A Syntactic Lexicon for Danish Adverbs

Sanni Nimb
Center for Sprogteknologi (CST)
University of Copenhagen
Njalsgade 80, DK-2300, 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 is integrated in the Danish STO lexicon and 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. The pattern descriptions in the lexicon make it easy to deduce syntactic classes of Danish adverbs and to examine possible links between the different syntactic properties. In the case of the group of polysemous adverbs, the encoding process has revealed that the different senses seldom display the same syntactic behavior. This means that in far most cases, a correct syntactic analysis of the sentence in which the polysemous adverb occurs is the key to disambiguation

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 program 2000-2004 (for more information see www.norfa.no). The lexicon is integrated in STO (SprogTeknologisk Ordbase, see Braasch & Olsen (2004) and http://cst.dk/sto), a large-sized Danish lexical database for natural language processing (NLP) and linguistic research. STO is a national follow-up to the former EU-funded lexicon-projects PAROLE and SIMPLE (see http://www.ub.es/gilcub/SIMPLE/ simple.html #Language). The adverb lexicon part gives syntactic information on the type of modification and position as well as on several other syntactic lexical properties of approx 800 Danish adverbs. The adverb lemmas have been selected from the lemma candidate list of approx. 1200 adverbs of The Danish Dictionary project (Lorentzen, 2004) on the basis of their frequency in a corpus of 40 mill tokens (the corpus of The Danish Dictionary, see Asmussen & Norling-Christensen, 1998). The information in the lexicon 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 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, such as semantic type (e.g. 'time', 'place') and selectional restrictions, is planned to be described afterwards at a semantic level in the STO lexicon with links to the syntactic entries, in accordance with the SIMPLE lexicon model (Nimb & Pedersen, 2000). 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. 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 individual syntactic properties of adverbs, 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. For further information, see the documentation reports for 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, however, that some adverbs with a time meaning have syntactic properties different from the main group of 'time' adverbs and display instead similar properties to the ones normally characterizing 'attitude' adverbs, indicating that a purely semantic sub-categorization as in the COMLEX lexicon is not desirable. Conlon & Evens (1994) describe another English adverb lexicon in the form of a database for linguistic research and NLP 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). As in the case of the Italian PAROLE lexicon, the syntactic information in the lexicon (except from the information on positional properties) has been coded "top-down" from general rules on semantic classes of adverbs, without specific examination of each word.

In the newly established Danish adverb lexicon we have instead, as already mentioned, based the encoding process on corpus examinations of each word and tried to keep a sharp distinction between the syntactic and the semantic properties of adverbs

2 The syntactic behaviour of Danish adverbs

The encoding principles in the lexicon part on adverbs in STO are developed on the basis of 1) the PAROLE lexicon coding formalism 2) a detailed corpus based examination of the syntactic behaviour of 49 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 49 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 the 49 adverbs, 15 were polysemous, meaning that they have more than one main sense in a medium-sized monolingual dictionary of modern Danish ('Nudansk Ordbog med etymologi', 1999, Politikens Forlag, Copenhagen). The 49 adverbs were studied 1) in concordance extractions of 100-120 lines (from "Berlingske Aviskorpus") for each adverb, which were tagged for syntactic behaviour 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. The study of the adverbs focused on their prototypical behaviour 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* (not), 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 being decisive for the node attachment of clause adverbs in the syntactic parse trees, is the distinction between V, VP and S modification. The syntactic principles used for this in the lexicon 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 valence 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 in the lexicon for the modification of the negation *ikke* (not) and for ADJP modification (always being pre-positional). When the adverb modifies NP's, adjectives, adverbs and PP's, 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. Figure 1 illustrates nearly all the position possibilities (marked below). The type of head of each separate adverb in the phrase is marked in brackets.

Nu (S)	*er*	*han*	*altså* (S)	*ikke*	*tit* (VP)	*lobet*	*hurtigt* (V)	*ud* (V)	*herfra* (VP)
NOW	IS	HE	REALLY	NOT	OFTEN	RUN	QUICKLY	OUT	HERE-FROM
fundament			**nexus/theme**	**negation**	**nexus/focus**		**manner**	**predicative**	**final**

Figure 1: ('He hasn't really that often left this place quickly'). The sentence positions of clause modifying adverbs.

The described positions for adverbs in Danish sentences are proposed by Hansen & Heltoft (b), not yet published). Only a special field for sentence adverbs, also proposed by them, has not been implemented since defined only by semantic properties. In the encoding process this field has instead been regarded as a part of the nexus/theme field. The only position used in the lexicon, but not mentioned in Figure 1, is the shared position for valence bound adverbials or prepositional objects right after the predicative field.

3 The lexical syntactic properties and their representation in the lexicon

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 described in the lexicon:

- their capability of being modified themselves by another adverb,
- their ability to combine with negation
- their capability of constituting the predicate in a predicative construction
- 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. This last principle is relevant to the cases of polysemous adverbs. Table 1 shows 5 lexical entries.

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 position and manner position
*her*_1 (here)	Dd1mFNS_VP	modifies VP in fundamental, nexus (theme as well as focus) and final position
her 2 (here)	Dd1m_PP	pre-modifies PP
her 3 (here)	Dd1mpost_NP	post-modifies NP

Table 1 Examples of lexical syntactic entries of adverbs. **Dd1** signifies in all cases Description of adverb with arity **1**, **m** signifies 'can itself be **m**odified by an adverb'.

4 The syntactic behaviour of polysemous adverbs

Within the field of computational linguistics 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, and in addition the adverb lexicon makes it possible to carry out many different types of researches on the syntactic behavior of single words or groups of adverbs. See Nimb (2004) for a more detailed description.

Looking more into the syntactic behavior of the group of polysemous adverbs, it became clear during the corpus examinations and the encoding process that they differ from other types of polysemous words. For the other word classes it is normally the words in the nearest context that enables the lexicographer to quickly distinguish and tag the different senses within a set of concordance lines. In the case of the polysemous adverbs, the different senses instead clearly appear once the tagging and the sorting of the lines with respect to the syntactic behavior of the adverb has been carried out. This experience led to a further study of polysemous adverbs. Apart from the initially studied 15 ones, the remaining polysemous adverbs from the dictionary 'Nudansk Ordbog med etymologi', i.e. those with more than one main sense, were found (57 more), and 24 of these were studied as well. Out of these 15 + 24 = 39 adverbs, 8 were sorted (those where only one sense was in fact present in the corpus). Of the remaining 31 adverbs, only for 7 adverbs the syntactic behavior of the different senses was identical. Lexical semantic information on e.g. the head of the adverb was in these cases required in order to be able to disambiguate between the senses. For the remaining adverbs (approx 75 %), the different syntactic behaviours displayed by the adverb were, on the contrary, each of them connected to only one of the two or more senses of the adverb.

These results might explain why automatic word sense disambiguation in a system using part-of-speech tagging extended with semantic knowledge (in the form of dictionary definitions, selectional restrictions and thesaural hierarchies), as described in Stevenson & Wilks (2000), does not obtain the same good results for adverbs (68.63 % correctness) as for other word categories (approx 90 % for nouns, verbs and adjectives). In the major part of the cases of adverbs it seems instead to be a detailed syntactic analysis of the phrase in which the adverb occurs that leads to word disambiguation.

5 Concluding remarks

In spite of the detailed syntactic information in the lexicon, more than one syntactically correct output will still often be produced in NLP systems using it, even if the system has a well-developed grammar. This is simply due to the many modification and position possibilities of adverbs. Furthermore, some phrases just are structural ambiguous. One example is the Danish sentence: *Han er vel ankommet* (*vel* understood as S modifying adverb: lit. HE HAS I SUPPOSE ARRIVED (He has arrived, I suppose); *vel* understood as V modifying manner adverb: lit. HE HAS WELL ARRIVED (He has arrived in good order)). It seems that rather than semantic lexical information, it is factors 'beyond' the lexical units, such as world knowledge, pragmatic phenomena, intonation and the placing of stress in the sentence (as in the case of *Han er vel ankommet* where intonation and stress changes the meaning of *vel*), which are relevant in order to solve the structural ambiguities. These factors are of course difficult to formalize in an NLP system, and the problem of deciding correctly between several analyses of phrases with adverbs purely by automatic means still remains to be solved

References

Asmussen, J. and Norling-Christensen, O. 1998. 'The Corpus of the Danish Dictionary' in Lexikos 8, AFRILEX Series 8:1998. Stellenbosch: Buro van die WAT.

Braasch, A. and Pedersen, B.S. 2004. 'Recent Work in the Danish Computational Lexicon Project "STO" 2002 in Braasch, A. & C. Povlsen (eds.), The Tenth EURALEX International Congress, Proceedings, Vol. I. Copenhagen, pp. 301-314

Conlon, S.P. & Evens, M. 1994. 'An Adverbial Lexicon for Natural Language Processing Systems' in International Journal of Lexicography Vol. 7 No. 3. Oxford: Oxford University Press.

Hansen, E. and Heltoft, L. a) 2003. Grammatik–syntaks. (Preliminary edition of 'Grammatik over det Danske Sprog', chapter 2). Skrifter fra Dansk og Public Relations, Roskilde University, Denmark and b) (not yet published) preliminary chapter 15 of 'Grammatik over det Danske Sprog'.

Lorentzen, H. 2004. 'The Danish Dictionary at large: Presentation, Problems and Perspectives' in The Eleventh EURALEX International Congress, Proceedings. Lorient: Université de Bretagne Sud.

Macleod, C., Meyers, A. and Grishman, R. 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.

Nimb, S. 2004. 'A Corpus-based Syntactic Lexicon for Adverbs' in Fourth International Conference on Language Resources and Evaluation, Proceedings, LREC 2004, Lissabon.

Nimb, S. and Pedersen, B.S. 2000. 'Treating Metaphoric Senses in a Danish Computational Lexicon' in The Ninth EURALEX International Congress, Proceedings, Vol. II. Stuttgart: Universität Stuttgart.

Stevenson, M. and Wilks, Y. 2000. 'Large Vocabulary Word Sense Disambiguation' in Ravin, Yael & Claudia Leacock (Eds.) Polysemy. Oxford: Oxford University Press.

Teleman, U. Hellberg, S. and Andersson, E. 1999. Svenska Akademiens Grammatik. Stockholm: Svenska Akademien.

Quirk, R. et al. 1972. A Grammar of Contemporary English. London: Longman.