL NOM
```

Figure 12: Three-layer annotation in FDG

based grammar described in Gambäck [8]. Figure 13 shows the output of the syntactic parser for the sentence *Bilen startar.* (The car starts.) Figure 14 shows the logical form for the same sentence.

```
[sigma_Declarative-1,
 [[s_np_vp_Normal-2,
   [[np_nbar_Definite-3,
      [[lex-3,[bilen]]]],
    [vp_v_Intransitive-5,
      [[lex-6,[startar]]]]
   ]]]]
```

Figure 13: Syntactic annotation in S-CLE

```
[dcl,
 form(l([bilen,startar]),verb(pres,no,no,no,y),Event,
      X^
      [X,
       [starta_2p,Event,
        term(l([bilen]),
             ref(def,bare,sing,l([])),_,
             Y^[bil1,Y],_,_)]],
      _)]
```

Figure 14: Semantic annotation in S-CLE

Preliminary work on constructing a Swedish treebank using S-CLE has been reported in Santamarta, Lindberg and Gambäck [25]. A major problem appears to be the limited coverage of S-CLE, which had a lexicon of some 1500 words at the time when the paper referred to was written. Although several techniques for bootstrapping the system are suggested in the paper, it is not clear to what extent they are feasible in practice. As far as I know, no further work has been carried out in this direction.

4 Discussion

Having reviewed four different annotation schemes for Swedish treebanks, I will now try to evaluate them along the following four dimensions:

11

- Descriptive adequacy and coverage with respect to written and spoken Swedish

- Relation to linguistic theory and annotation schemes for other languages

- Appropriateness for the needs of different treebank user groups

- Availability of resources for automatic annotation

4.1 Descriptive Adequacy and Coverage

If we start by considering the schemes designed for manual annotation, it is clear that MAMBA is well integrated in the Swedish tradition of descriptive grammar and has been used in corpus annotation on a large scale. It can therefore be expected to have good descriptive adequacy and coverage. Another point in favor of MAMBA is that it is the only scheme which has been tested on, and to some extent designed for, the analysis of spoken language.

The SynTag scheme has also been well tested in large scale corpus annotation and should at least have good coverage. The descriptive adequacy is harder to assess without a more in-depth analysis of the annotated resources, since the introduction of new and more abstract categories obscures the relation to the descriptive tradition. As pointed out by Järborg [13], this may be one of the reasons why the SynTag scheme has not enjoyed the same popularity as MAMBA.

If we turn to the parsing systems, SWECG is known to have good coverage and precision with respect to morphological analysis. With regard to the syntactic analysis, the descriptive adequacy is less well tested, but the large lexicon and robust parsing methodology of SWECG should at least guarantee good coverage. By contrast, it seems that S-CLE at present has too limited coverage to be practically useful for treebank annotation, which also means that it is difficult to assess its descriptive adequacy.

4.2 Theoretical Orientation and Universality

In section 2.2, I distinguished three main kinds of treebank annotation with respect to theoretical orientation:

- Annotation of constituent structure

- Annotation of functional structure

- Theory-specific annotation

It seems that the annotation schemes considered for Swedish fall mainly in the second category. Thus, the schemes of MAMBA, SynTag and SWECG all have in common that the syntactic analysis focuses on grammatical functions and syntactic dependencies, rather than phrase structure and bracketing. Moreover, none of these schemes can be described as theory-specific, unless we want to regard Constraint Grammar as a theory in itself. And both MAMBA and SynTag are intentionally designed to be theory-neutral, although they are both clearly influenced by dependency grammar and by Diderichsen's field model. The connection to dependency grammar is even stronger for SWECG, especially in the recent development of FDG.

The only exception from the functionally oriented trend in Swedish annotation schemes is S-CLE, which is firmly based in the tradition of generative grammar and influenced by theories such as GPSG and HPSG. As can be seen from the example in Figure 13, the syntactic analysis produced by S-CLE

is a traditional phrase structure analysis. The logical forms used in the semantic annotation belong to the theoretical tradition of formal semantics using underspecified representations.

As regards the relation to annotation schemes for other languages, we may first note that both SWECG and S-CLE have direct counterparts for the English language, namely ENGCG (Karlsson et al. [16]) and CLE (Alshawi [1]), although none of these systems have as far as I know been used in treebank annotation.

Of the schemes for manual annotation, MAMBA is clearly the most language-specific, although the actual information contained in the annotation is quite similar to what is found in mainstream annotation schemes such as Penn Treebank II. But whereas the latter can be described as a constituency-based scheme with added functional annotation, MAMBA is rather a functionally based scheme with added constituency annotation.

SynTag is potentially more universal than MAMBA, given its more abstract categories. However, the lack of connection to existing annotation schemes for other languages makes this universality more potential than actual.

4.3   User Requirements

Broadly speaking, we may distinguish at least three different areas of use for corpora in general and treebanks in particular:

- Linguistic research

- Language teaching

- Language technology

It goes without saying that the users in these areas have different requirements on annotation schemes. For example, theory-specific annotation may be required both in the research context and in language technology projects but is less likely to be important for language teaching. On the other hand, perspicuity of notation and ease of access is crucial in the latter context.

Among the annotation schemes considered for Swedish, MAMBA and SWECG have the advantage of using grammatical concepts that are familiar to a larger group of users, and would therefore seem to be preferable, *ceteris paribus*, to both SynTag and S-CLE. On the other hand, the kind of annotation (potentially) provided by S-CLE could be very useful in certain language technology applications.

4.4   Resources for Automatic Annotation

The availability of resources for automatic annotation is quite limited for Swedish. Although parsing systems exist for both SWECG and S-CLE, none of them is freely available. For S-CLE there is the added problem that the coverage of the parser is very limited.

For MAMBA and SynTag, there are currently no parsers available, but since annotated corpora exist for both schemes, it is at least conceivable that machine learning techniques could be used to construct software for automatic annotation.

5 Conclusion

In conclusion, MAMBA and SWECG emerge as the strongest candidates for use in the annotation of a Swedish treebank. The other two schemes considered, SynTag and S-CLE, are interesting in their own right but are on the whole less suitable for adoption in a large-scale treebank project.

MAMBA and SWECG have the advantage of being firmly based in the Swedish tradition of descriptive grammar and can therefore be expected to have good descriptive adequacy and coverage. This is true especially for MAMBA, which has been designed especially to handle spoken language as well as written language. Moreover, the fact that these schemes are based on notions of traditional grammar means that they provide an annotation which may be more accessible to non-expert treebank users.

The main weakness of SWECG is that the annotation contains little or no information about phrase structure and is therefore difficult to relate to many current linguistic theories. However, this situation has clearly improved with the development of FDG, which establishes a more direct connection to dependency-based theories of syntax and also provides a better basis for the reconstruction of phrase structure from dependency structure if this is required.

For MAMBA the biggest problem is instead the lack of resources for automatic annotation, although it may be possible to improve the situation by using the available annotated corpora for bootstrapping a parsing system.

References

[1] Alshawi, Hiyan (1992) *The Core Language Engine*. Cambridge, MA: MIT Press.

[2] Birn, Juhani (1998) Swedish Constraint Grammar. Lingsoft Inc. (URL: http://www.lingsoft.fi/doc/swecg/intro/).

[3] Brants, Sabine and Hansen, Silvia (2002) Developments in the TIGER Annotation Scheme and their Realization in the Corpus. In *Proceedings of the Third Conference on Language Resources and Evaluation (LREC 2002)*, pp. 1643–1649, Las Palmas.

[4] Charniak, Eugene (1996) Tree-Bank Grammars. In *AAAI/IAAI*, Vol. 2, pp. 1031–1036.

[5] Chomsky, Noam (1965) *Aspects of the Theory of Syntax*. MIT Press.

[6] Diderichsen, Paul (1946) *Elementær dansk grammatik*. Copenhagen: Gyldendal.

[7] Gambäck, Björn and Rayner, Manny (1992) The Swedish Core Language Engine. In *Papers from the 3rd Nordic Conference on Text Comprehension in Man and Machine*, Linköping University, Linköping, Sweden, pp. 71–85.

[8] Gambäck, Björn (1997) *Processing Swedish Sentences: A Unification-Based Grammar and Some Applications*. Royal Institute of Technology: Department of Computer and Systems Sciences.

[9] Garside, R., Leech, G. and Varadi, T. (compilers) (1992) *Lancaster Parsed Corpus*. A machine-readable syntactically-analysed corpus of 144,000 words, available for distribution through ICAME, The Norwegian Computing Centre for the Humanities, Bergen.

[10] Hajic, Jan (1998) Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In *Issues of Valency and Meaning*, pp. 106–132. Prague: Karolinum.

[11] Hockett, Charles F. (1958) *A Course in Modern Linguistics.* New York: MacMillan.

[12] Järborg, Jerker (1986) Manual för syntaggning [Manual for syntagging]. Göteborgs universitet: Institutionen för språkvetenskaplig databehandling.

[13] Järborg, Jerker (1990) Användning av SynTag [Use of SynTag]. Göteborgs universitet: Institutionen för språkvetenskaplig databehandling.

[14] Jörgensen, Nils and Svensson, Jan (1986) *Nusvensk grammatik* [Contemporary Swedish Grammar]. Malmö: Gleerups.

[15] Karlsson, Fred (1990) Constraint Grammar as a Framework for Parsing Running Text. In Karlgren, Hans (ed.) *Papers presented to the 13th International Conference on Computational Linguistics*, Vol. 3. Helsinki, pp. 168–173.

[16] Karlsson, Fred, Voutilainen, Atro, Heikkilä, Juha and Anttila, Arto (eds.) (1995) *Constraint Grammar: A Language-independent System for Parsing Unrestricted Text*. Berlin and New York: Mouton de Gruyter.

[17] Loman, Bengt and Jörgensen, Nils (1971) Manual för analys och beskrivning av makrosyntagmer [Manual for analysis and description of macrosyntagms]. Lund: Studentlitteratur.

[18] Marcus, Mitchell P., Santorini, Beatrice and Marcinkiewicz, Mary Ann (1993) Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19, 313–330. [Reprinted in Armstrong, Susan (ed.) (1994) *Using large corpora*, pp. 273–290. Cambridge, MA: MIT Press.]

[19] Marcus, Mitchell P., Kim, Grace, Marcinkiewicz, Mary Ann, MacIntyre, Robert, Bies, Ann, Ferguson, Mark, Katz, Karen and Schasberger, Britta (1994) The Penn Treebank: Annotating Predicate Argument Structure", In *ARPA Human Language Technology Workshop*.

[20] Marciniak, Malgorzata, Mykowiecka, Agniezska, Kupść, Anna and Przepiórkowski, Adam (2000) An HPSG-Annotated Test Suite for Polish. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*.

[21] Mel'čuk, Igor (1988) *Dependency Syntax: Theory and Practice*. State University of New York Press.

[22] Oflazer, Kemal, Say, Bilge and Hakkani Tur, Dilep (2000) A Syntactic Annotation Scheme for Turkish. In *Proceedings of 10th International Conference on Turkish Linguistics (ICTL-2000)*.

[23] Sampson, Geoffrey (1995) *English for the Computer*. Oxford University Press.

[24] Sadler, Louisa, von Genabith, Josef and Way, Andy (2000) Automatic F-Structure Annotation from the AP Treebank. In Butt, Miriam and Holloway King, Tracy (eds.) *Proceedings of the Fifth International Conference on Lexical-Functional Grammar*, The University of California at Berkeley, 19 July – 20 July 2000. Stanford, CA: CSLI Publications.

[25] Santamarta, Lena, Lindberg, Nikolaj and Gambäck, Björn (1995) Towards Building a Swedish Treebank. In *Proceedings of the 10th Nordic Conference of Computational Linguistics*, University of Helsinki, Helsinki, Finland, May 1995. Short Papers, pp. 37–40.

[26] Sgall, Petr, Hajicova, Eva and Panevova, Jarmila (1986) *The Meaning of the Sentence in Its Pragmatic Aspects*. Reidel.

[27] Simov, Kiril, Popova, Gergana, Osenova, Petya (forthcoming) HPSG-Based Syntactic Treebank of Bulgarian (BulTreeBank). In Wilson, Andrew, Rayson, Paul, McEnery, Tony (eds.) *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, pp. 135-142. Munich: Lincom-Europa.

[28] Tapanainen, Pasi and Järvinen, Timo (1997) A Non-Projective Dependency Parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*. Association for Computational Linguistics, Washington, D.C.

[29] Teleman, Ulf (1974) *Manual för grammatisk beskrivning av talad och skriven svenska [Manual for grammatical description of spoken and written Swedish]*. Lund: Studentlitteratur.

[30] Teleman, Ulf, Hellberg, Staffan and Andersson, Erik (1999) *Svenska Akademiens Grammatik [Swedish Academy Grammar]*. Stockholm: Svenska Akademien.