FUNCTIONAL FEATURES IN DATA-DRIVEN DEPENDENCY PARSING

Lilja Øvrelid
Potsdam University

Proceedings of the LFG08 Conference

Miriam Butt and Tracy Holloway King (Editors)

2008

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

This article relates results in data-driven dependency parsing of Swedish to linguistic generalizations regarding syntactic argument status, such as tendencies regarding animacy and definiteness, as well as properties more specific to the Scandinavian languages, such as finiteness. We show how data-driven modeling in combination with labeled dependency representations enable the acquisition of functional preferences that are evident as statistical tendencies in language data. We present an in-depth error analysis of a data-driven dependency parser with a particular focus on assignment of core syntactic arguments and show how a data-driven parser provides an experimental setting where the influence of various linguistic properties may be evaluated and investigated further.

1 Introduction

The separation of functional structure from constituent structure is motivated largely by cross-linguistic variation in degree of configurationality: languages differ in the extent to which grammatical functions may be equated with a specific structural position. F-structure constraints capture generalizations regarding grammatical functions regardless of their c-structure realization. In functional-typological Optimality Theory (Aissen, 2003; Bresnan and Aissen, 2002) constraints targeting grammatical functions have been centred around a notion of prominence and harmony, which have been shown to capture both categorical generalizations, as well as frequency effects observed in a range of languages (Bresnan et al., 2001). The idea that grammars are inherently probabilistic in nature has been motivated by empirical evidence observed as frequency effects in linguistic studies ranging from computational, psycholinguistic, typological to more theoretical (Bresnan, 2006; Manning, 2003). In computational linguistics, data-driven, statistical methods show impressive results for a range of NLP tasks, including syntactic parsing. There exists an expressed interest in a deeper understanding of the results obtained using data-driven methods and how these relate to generalizations from more theoretically oriented work.

Syntactic arguments express the main participants in an event, and hence are intimately linked to the semantics of a sentence. Syntactic arguments also occur in a specific discourse context where they convey linguistic information. For instance, the subject argument often expresses the agent of an action, and will therefore tend to refer to a human being. Moreover, subjects typically express the topic of the sentence and will tend to be realized by a definite nominal. These types of generalizations regarding the linguistic properties of syntactic arguments express soft constraints, rather than absolute requirements on syntactic structure. In language data, we observe frequency effects in the realization of syntactic arguments and a range of linguistic studies emphasize the correlation between syntactic function and various linguistic properties, such as animacy and definiteness.

The realization of a predicate-argument structure is furthermore subject to surface-oriented and often language-specific restrictions relating to word order and morphology. In many languages, the structural expression of syntactic arguments exhibits variation. The Scandinavian languages, for instance, are characterized by a rigid verb placement and a certain degree of variation in the positioning of syntactic arguments. Work in syntactic theory which separates the function-argument structure from its structural realization highlights exactly the mediating role of arguments between semantics and morphosyntax.

The use of distinguishing, linguistic properties of arguments, such as animacy, definiteness and finiteness, in automatic analysis of syntactic arguments has been shown to give improved results for Swedish (Øvrelid and Nivre, 2007; Øvrelid, 2008c). In this article, we relate these results, which were obtained using a data-driven dependency parser for Swedish to linguistic generalizations regarding argumenthood and the expression of syntactic arguments in Scandinavian. In particular, we propose that the use of dependency representations, which operate on a flat structure and a separate level of grammatical functions, allows for the acquisition of linguistic generalizations regarding syntactic argumenthood, irrespective of structural realization. A detailed error analysis is provided in order to pinpoint the effect of the various, linguistically motivated features during parsing. We investigate the relation of syntactic arguments to semantic interpretation, as well as to explicit, formal marking such as case and word order.

2 Arguments

A distinction between *arguments* and *non-arguments* is made in some form or other in all syntactic theories.[1] The distinction can be expressed through structural asymmetry or stipulated for theories where grammatical functions are primitives in representation. For instance, in LFG (Kaplan and Bresnan, 1982; Bresnan, 2001), grammatical functions are primitive concepts and arguments or governable functions (SUBJ, OBJ, OBJ_θ, OBL_θ, COMP, XCOMP) are distinguished from non-arguments or modifiers (ADJ, XADJ). HPSG (Pollard and Sag, 1994) similarly distinguishes the valency features (SPR, COMPS) from modifiers (MOD). In most versions of dependency grammar, (see, e.g, Mel'čuk, 1988), grammatical functions are also primitive notions and not derived through structural position.

2.1 Argument differentiation

Syntactic arguments may be distinguished by a range of linguistic factors related to structural, semantic as well as more discourse-related properties.

The dimension of **animacy** roughly distinguishes between entities which are alive and entities which are not; however, other distinctions are also relevant and

[1] We adopt the more theory-neutral term of 'non-argument', rather than 'adjunct', which is closely connected to the structural operation of adjunction.

the animacy dimension is often viewed as a continuum. Animacy is a grammatical factor in a range of languages and is closely related to argument realization and differentiation. A recent special issue of the linguistic journal *Lingua* was dedicated to the topic of animacy and discusses the role of animacy in natural language from rather different perspectives, ranging from theoretical and typological to experimental studies (de Swart et al., 2008). These various perspectives all highlight animacy as an influencing factor in argument differentiation. For instance, in the Mayan language MamMaya, a transitive sentence is ungrammatical if the object is higher in animacy than the subject, as in *The dog sees the woman* (de Swart et al., 2008). In Navajo, such a construction is clearly avoided and an alternative construction (*The woman is seen by the dog*) is chosen instead.[2] In many languages this tendency is reflected in language data as a frequency effect, even though these types of transitive constructions are perfectly grammatical (Dahl and Fraurud, 1996).

The property of **definiteness** is not as commonly recognized as a factor in argument differentiation as animacy. A tendency towards definite subjects has, however, been noted for several languages, both as a categorical constraint influencing morphological marking and as a statistical tendency. Common to these is the same generalization, namely a tendency for subjects to be definite or specific and for objects to be indefinite. In Turkish and Persian, we find Differential Object Marking which is sensitive to definiteness and where definite objects are marked with accusative case, but indefinite objects are not (Croft, 2003). A range of languages have been noted to categorically exclude or strongly disprefer non-specific indefinite subjects (Aissen, 2003).

(1) pronoun $>$ proper name $>$ common noun

This sense of referentiality, then, relates to the extent to which semantic interpretation requires access to the context of the utterance. This is related to the expression of definiteness, or level of cognitive status. Pronouns have to be resolved by the context, proper nouns rely on a conventional mapping to a referent, whereas the interpretation of common nouns relies the least on context and more on denotation.

Syntactic arguments differ with respect to their referentiality. As mentioned earlier the definiteness or cognitive status of an element influences its referentiality. In particular, subjects are likely to be pronominal and objects are more likely to express a lower referentiality (Keenan, 1976). The category of pronouns may be further subdivided along the dimension of *person* which distinguishes reference to the speaker and hearer (i.e. discourse participants) from others (Croft, 2003, 130).

[2]The *inverse* construction in Navajo can be paraphrased by the English passive construction and is expressed by the verbal affix *bi* and employed when the subject is lower in animacy than the object (Dahl and Fraurud, 1996).

2.2 Arguments in Scandinavian

Scandinavian languages have limited morphological marking of syntactic functions, but allow for variation in word order which makes for an interesting comparison with more configurational languages, like English.

2.2.1 Morphology

The Scandinavian languages make limited use of case marking, and, in this respect, resemble English. Pronouns are marked for case, but exhibit syncretism and syntactic variation, whereas nouns distinguish only genitive case and are otherwise invariant for case. The distinction between various types of arguments is, however, partially encoded through *case* marking in Scandinavian. Nominal arguments are furthermore inflected for other categories, such as definiteness.

2.2.2 Word order

The classical descriptive model for Scandinavian word order is based around organization into so-called *topological fields* . The topological fields approach separates the clause into, roughly speaking, three parts: the *initial* field, the *mid field* and the *end field*:

(2)

	Initial	**Mid**	**End**
MAIN	*I morgon* tomorrow	*kan hon inte* can she not	*vara med vid sammanträdet.* be with at meeting-DEF
SUBORD	*eftersom* since	*hon inte kan* she not can	*vara med vid sammanträdet.* be with at meeting-DEF

Initial variation The initial position is characterized by a great deal of variation. It has been claimed to mark the syntactic-semantic type of the clause and is closely related to the speech act expressed by the clause (Platzack, 1987). Moreover, the initial constituent is often topical, in the sense that it links the sentence to the preceding context. Most clausal constituents may occupy initial position in declarative main clauses, e.g., subjects (3), direct objects (4) and adverbials (5).[3]

(3) *Statsministern* *håller* *talet* *i* *morgon.*
prime minister-DEF holds speech-DEF in tomorrow
'The prime minister gives the speech tomorrow.'

(4) *Talet* *håller* *statsministern* *i* *morgon.*
speech-DEF holds prime minister-DEF in tomorrow
'The speech, the prime minister gives tomorrow.'

[3]The examples in the current section (section 2) are constructed. All other examples in the article are authentic and taken from the Talbanken05 treebank of Swedish, see section 3.1.

(5) *I morgon håller statsministern talet.*
in tomorrow holds prime minister-DEF speech-DEF
'Tomorrow, the prime minister gives the speech.'

Rigid verb placement Like the majority of Germanic languages, but unlike English, the Scandinavian languages are *verb second (V2)*; the finite verb is the second constituent in declarative main clauses, see (3)–(5) above. Non-finite verbs follow the finite verb, but precede their complements.[4] The presence of a non-finite verb introduces a greater rigidity in terms of positioning and interpretation of the clausal constituents. Main clauses consisting of a finite, transitive verb along with its arguments are structurally ambiguous, as in (6), whereas the placement of a non-finite verb in the same clause clearly indicates syntactic functions, as in (7)–(8):

(6) *Vem såg Ida?*
who saw Ida
'Who saw Ida / Who did Ida see?'

(7) *Vem har sett Ida?*
who has seen Ida
SUBJ OBJ
'Who has seen Ida?'

(8) *Vem har Ida sett?*
who has Ida seen
OBJ SUBJ
'Who has Ida seen?'

Variable argument placement The generalization that most constituents may occupy sentence-initial position entails that arguments have two alternative positions – initial position and a non-initial position. A schematized version of the predictions of the fields analysis with respect to the linearization of verbs and (non-initial) arguments in main clauses is provided in (9) below (Engdahl et al., 2004):

(9) Linearization of grammatical functions in declarative, main clauses:
XP | V_{fin} SUBJ S-ADV | $V_{non-fin}$ OBJ_{ind} OBJ_{dir} ADV

The subject, for instance, may occupy either the initial position or the position immediately following the verb. Note that the fields analysis does not capture the generalization that the subject is the most common initial constituent.

In recent years, proposals have been made for a considerably flatter c-structure representation for Scandinavian, due partly to the variation described above (Börjars et al., 2003; Engdahl et al., 2004; Andréasson, 2007). In these proposals, the ordering of arguments is rather determined by OT-like constraints expressing the interaction of various structural, semantic and pragmatic generalizations.

[4] In this respect Scandinavian differs from German, which positions non-finite verbs in clause final position.

3 Data-driven dependency parsing

A distinction is often made between grammar-driven and data-driven approaches to parsing, where the former is characterized by a generative grammar which defines the language under analysis and the latter is not. This distinction has, however, become less clear-cut due to the extensive use of empirical methods in the field in recent years. Most current parsers are data-driven in the sense that they employ frequencies from language data to induce information to improve parsing. Data-driven parsing may thus be characterized, first and foremost, by the use of inductive inference, rather than by the use or dispension of a grammar in the traditional sense (Nivre, 2006).

The availability of treebanks has been crucial to the development of data-driven parsing, supplying data for inductive inference in terms of estimation of parameters for statistical parse models or even for the induction of whole grammars (Charniak, 1996; Cahill et al., 2008). A system for data-driven parsing of a language L may be defined by three components (Nivre, 2006, 27):

1. A formal model M defining permissible analyses for sentences in L.
2. A sample of text $T_t = (x_1, \ldots, x_n)$ from L, with or without the correct analyses $A_t = (y_1, \ldots, y_n)$.
3. An inductive inference scheme I defining actual analyses for the sentences of any text $T = (x_1, \ldots, x_n)$ in L, relative to M and T_t (and possibly A_t).

In strictly data-driven approaches, a grammar, whether hand-crafted or induced, does not figure at all. It follows that the formal model M is not a grammar and the sample of text T_t is a treebank containing the correct analyses with respect to M, which constitutes the training data for the inductive inference scheme I. Parsing in this respect does not rely on a definition of the language under analysis independently of the input data. Without a formal grammar, data-driven models condition on a rich context in the search for the most probable analysis.

The use of dependency representations in syntactic parsing has recently received extensive attention in the NLP community (Buchholz and Marsi, 2006; Nivre et al., 2007). One of the arguments in favour of parsing with dependency representations is that dependency relations are much closer to the semantic relations which figure between words in a sentence than a tree is. As automatic parsing often is viewed as a means to a semantic interpretation of a sentence, dependency analysis represents a step in the right direction.

Common to all dependency-based grammar theories is the notion of *dependency* – a binary, asymmetrical relation between lexical items or words. Each word in a sentence has a head or governor of which it is a dependent (Mel'čuk, 1988). The dependency relation which holds between two words may or may not be labeled and its participants, the head and dependent, may or may not be ordered. Many of the theoretical proposals of dependency grammar separate dependency

structure from word order (Mel'čuk, 1988). Figure 1 shows the labeled dependency graph of example (10), taken from the Swedish treebank, Talbanken05, described in section 3.1 below.

(10) *Därefter betalar patienten avgift med 10 kronor om dagen.*
thereafter pays patient-DEF fee with 10 kronas in day-DEF
'Thereafter, the patient pays a fee of 10 kronas a day.'

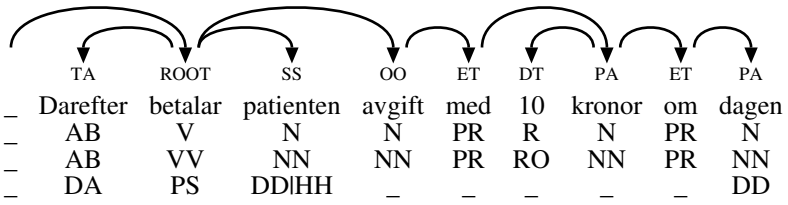

Figure 1: Dependency representation of example from Talbanken05.

3.1 Parsing Swedish

In the remaining sections we will present experiments in data-driven dependency parsing of Swedish. The focus will be on the analysis of syntactic arguments and, in particular, on argument differentiation: the process by which functional arguments are distinguished along one or more linguistic dimensions. In a data-driven parser, parsing is by definition guided by frequencies in language and there is no explicit grammar. This allows us to make as few assumptions as possible with respect to formulations of constraints on arguments, as well as their interaction. Due to the variation identified above, we do not want to commit to a strictly structural definition of argument status. Rather, a view of grammatical functions as primitive notions, separated from surface linguistic properties, enables investigations also into mismatches between levels of linguistic analysis.

Talbanken05 is a Swedish treebank converted to dependency format, containing both written and spoken language (Nivre et al., 2006a).[5] For each token, Talbanken05 contains information on word form, part of speech, head and dependency relation, as well as various morphosyntactic and/or lexical semantic features. The nature of this additional information varies depending on part of speech:

NOUN:	*definiteness, animacy, case* (Ø/GEN)
PRO:	*animacy, pronoun type, case* (Ø/ACC)
VERB:	*tense, voice* (Ø/PA)

We use the freely available **MaltParser**,[6] which is a language-independent system for data-driven dependency parsing. MaltParser is based on a deterministic

[5]The written sections of the treebank consist of professional prose and student essays and amount to 197,123 running tokens, spread over 11,431 sentences.

[6]http://w3.msi.vxu.se/users/nivre/research/MaltParser.html

parsing strategy, first proposed by Nivre (2003), in combination with treebank-induced classifiers for predicting the next parsing action. Classifiers can be trained using any machine learning approach, but the best results so far have been obtained with support vector machines, using LIBSVM (Chang and Lin, 2001). MaltParser has a wide range of parameters that need to be optimized when parsing a new language. As our baseline, we use the settings optimized for Swedish in the CoNLL-X shared task (Nivre et al., 2006b), where this parser was the best performing parser for Swedish. The only parameter that will be varied in the later experiments is the feature model used for the prediction of the next parsing action. We will therefore describe the feature model in a little more detail.

MaltParser uses two main data structures, a stack (S) and an input queue (I), and builds a dependency graph (G) incrementally in a single left-to-right pass over the input. The decision that needs to be made at any point during this derivation is (a) whether to add a dependency arc (with some label) between the token on top of the stack (*top*) and the next token in the input queue (*next*), and (b) whether to pop *top* from the stack or push *next* onto the stack. The features fed to the classifier for making these decisions naturally focus on attributes of *top*, *next* and neighbouring tokens in S, I or G. In the baseline feature model, these attributes are limited to the word form (FORM), part of speech (POS), and dependency relation (DEP) of a given token, but in later experiments we will add other linguistic features (FEATS). The baseline feature model is depicted as a matrix in Table 1, where rows denote tokens in the parser configuration (defined relative to S, I and G) and columns denote attributes. Each cell containing a plus sign (+) corresponds to a feature of the model. Examples of the features include part-of-speech for the top of the stack, lexical form for the next and previous (*next-1*) input tokens and the dependency relation of the rightmost sibling of the leftmost dependent of *top*.

	FORM	POS	DEP	FEATS
S:*top*	+	+	+	+
S:*top*+1		+		
I:*next*	+	+		+
I:*next*−1	+			+
I:*next*+1	+	+		+
I:*next*+2		+		
G: head of *top*	+			+
G: left dep of *top*			+	
G: right dep of *top*			+	
G: left dep of *next*	+		+	+
G: left dep of head of *top*			+	
G: left sibling of right dep of *top*			+	
G: right sibling of left dep of *top*	+			+
G: right sibling of left dep of *next*		+	+	

Table 1: Baseline and extended (FEATS) feature model for Swedish; S: stack, I: input, G: graph; $\pm n$ = n positions to the left ($-$) or right ($+$)

4 Error analysis of baseline

An error analysis is crucial for obtaining a better understanding of the types of generalizations regarding syntactic argumenthood that are being acquired by our data-driven parser. The data for the error analysis of argument assignment in Swedish was obtained by parsing the written part of Talbanken05 with MaltParser. We employed the settings optimized for Swedish in the CoNLL-X shared task (Nivre et al., 2006b), with the feature model presented in the first three columns in Table 1. As we can see, the features employed during parsing are part-of-speech (POS), lexical form (FORM) and structural properties of the dependency graph under construction (DEP).

Table 2 provides an overview of the parser performance for the various argument relations in the treebank[7] It is quite clear that there is a direct relation between the frequency of the dependency relation in the treebank and the parser performance. The most frequent relations are also the relations for which the parser performs best – subject SS (90.25), predicative SP (84.82), and object OO (84.53).

Deprel		Gold	Correct	System	Recall	Precision	F-score
SS	subject	19383	17444	19274	90.00	90.51	90.25
SP	subject predicative	5217	4416	5196	84.65	84.99	84.82
OO	direct object	11089	9639	11718	86.92	82.26	84.53
IO	indirect object	424	276	301	65.09	91.69	76.14
AG	passive agent	334	249	343	74.55	72.59	73.56
VO	object infinitive	121	84	112	69.42	75.00	72.10
ES	logical subject	878	562	687	64.01	81.80	71.82
FS	formal subject	884	578	737	65.38	78.43	71.31
VS	subject infinitive	102	47	58	46.08	81.03	58.75
FO	formal object	156	70	91	44.87	76.92	56.68
OP	object predicative	189	42	112	22.22	37.50	27.91
EO	logical object	22	2	3	9.09	66.67	16.00

Table 2: Dependency relation performance: total number of gold instances (Gold), system correct (Correct), system proposed (System), recall, precision and F-score

Table 3 shows the most frequent error types involving argument relations. We find frequent error types involving different kinds of subjects (SS, FS , ES), objects (OO, IO) and predicatives (SP). We find that the two most frequent error types involving argument relations are errors analyzing subjects as objects (SS_OO) and vice versa (OO_SS).

In addition to the confusion of subjects and objects, which constitutes the most common error type for both relations, we find that both subjects and objects are quite commonly assigned status as the root of the dependency graph (ROOT). For both argument relations we also observe error types indicating confusion with other argument relations. For subjects we observe confusion with the other main argu-

[7]These are evaluated by the standard class-based evaluation measures of precision, recall and a balanced F-score: 2PR/P+R (P=precision: true positives / true positives + false positives, R=recall: true positives / true positives + false negatives)

Gold	System	#
SS	OO	446
OO	SS	309
FS	SS	281
SS	ROOT	265
SP	SS	240
SS	DT	238
OO	ROOT	221
SS	SP	206
DT	SS	146
SS	CC	137

Table 3: 10 overall most frequent argument error types.

ment functions, such as subject predicatives (SP) and expletive subjects (FS), as well as confusion with determiners (DT) and prepositional complements (PA). For objects we observe primarily confusion with various oblique, adverbial relations (AA, ET, OA), as well as confusion with prepositional complements (PA) and determiners (DT).

There are various sources of errors in subject/object assignment. Common to all of them is that the parts of speech that realize subjects and objects are compatible with a range of dependency relations. Pronouns, for instance, may function as subjects, objects, determiners, predicatives, conjuncts, prepositional objects, etc. In addition, we find "traditional" attachment ambiguity errors, for instance in connection with coordination, subordination, particle verbs, etc. These represent notorious phenomena in parsing, and are by no means particular to Swedish. Scandinavian type languages, however, also exhibit ambiguities in morphology and word order which complicate the picture further. The confusion of subjects and objects follows from lack of sufficient formal disambiguation, i.e., simple clues such as word order, part-of-speech and word form do not clearly indicate syntactic function. The reason for this can be found in ambiguities on several levels.

With respect to word order, we have seen that subjects and objects may both precede or follow their verbal head, but these realizations are not equally likely. Subjects are more likely to occur preverbally, whereas objects typically occupy a postverbal position. Based only on the word order preferences discussed above, we would expect postverbal subjects and preverbal objects to be more dominant among the errors than in the treebank as a whole (23% and 6% respectively), since they display word order variants that depart from the canonical, and hence most frequent, ordering of arguments. This is precisely what we find. Table 4 shows a breakdown of the errors for confused subjects and objects and their position with respect to the verbal head.

We find that postverbal subjects (After) are in a clear majority among the subjects erroneously assigned the object relation. Due to the V2 property of Swedish,

		Before		After		Total	
Gold	System	#	%	#	%	#	%
SS	OO	103	23.1	343	76.9	446	100.0
OO	SS	103	33.3	206	66.7	309	100.0

Table 4: Ordering relative to verb for the SS_OO and OO_SS error types.

the subject must reside in a position following the finite verb whenever another constituent occupies the preverbal position, as in (11) where a direct object resides sentence-initially or (12) where we find a sentence-initial adverbial:

(11) *Samma erfarenhet gjorde* **engelsmännen**.
same experience made Englishmen-DEF
'The same experience, the Englishmen had.'

(12) *År 1920, och först då, fick den gifta* **kvinnan**
year 1920, and first then, got the married woman-DEF
fullständig myndighet.
complete rights
'It was not until 1920 that the married woman recieved full civil rights.'

For the confused objects we find a larger proportion of preverbal elements than for subjects, which is the mirror image of the normal distribution of syntactic functions among preverbal elements. As table 4 shows, the proportion of preverbal elements among the subject-assigned objects (33.3%) is notably higher than in the corpus as a whole, where preverbal objects account for a miniscule 6% of all objects.

The preverbal objects are topicalized elements which precede their head verb as in (13)–(14).

(13) **Detta** *anser tydligen inte Stig Hellsten.*
this means apparently not Stig Hellsten
'This, Stig Hellsten apparently does not believe.'

(14) *Kärlekens innersta* **väsen** *lär inte något politiskt parti*
love-DEF.GEN inner nature seems not any political party
kunna påverka.
can-INF influence
'The inner nature of love, it seems that no political party can influence.'

Contrary to our initial hypothesis, however, we find a majority of postverbal objects among the objects confused for subjects. These objects are interpreted as subjects because the local preverbal context strongly indicates a subject analysis. This includes verb-initial clauses as in (15), as well as constructions where the immediate preverbal context consists of an adverbial and the subject is non-local, as in (16) and (17) below.

(15) *Glöm aldrig det* **löfte** *om trohet för livet.*
forget never that promise of faithfulness for life-DEF
'Never forget that promise of faithfulness for life.'

(16) *Ungdomarna blir med barn och det sociala trycket*
teenagers become with child and the social pressure-DEF
nästan tvingar **dem** *att gifta sig.*
almost forces them to marry themselves
'The teenagers become pregnant and social pressure almost forces them to get married.'

(17) *Eftersom man har full frihet att enkelt och snabbt ingå*
because one has full freedom to easily and quickly enter
äktenskap.
marriage
'Because one has the freedom to easily and quickly get married.'

The example in (16) is particularly interesting as it violates the V2-property, assumed to be a categorical constraint of Swedish. We may note that the examples in (15)–(17) above indicate acquisition of argument ordering resulting from the V2 requirement; when there is no preverbal argument or when the preverbal argument is not a good subject candidate, the argument following the verb is analyzed as subject. Recall, however, that the parser does not have information on tense or finiteness, and hence it overgeneralizes for examples like (17), where the verb is non-finite.

In addition to the word order variation discussed above, Swedish also has limited morphological marking of syntactic function. Recall that nouns are only marked for genitive case and only pronouns are marked for accusative case. There is also syncretism in the pronominal paradigm. There are pronouns which are invariant for case, e.g. *det/den* 'it', *ingen/inga* 'no', and furthermore may function as determiners. This means that with respect to word form, only the set of unambiguous pronouns clearly indicates syntactic function. We may predict that subject/object confusion errors frequently involve elements whose syntactic category and/or lexical form does not disambiguate, i.e., nouns or ambiguous pronouns. Table 5 shows the distribution of nouns, functionally ambiguous and unambiguous pronouns and other parts of speech for confused subjects/objects.[8] Indeed, we find that nouns and functionally ambiguous pronouns dominate the errors where subjects and objects are confused. Since case information is not explicitly represented in the input, this indicates that case is acquired quite reliably through lexical form. The fact that we find a higher proportion of ambiguous pronouns among the objects erroneously assigned subject status indicates that the parser has acquired a

[8]The 'other' category consists mainly of verbs (heads of subordinate clauses), adjectives, participles and numerals functioning as nominal heads.

Gold	System	Noun		Pro_{amb}		Pro_{unamb}		Other		Total	
SS	OO	324	72.6%	53	11.9%	29	6.5%	40	9.0%	446	100%
OO	SS	215	69.6%	74	23.9%	9	2.9%	11	3.6%	309	100%

Table 5: Part of speech for the SS_OO and OO_SS error types – nouns, ambiguous pronouns, unambiguous pronouns and other parts of speech.

preference for subject assignment to pronouns compatible with the difference in frequency for pronominal realization (SS_{pro} 49.2%, OO_{pro} 10.1%).

The initial error analysis shows that the confusion of different types of argument relations, in particular subjects and objects, constitutes a frequent and consistent error during parsing. It is caused by ambiguities in word order and morphological marking and we find cases that deviate from the most frequent word order patterns and are not formally disambiguated by part-of-speech information. In order to resolve these ambiguities, we have to examine features beyond part-of-speech category and linear word order.

5 Parse experiments

In the following we will experiment with the addition of morphosyntactic and lexical semantic features that approximate the distinguishing properties of the core argument functions discussed earlier. We will isolate features of the arguments and the verbal head, as well as combinations of these, and evaluate their effect on overall parsing results as well as on subject/object disambiguation specifically.

5.1 Linguistic features for argument disambiguation

Argument relations tend to differ along several linguistic dimensions. These differences are found as statistical tendencies, rather than absolute requirements on syntactic structure, and are therefore highly suitable for data-driven modeling.

In table 6 we find an overview of the linguistic dimensions discussed above with their corresponding treebank feature. It distinguishes between the features discussed earlier, representing soft, cross-linguistic tendencies in argument differentiation, and the more language-specific features of Scandinavian discussed in section 2. We map the linguistic features to a set of empirical features representing information which is found in the annotation of the Talbanken05 treebank (see section 3.1 above).

Recall that the Talbanken05 treebank explicitly distinguishes between person- and non-person referring nominal elements, a distinction which overlaps fairly well with the traditional notion of animacy.[9] Morphological definiteness is marked for

[9]See Øvrelid (2008a) for a more detailed overview of the information on person reference in

Linguistic feature	Treebank feature
animacy	person reference
definiteness	morphological definiteness
referentiality	pronoun type, part-of-speech
finiteness	tense
case	morphological case

Table 6: Linguistic features and their empirical counterparts.

all common nouns in Talbanken05 and the treebank also contains morphological case annotation for pronouns, distinguishing nominative and accusative case, as well as genitive case for common nouns. The morphosyntactic features which are expressed for the part-of-speech of verbs in Talbanken are tense (present, past, imperative, past/present subjunctive, infinitive and supine) and voice (Ø/passive; PA).

Pronouns are furthermore annotated with a set of pronominal classes which distinguish between e.g. 1st/2nd person and 3rd person pronouns, reflexive, reciprocal, interrogative, impersonal pronouns, etc. For the third person neuter pronoun *det* 'it' and demonstrative *detta* 'this', the annotation in Talbanken05 distinguishes between an impersonal and a personal or "definite" (DP) usage. The impersonal pronominal class is employed for *non-referential* pronouns.[10] The two classes of pronouns have quite distinct syntactic behaviours. The impersonal pronouns never function as determiners (DT), whereas the definite pronouns often do (71.4%). Also, the impersonal pronouns are more likely to function as formal subjects FS (32.4%) than the definite pronoun (1.1%).

5.2 Experimental methodology

All parsing experiments are performed using 10-fold cross-validation for training and testing on the entire written part of Talbanken05. The feature model used throughout is the extended feature model depicted in Table 1, including all four columns. What is varied in the experiments is only the information contained in the FEATS features (animacy, definiteness, etc.), while the tokens for which these features are defined remains constant. Overall parsing accuracy will be reported using the standard metrics of *labeled attachment score* (LAS) and *unlabeled attachment score* (UAS), i.e. the percentage of tokens that are assigned the correct head *with* (labeled) or *without* (unlabeled) the correct dependency label. Statistical significance is checked using Dan Bikel's randomized parsing evaluation compara-

Talbanken05.

[10]Note that we here employ 'referential' in a narrow sense, which only includes reference to entities. The category of 'non-referential pronouns' consequently includes pronouns which do not refer, i.e., expletives, as well as pronouns which refer to propositions.

tor.[11] Since the main focus of this article is on the disambiguation of grammatical functions, we report accuracy for specific dependency relations, measured as a balanced F-score.

We perform a set of experiments with an extended feature model and added information on animacy, definiteness, case, finiteness and voice, where the features are employed individually as well as in combination.

5.3 Results

The overall results for these experiments are presented in table 7, along with p-scores indicating statistical significance of the difference compared to the baseline parser (NoFeats). The experiments show that each feature individually causes a significant improvement in terms of overall labeled accuracy as well as performance for argument relations. Error analysis comparing the baseline parser (NoFeats) with new parsers trained with individual features reveal the influence of these features on argument disambiguation.

	UAS	LAS	p-value
NoFeats	89.87	84.92	–
Anim	89.93	85.10	$p<.0002$
Def	89.87	85.02	$p<.02$
Pro	89.91	85.04	$p<.01$
Case	89.99	85.13	$p<.0001$
Verb	90.15	85.28	$p<.0001$
ADPC	90.13	85.35	$p<.0001$
ADPCV	90.40	85.68	$p<.0001$

Table 7: Overall results in gold standard experiments expressed as unlabeled and labeled attachment scores.

As Table 7 shows, the addition of information on animacy (Anim) for nominal elements causes an improvement in overall results ($p<.0002$). The subject and object functions are the dependency relations whose assignment improves the most when animacy information is added. There is also an effect for a range of other functions where animacy is not directly relevant, but where the improved analysis of arguments contributes towards correct identification (e.g., adverbials and determiners). If we take a closer look at the individual error types involving subjects and objects, we find that the addition causes a reduction of errors confusing subjects with objects (SS_OO), determiners (SS_DT) and subjects predicatives (SS_SP) – all functions which do not embody the same preference for animate reference as subjects.

The addition of information on definiteness (Def) during parsing causes a slight (at the $p<.03$ level) improvement of overall results. Most noteworthy is an improvement in the identification of subject predicatives (SP), which are often confused with subjects. Nominal predicatives in Swedish usually stand in a classifying

[11] http://www.cis.upenn.edu/∼dbikel/software.html

relation to their subject and are often realized by an indefinite noun (89.4%). If we examine the set of corrected errors compared to the baseline, we find that the added information causes a 14.2% reduction of the SP_SS errors, all of which are indefinite nouns.

The addition of pronoun type (Pro) information causes a general improvement in overall parsing results ($p<.01$), as we can see from Table 7. The dependency relations whose assignment improves the most are, once again, the core argument functions (SS, OO), as well as determiners (DT). We also find a general improvement in terms of recall for the assignment of the formal subject (FS) and object (FO) functions, which are both realized by the third person neuter pronoun *det* 'it', annotated as non-referential in the treebank.

When we employ case (Case) information during parsing we find a clear improvement in results ($p<.0001$). However, the improvement is not first and foremost caused by improvement in assignment of subjects and objects, but rather, the assignment of determiners and prepositional objects.

As Table 7 shows, the addition of morphosyntactic information for verbs (Verb) also causes a clear improvement in overall results ($p<.0001$). The added information has a positive effect on the verbal dependency relations – relations for finite (ROOT, MS) and non-finite verbs (VG, IV), as well as an overall effect on the assignment of subjects and objects. Information on voice also benefits the relation expressing the demoted agent (AG) in passive constructions. We experimented with the use of tense as well as finiteness, a binary feature which was obtained by a mapping from tense to a binary feature finite/non-finite. Finiteness gave significantly better results ($p<.03$) and was therefore employed in the following. See Øvrelid (2008b) for details.

		NoFeats	ADPCV
SS	subject	90.25	91.87
OO	object	84.53	86.38
SP	subj.pred.	84.82	86.10
FS	formal subj.	71.31	74.09
AG	pass. agent	73.56	79.75
ES	logical subj.	71.82	73.67
FO	formal obj.	56.68	67.65
VO	obj. small clause	72.10	84.72
VS	subj. small clause	58.75	65.56
IO	indir. obj.	76.14	77.09

Table 8: F-scores for argument relations with combined features (ADPCV).

The ADPCV experiment which combines information on animacy, definiteness, case and verbal features shows a cumulative effect of the added features with results which differ significantly from the baseline, as well as from each of the individual experiments ($p<.0001$). We observe clear improvements for the analysis of all argument relations, as shown by the third column in table 8 which presents F-scores for the various argument relations. In the error analysis of the baseline parser in section 4, we concluded that word order and morphology does not provide

sufficient information for argument disambiguation in all cases.

In tables 9 and 10 we examine word order and part-of-speech for the corrected SS_OO and OO_SS errors in the ADPCV experiment. We see that the added information contributes to the reduction of precisely the types of errors which were identified in the error analysis. In particular, improvement is centered in postverbal positions, largely occupied by nouns and case ambiguous pronouns.

		Before		After		Total	
Gold	System	#	%	#	%	#	%
SS	OO	21	10.6	178	89.4	199	100.0
OO	SS	15	10.6	127	89.4	142	100.0

Table 9: Order relative to verb for corrected SS_OO and OO_SS errors in the ADPCV experiment.

		Noun		Pro_{amb}		Pro_{unamb}		Other		Total	
Gold	System	#	%	#	%	#	%	#	%	#	%
SS	OO	144	72.4	23	11.6	18	9.0	14	7.0	199	100.0
OO	SS	111	78.2	21	14.8	6	4.2	4	2.8	142	100.0

Table 10: Part of speech for corrected SS_OO and OO_SS errors in the ADPCV experiment.

Figure 2 shows the total number of SS_OO and OO_SS errors in the various experiments and clearly illustrates the observed reduction for this error type with the chosen set of linguistic features. If we examine confusion matrices for the assignment of the subject and object relations, we find a reduction of total errors for the SS_OO and OO_SS error types with 34.3% and 30.4% respectively. With respect to the specific errors performed by the baseline parser, we observe a substantial reduction of 44.6% for SS_OO and 46.0% for OO_SS.

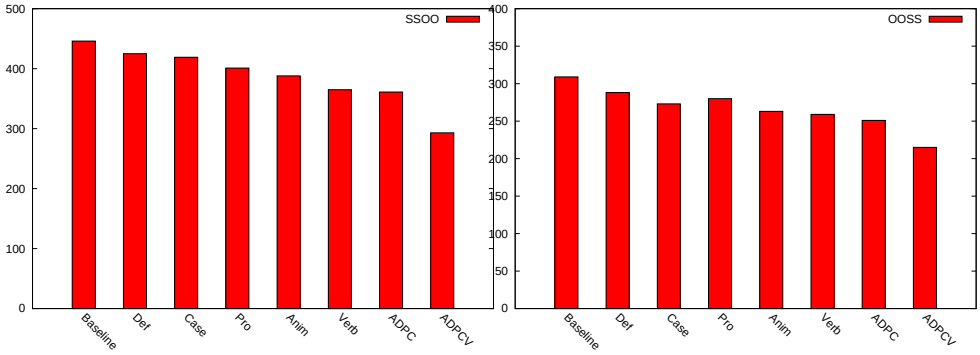

Figure 2: Total number of SS_OO errors (left) and OO_SS errors (right) in the various experiments.

6 Discussion and conclusion

An error analysis of a state-of-the-art data-driven dependency parser for Swedish revealed consistent errors in dependency assignment, namely the confusion of argument functions. The error analysis showed that further improvement of argument analysis was partly dependent on properties of argument realization other than word order and morphology. The separation of functional arguments from structural position which characterizes dependency analysis enabled the acquisition of functional generalizations irrespective of structural realization. For Scandinavian type languages, which are characterized by considerable word order variation and lack of morphological marking, the separation of function from structural realization constitutes an important property. Furthermore, the acquisition of soft, functional constraints is clear from the type of improvement which the added information incurred. We found improvement largely in labeled results caused by disambiguation of grammatical functions, rather than structural positions (attachment). For instance, for the errors confusing subjects for objects and vice versa, which were largely errors in labeling, we observed an error reduction of 44–46% of the baseline errors in the experiments combining all features. We found that a majority of the improved errors were arguments which were non-canonical in some sense, i.e., departing from the most frequent structural and morphological properties. Improvement thus relied on other properties of argument relations and the abstraction over specific realization in terms of dependency relations. The results are in line with recent proposals for a considerably flatter analysis of Scandinavian where ordering is determined by OT constraints (Engdahl et al., 2004; Andréasson, 2007).

We established a set of features expressing distinguishing semantic and structural properties of arguments such as animacy, definiteness and finiteness and performed a set of experiments with gold standard features taken from a treebank of Swedish. The experiments showed that each feature individually caused an improvement in terms of overall labeled accuracy and performance for the argument relations, in line with linguistic generalizations.

Properties of the Scandinavian languages connected with errors in argument assignment are not isolated phenomena. A range of other languages exhibit similar properties, for instance, Italian exhibits word order variation, little case, syncretism in agreement morphology, as well as pro-drop; German exhibits a larger degree of word order variation in combination with quite a bit of syncretism in case morphology; Dutch has word order variation, little case and syncretism in agreement morphology. These are all examples of other languages for which the results described here are relevant. Future work naturally extends to a multilingual setting, where similar experiments may be performed for these languages and the results may be evaluated and analyzed further.

References

Aissen, Judith. 2003. Differential Object Marking: Iconicity vs. Economy. *Natural Language and Linguistic Theory* 21(3), 435–483.

Andréasson, Maia. 2007. *Satsadverbial, Ledföljd och Informationsdynamik i Svenskan*. Göteborgsstudier i Nordisk Språkvetenskap, Göteborg University.

Börjars, Kersti, Engdahl, Elisabet and Andréasson, Maia. 2003. Subject and Object Positions in Swedish. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG03 Conference*, Stanford, CA: CSLI Publications.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Malden, Mass.: Blackwell Publishers.

Bresnan, Joan. 2006. Is Syntactic Knowledge Probabilistic? Experiments with the English Dative Alternation. In Sam Featherston and Wolfgang Sternefeld (eds.), *Roots: Linguistics in search of its evidential base*, Berlin: Mouton de Gruyter.

Bresnan, Joan and Aissen, Judith. 2002. Optimality and Functionality: Objections and Refutations. *Natural Language and Linguistic Theory* 20(1), 81–95.

Bresnan, Joan, Dingare, Shipra and Manning, Christopher D. 2001. Soft Constraints Mirror Hard Constraints: Voice and Person in English and Lummi. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG01 Conference*, Stanford, CA: CSLI Publications.

Buchholz, Sabine and Marsi, Erwin. 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164.

Cahill, Aoife, Burke, Michael, ODonovan, Ruth, Riezler, Stefan, van Genabith, Josef and Way, Andy. 2008. Wide-Coverage Deep Statistical Parsing Using Automatic Dependency Structure Annotation. *Computational Linguistics* 34(1), 81–124.

Chang, Chih-Chung and Lin, Chih-Jen. 2001. LIBSVM: A Library for Support Vector Machines, software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

Charniak, Eugene. 1996. Treebank Grammars. In *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI)*, pages 1031–1036.

Croft, William. 2003. *Typology and Universals*. Cambridge: Cambridge University Press, second edition.

Dahl, Östen and Fraurud, Kari. 1996. Animacy in Grammar and Discourse. In Thorstein Fretheim and Jeanette K. Gundel (eds.), *Reference and referent accessibility*, pages 47–65, Amsterdam: John Benjamins.

de Swart, Peter, Lamers, Monique and Lestrade, Sander. 2008. Animacy, Argument Structure and Argument Encoding: Introduction to the Special Issue on Animacy. *Lingua* 118(2), 131–140.

Engdahl, Elisabet, Andréasson, Maia and Börjars, Kersti. 2004. Word Order in the Swedish Midfield – an OT Approach. In Fred Karlsson (ed.), *Proceedings of the 20th Scandinavian Conference of Linguistics*.

Kaplan, Ronald M. and Bresnan, Joan. 1982. Lexical-Functional Grammar: A Formal System for Grammatical Representation. In Joan Bresnan (ed.), *The mental representation of grammatical relations*, pages 173–281, Cambridge, MA: MIT Press.

Keenan, Edward L. 1976. Towards a Universal Definition of "Subject". In Charles N. Li (ed.), *Subject and Topic*, pages 303–333, Cambridge, MA: Academic Press.

Manning, Christopher D. 2003. Probabilistic Syntax. In Rens Bod, Jennifer Hay and Stefanie Jannedy (eds.), *Probabilistic Linguistics*, pages 289–341, Cambridge, MA: MIT Press.

Mel'čuk, Igor. 1988. *Dependency Syntax: Theory and Practice*. Albany: State University of New York Press.

Nivre, Joakim. 2006. *Inductive Dependency Parsing*. Dordrecht: Springer.

Nivre, Joakim, Hall, Johan, Kübler, Sandra, McDonald, Ryan, Nilsson, Jens, Riedel, Sebastian and Yuret, Deniz. 2007. CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932.

Nivre, Joakim, Nilsson, Jens and Hall, Johan. 2006a. Talbanken05: A Swedish Treebank with Phrase structure and Dependency Annotation. In *Proceedings of the fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1392–1395.

Nivre, Joakim, Nilsson, Jens, Hall, Johan, Eryiğit, Gülşen and Marinov, Svetoslav. 2006b. Labeled Pseudo-Projective Dependency Parsing with Support Vector Machines. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*.

Øvrelid, Lilja. 2008a. *Argument Differentiation. Soft constraints and data-driven models*. Ph. D.thesis, University of Gothenburg.

Øvrelid, Lilja. 2008b. Finite Matters: Verbal Features in Data-Driven Parsing of Swedish. In Aarne Ranta & Bengt Nordström (ed.), *Proceedings of the International Conference on NLP, GoTAL 2008*, LNCS/LNAI Volume 5221, Springer.

Øvrelid, Lilja. 2008c. Linguistic features in data-driven dependency parsing. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL 2008)*.

Øvrelid, Lilja and Nivre, Joakim. 2007. When Word Order and Part-of-Speech Tags are not Enough – Swedish Dependency Parsing with Rich Linguistic Features. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 447–451.

Platzack, Christer. 1987. Huvudsatsordföljd och Bisatsordföljd. In Ulf Teleman (ed.), *Grammatik på villovägar*, pages 87–96, Solna: Esselte Studium.

Pollard, Carl and Sag, Ivan A. 1994. *Head-driven Phrase Structure Grammar*. Chicago, IL: University of Chicago Press.