A Corpus-based Syntactic Lexicon for Adverbs

Sanni Nimb

Center for Sprogteknologi (CST)
Njalsgade 80, Copenhagen, Denmark
sanni@cst.dk

Abstract

A word class often neglected in the field of NLP resources, namely adverbs, has lately been described in a computational lexicon produced at CST as one of the results of a Ph.D.-project. The adverb lexicon, which is integrated in the Danish STO lexicon, gives detailed syntactic information on the type of modification and position, as well as on other syntactic properties of approx 800 Danish adverbs. One of the aims of the lexicon has been to establish a clear distinction between syntactic and semantic information - where other lexicons often generalize over the syntactic behavior of semantic classes of adverbs, every adverb is described with respect to its proper syntactic behavior in a text corpus, revealing very individual syntactic properties. Syntactic information on adverbs is needed in NLP systems generating text to ensure correct placing in the phrase they modify. Also in systems analyzing text, this information is needed in order to attach the adverbs to the right node in the syntactic parse trees. Within the field of linguistic research, several results can be deduced from the lexicon, e.g. knowledge of syntactic classes of Danish adverbs.

1. Introduction

At CST, University of Copenhagen, a syntactic lexicon for Danish adverbs has newly been produced as one of the results of a Ph.D.-project financed by the Nordic language technology research programme 2000-2004 (for more information see www.norfa.no). The lexicon is integrated in STO[1] which is a national Danish follow-up to the former EU-funded lexicon-projects PAROLE and SIMPLE (see http://www.ub.es/gilcub/SIMPLE/simple.html#Language), and gives syntactic information on the type of modification and position as well as on several other syntactic lexical properties of approx 800 Danish adverbs, selected on the basis of their frequency in a text corpus. The lexical information is based on a series of syntactic tests as well as an individual examination of each adverb in a newspaper corpus of 30 mill. tokens ("Berlingske Aviskorpus", Berlingske Tidende & Weekendavisen 1999).

The lexicon, which can be used in NLP systems as well as for linguistic research, differs in several ways from earlier large-scaled computational adverb lexicons. First of all it is established by a corpus based study of the syntactic behavior of each adverb; secondly it focuses on properties which can be tested purely syntactically in order to keep a sharp distinction in the lexicon between syntax and semantics - semantic information on the adverbs is planned to be described afterwards at a semantic level in the STO lexicon, with links to the syntactic entries. For the human user, the lexicon furthermore contains a corpus example in every entry to illustrate one or more of the syntactic properties covered by the entry in question.

In the PAROLE project, which STO builds upon, the syntax of adverbs was not included in the Danish lexicon at all, since the project concentrated on the complement taking word categories: verbs, nouns and adjectives. In the Swedish PAROLE lexicon, adverbs were encoded with respect to which type of head they modify, but information on their position as well as on other syntactic properties was left out. The Italian and the Spanish PAROLE lexicons are the ones including the highest amount of syntactic information on adverbs within the PAROLE lexicon project. The Italian lexicon relies, however, on the general syntactic behavior of semantic classes of adverbs instead of examinations of the individual behavior of each adverb, and gives no information on word order behaviour. The Spanish lexicon describes more individual syntactic properties of adverbs than the Italian one does, such as their capability to be modified or to take a complement or an apposition, but still gives no information on position possibilities in the sentence, even though the position of adverbs plays a role (which is also mentioned in both the Italian and the Spanish documentation reports). For further information, see the documentation reports for Danish, Swedish, Italian and Spanish (http://www.ub.es/gilcub/SIMPLE/simple.html #Language).

English adverbs have been treated in two American computational lexicon projects. The first one, COMLEX (Macleod et al., 1998) gives very detailed information on syntactic properties as well as on semantic type in the lexical entry, but avoids to make the difficult, but in relation to NLP systems absolutely necessary distinction between V, VP, and S modification by grouping these under the same label 'clause-modifying'. Semantic features assigned afterwards divide this main group into subtypes such as 'time' adverbs, 'attitude' adverbs etc. The study of Danish adverbs have shown that some adverbs with a time meaning have syntactic properties different from the main group of 'time' adverbs and display instead

[1] STO (SprogTeknologisk Ordbase) has been funded by The Danish Ministry of Research. STO constitutes a large-sized Danish lexical database for NLP, linguistic research etc. and contains 45,000 entries which for all word classes give detailed morphological and syntactic information, and to some extent also semantic information, organized at three separate levels. For more information on STO, see Braasch & Olsen (2004) and http://cst.dk/sto.

similar properties to the ones normally characterising 'attitude' adverbs, indicating that a purely semantic subcategorization as in the COMLEX lexicon is not desirable. It might also be a disadvantage that there is no clear distinction between syntax and semantics, a distinction which is often aimed at within the field of NLP where the systems normally are constructed to manage the syntactic and the semantic processes in two separate steps.

Conlon & Evens (1994) describe another English adverb lexicon in the form of a database containing multiple kinds of information on English adverbs. The information is partly deduced semi-automatically from printed dictionaries (the lemmas and the semantic types), partly collected from the linguistic research on semantic groups of English adverbs over time (e.g. syntactic properties). The different types of information are organized and systematized in a way that should make them easily available for use in NLP systems, although many of the information types are made for human users, e.g. for linguistic research. As in the case of the Italian PAROLE lexicon, the syntactic information in the lexicon has been coded "top-down" from general rules on semantic classes of adverbs, without specific examination of each word. The consequence is that individual lexical properties of single words might very well be missing. The information on positional properties, however, is based on detailed corpus examinations of each single adverb, being extracted from an earlier printed lexicon of adverb positions in English from 1964 (by Sven Jacobson).

1. Linguistic Background for the STO Adverb Lexicon

The encoding principles for adverbs in the newly established Danish computational lexicon STO are developed on the basis of 1) the PAROLE lexicon coding formalism 2) a detailed corpus based examination of the syntactic behavior of 73 Danish adverbs, 3) studies of the information types in former NLP lexica for adverbs and 4) studies of literature on adverbs, especially Telemann et al. (1999), Quirk et al. (1972) and Hansen & Heltoft (2003). The 73 adverbs which were studied carefully as a starting point, represented all semantic adverb types as described in Telemann et al. (1999), namely: degree, manner, time and place adverbs, adverbs representing a valence bound actant, adverbs representing a logic relation (this group covers conjuncts and focus adverbs in the English literature (Quirk et al., 1972)), adverbs expressing negation, and finally adverbs expressing speaker attitude (also called disjuncts or sentence adverbs). Of these 73 adverbs, 39 were selected because of their polysemous properties, meaning that they have more than one main sense in a middle sized monolingual dictionary of modern Danish ('Nudansk Ordbog med etymologi', 1999). The 73 adverbs were studied 1) in concordance extractions of 100-120 lines (from "Berlingske Aviskorpus") for each adverb, which were tagged for syntactic behavior and afterwards sorted on the tags and 2) in a number of different syntactic surroundings set up to test the syntactic potential of each adverb.

2. Syntactic Properties of Danish Adverbs and Corresponding Encoding Solutions

The study focused on the prototypical behavior of the adverbs as individual words in the corpus, not taking into account how they behave interacting with other adverbs in the same phrase. One of the conclusions was that adverbs, not surprisingly, constitute a syntactically extremely eclectic word class, since they can modify all kinds of words and phrases and occur in many different positions. The different types of heads that adverbs can modify in the lexicon were finally defined as being the following: adjectives, adjective phrases, adverbs, the negation 'ikke', noun phrases, prepositional phrases, lexical verbs, verb phrases and sentences, leaving out quantifier modification (included in the NP modification) and infinitive modification (described indirectly by other properties). An often discussed problem within the field of formal linguistics, since decisive for the node attachment of clause adverbs in the syntactic parse trees, is the distinction between V, VP and S modification. The principles used in the lexicon in order to deal with this problem are the following:

An adverb modifes the lexical verb V
i) when it occurs in the so-called manner field in Danish sentences, between the object and the particle of a transitive phrasal verb (*Han har læst bogen omhyggeligt igennem* (Lit. HE HAS READ BOOK-THE CAREFULLY THROUGH, He has carefully read the book from end to end);
ii) when it constitutes a predicative adverbial (in the position for these in the Danish sentence, before a prepositional object): *De gav bogen sammen til ham* (Lit. THEY GAVE BOOK-THE TOGETHER TO HIM, They gave him the book together);
iii) when it constitutes a valency-bound adverbial: *Han tog derhen* (He went there) or replaces a prepositional object: *Han tænkte derover* (*derover* replacing *over det*) (Lit. HE THOUGHT THERE-OVER, He thought about it).

An adverb modifes the verbal phrase VP
when it does not satisfy the criteria for being a V-modifying adverb but is, as in the case of the V modifying adverbs, still able to occur in an independent infinitive construction with the verb: *At rejse senere / er dumt* (To travel later /is stupid); *Kun at rejse / er sjovt* (Only to travel / is amusing). It is especially marked in the entry when the adverb occurs outside (pre-modifies) the infinitive phrase, as in the case for *kun* (only).

Finally an adverb modies the whole sentence S when it cannot occur inside, nor outside an independent infinitive phrase, but only in inflected verb phrases or full sentences:
At rejse er sandelig dumt (To travel is indeed stupid)
* *At rejse sandelig /er dumt* (To travel indeed is stupid)
* *Sandelig at rejse /er dumt* (Indeed to travel is stupid).

As regards the position possibilities, these are not marked for the modification of the negation 'ikke' and for ADJP modification, since the position of the adverbs in both cases is always pre-positional. When the adverb modifies noun phrases, adjectives, adverbs and prepositional phrases we distinguish in the lexicon between pre- and postpositions (or both possibilities) For the clause modifying adverbs, we operate with 5 positions in the case of V modification and 4 positions for the cases of VP and S modification. The sentence in Figure 1 *Nu er han altså ikke tit løbet hurtigt ud herfra* (lit. NOW HAS HE REALLY NOT OFTEN RUN FAST OUT HERE-FROM, meaning 'He hasn't really that often left this place quickly') illustrates nearly all the position and modification possibilities for the three types of phrase modifying adverbs. Only the position for valency-bound adverbials or adverbs replacing a prepositional object is empty (BA).

Nu (S)	*er*	*han*	*altså* (S)	*ikke*	*tit* (VP)	*løbet* ->
NOW	IS	HE	REALLY	NOT	OFTEN	RUN ->
fundament			**nexus/theme**		**nexus/focus**	->

hurtigt (V)	*ud* (V)	Ø	*herfra* (VP)	
QUICKLY	OUT		HERE-FROM	
manner field	**predicative field**	**BA**	**final field**	

Figure 1: example of sentence with all adverb positions filled out, except Bound Adverbial field (BA). (S): modifies sentence, (VP): modifies verb phrase, (V): modifies lexical verb. Name of sentence position (Hansen & Heltoft, chapter 15, not yet published) is marked below each adverb.

Both the modification and the position capabilities of an adverb are conceived of as individual lexical properties, since the meaning of a polysemous adverb often depends on these two things, and since synonymous adverbs do not necessarily share the same modifying and positional characteristics. Furthermore, we define the following syntactic characteristics of adverbs to be lexical properties and therefore to be decribed in the lexicon:

- their capability of being modified themselves by another adverb,
- their ability to combine with negation[2],
- their capability of constituting the predicate in a predicative construction (also claimed by Zinsmeister & Heid (2003) to be a lexical property)
- their capability of subcategorizing for a prepositional phrase or a noun phrase, and
- their capability of occurring in a cleft sentence.

Finally it is worth mentioning that the two overall principles for establishing a syntactic entry in the lexicon are 1. type of head: one new entry per type, 2. word sense: one new entry per sense, even if the type of head is the same for the two senses. Table 1 shows 7 lexical entries.

[2] It could be discussed whether the requirement of negative or positive context is not semantic in its character, rather than syntactic.

Adverb	**Coding**	**Explanation of coding**
afgjort (definitely)	Dd1mTe_S	modifies sentence in nexus/theme position
åbenhjertigt (openly, frankly)	Dd1mFFoMå_V	modifies verb in fundament, nexus/focus and manner position
*meget*_1(very)	Dd1_ADJ	modifies adjective
*meget*_2 (very)	Dd1_ADV	modifies adverb
*her*_1 (here)	Dd1mFNS_VP	modifies VP in fundamental, nexus (theme as well as focus) and final position
*her*_2 (here)	Dd1m_PP	premodifies PP
*her*_3 (here)	Dd1mpost_NP	postmodifies NP

Tabel 1. Examples of lexical syntactic entries of adverbs. **Dd1** signifies in all cases **D**escription of a**d**verb with arity **1**, **m** signifies 'can itself be **m**odified by an adverb'.

2. Results Deduced from the Lexicon

The pattern descriptions in the lexicon make it very easy to deduce syntactic main classes of Danish adverbs and to examine possible links between the different syntactic properties. E.g. one main syntactic class is the class of adverbs able to occur in the manner field. The adverbs in this class share the capability of modifying the lexical V, of being able to modify a past participle, of being themselves modifiable by other adverbs and the missing ability to occur in the final position of the sentence (unless they are part of a coordinated adverbial). A subclass of this group of adverbs can also occur in the nexus/focus field, and a third subclass can occur both in manner, nexus/focus and fundamental field.

In the area of computational linguistics, the information on positions in the lexicon enables systems generating Danish text to place the adverbs correctly in the phrase. Furthermore information on the internal order in the nexus/theme field left of the negation can be indirectly deduced from the lexicon. The encodings show that within the group of adverbs which occur in the nexus/theme position, some are S modifying, (not able to occur in infinitives) while others are VP modifying (able to occur in infinitive). Looking closer into this, the S modifying adverbs always occur before the VP modifying ones in the nexus/theme field. We can see this by the fact, that only 'time' adverbs which cannot occur in an infinitive, are able to occur before an S modifying 'attitude' adverb. One example is *hidtil* (till now/till then), not able to occur in an infinitive phrase: * *At bo her hidtil har været rart* (To live here till now has been nice), but, opposite to most 'time' adverbs, able to occur in front of the 'attitude' adverb *desværre* (unfortunately):

Potentialet for at skabe … arkitektur af verdensklasse er hidtil desværre ikke blevet overbevisende realiseret.

(The potential for creating ... world-class architecture has till now unfortunately not been carried out in a convincing way) (Berlingske newspaper corpus 1999)

This makes us propose that the position for adverbs in Danish sentences, between the auxiliary and the main verb and left of the negation, can be divided in 2 sub-positions of which only the second one allows adverbs that are able to occur in infinitive phrases:

[S adverbs, not accepted in infinitive phrases
[VP adverbs, accepted in infinitive phrases
[Negation]]]

The encodings in the lexicon, however, do not tell us anything about the internal order within each of these two groups, e.g. in the case where two S modifying adverbs in the same sentence are to be generated in the nexus/theme field. The principles behind this order are semantic in nature and cannot be described by means of syntax.

In systems for text analysis, the information in the lexicon ensures the attachment of the adverb to the correct node in the syntactic parse trees. Within the framework of e.g. GB, where adverbs are assumedn to be base generated in a certain position and afterwards moved to other positions in the sentence, the difficult case of adverbs modifying the verb or the sentence (the 'clause-modifying adverbs' in COMLEX (Macleod et al., 1998)), is much discussed. Based on the encodings in the lexicon we propose the following analysis: All the V modifying adverbs are base generated somewhere to the right of the verb, e.g. the manner adverbs after the objects but before the bound adverbials and prepositional objects. We consider them complements to the lexical verb in these positions. This includes also the non obligatory manner adverbs, as proposed by (Abeillé & Godard, 2003). When the manner adverb, as a lexical property, is also able to occur to the left of the main verb (in the nexus/focus position, after the auxiliary), it is instead analyzed as an adjunct in the specifier position of the lexical verb V. The encodings reveal that the VP modifying adverbs display a much more diverse positional behavior. A large number of them are able to occur in the final field but some of them are only able to occur in the nexus and fundamental field. Others, however, never occur in the nexus field, and others again only in nexus/theme or in nexus/focus (though always in a position before the V modifying manner adverbs in nexus/focus). Consequently they are assumed to be base generated in different, and maybe more than one position, as proposed by Frey (2003). As regards the relation between position and modification, the following analysis for VP modifying adverbs is proposed: in the nexus field left of the main verb they pre-modify the whole VP inclusive all the V modifying adverbs, corresponding to left-adjunction of VP, whereas they post-modify the VP (right adjunction) when they occur in the final position of the sentence, after the verbal complements. If it is stated in the lexicon entry they can then move to the fundamental field. The S modifying adverbs are base generated in the nexus/theme position, left of the VP modifying adverbs. They can then be moved into the fundamental field and to a position at the end of the sentence if it is stated in the lexicon. Figure 3 illustrates the proposed basis positions.

Han har (HE HAS)
[**S adv** (*nok* PROBABLY)
[**VP adv** (*atter* AGAIN)
[V OBJ **V Adv** Particle] **VP adv**]]]
pakket den hurtigt ind derhjemme
WRAPPED IT QUICKLY UP AT HOME

Figure 3.Base generated positions for *nok* (probably), *atter* (again), *hurtigt* (quickly), *derhjemme* (at home).

3. Conclusions

The adverb lexicon constitutes the basis for many types of examinations of the syntactic behaviour of groups of adverbs. Furthermore, the use of the lexicon in NLP systems handling adverbs can improve the results in the parsing process as well as in the text generating process. It should be noted, however, that in spite of the detailed information in the lexicon, more than one syntactically correct output will still be produced in many cases, simply due to the many modification and position possibilities of the wordclass.

4. References

A. Abeillé & D. Godard (2003). The syntax of French adverbs without functional projections. In Martine Coene, Gretel De Cuyper, Yves D'hulst (Eds.), Current studies in Comparative Romance Linguistics. J. Benjamins, forthcoming.

Braasch, Anna & Sussi Olsen (2004). STO: A Danish Lexicon Resource - ready for Applications. In Fourth International Conference on Language Resources and Evaluation, Proceedings, LREC 2004, Lissabon.

Conlon, Sumali Pin-Ngern & Martha Evens (1994). An Adverbial Lexicon for Natural Language Processing Systems. In International Journal of Lexicography Vol. 7 No. 3, (pp197—221). Oxford University Press.

Frey, Werner (2003). Syntactic conditions on adjunct classes. In Ewald Lang, Claudia Maienborn, Cathrine Fabricius-Hansen (Eds.), Modifying Adjuncts. (pp. 163—209). Mouton de Gruyter, Berlin.

Hansen, Erik & Lars Heltoft (2003):"Grammatik – syntaks" (preliminary edition of Grammatik over det Danske Sprog, kap. 2), Skrifter fra Dansk og Public Relations, Roskilde University, Denmark, and preliminary chapter 15, not yet published.

Macleod, Catherine, Adam Meyers & Ralph Grishman (1998). The Syntactic Classification of Adverbs as an Update to COMLEX Syntax: An Addition to an On-line Ressource for Research in Syntax. In Proceedings of the ALLC/ACH'98, Hungary.

Nudansk Ordbog med etymologi (1999), Politikens forlag, Copenhagen.

Teleman, Ulf, Staffan Hellberg & Erik Andersson (1999). Svenska Akademiens Grammatik, Svenska Akademien, Stockholm.

Quirk, Randolph et al. (1972). A Grammar of Contemporary English, Longman.

Zinsmeister, Heike & Ulrich Heid (2003). Identifying predicatetively used adverbs by means of a statistical grammar model. In CL2003, Proceedings, Lancaster.