Heuristic Schema Parsing of Swedish Text

Kenneth Wilhelmsson
Department of Linguistics
University of Gothenburg
kw@ling.gu.se

Abstract

A method for identification of the *primary* (main clause) functional constituents of Swedish sentences is outlined. The method gives a robust analysis of the unbounded constituents (phrases which do not have an upper bound on their length: subjects, objects/predicatives and adverbials) by first identifying bounded constituents. Diderichsen's sentence schema, chunking, syntactic valency data and heuristics are used for the delimitation of the constituents and labelling with grammatical functions.

1 Introduction

Description of the moderately fixed word order of the main clause in Nordic languages has, since the work by the Danish linguist Paul Diderichsen (1946), often taken on a field-oriented form (see Table 1). Diderichsen's sentence schema is used here to provide "short-cuts" in the steps of the heuristic parsing, through its stating that schema positions mostly have restricted numbers of contents, of restricted forms, and if localized, need not always be thoroughly analyzed, for marking up segments with syntactic functions. Being V2 languages, Nordic languages have declarative main clauses where the second position is occupied by a finite verb. This delimits the preceding field, the fundament field – which may, however, also be preceded by a *förfält*[1]. As compulsory components of main clauses, with a fixed position in the clause and the length of one word, primary finite verbs play a crucial role in the parsing method presented.

Fundamental field	**Nexus field**			**Content field**		
Fundament	v	n	a	V	N	A
(Fronted constituent)	Finite	Subject	Adv	Non-finite	Obj	Adv
Trots vädret	*ska*	*han*	*inte*	*ta*	*bilen*	*idag*
Despite the weather	*will*	*he*	*not*	*take*	*the car*	*today*
B/U	B	U	B/U[2]	B	U	U

Table 1: An adaptation of Diderichsen's main clause schema showing basic Swedish declarative word order and boundedness of positional content. (B: Bounded, U: Unbounded).

Starting with the obligatory primary finite verbs, the parsing method seeks to delimit and identify contents of fields through correct identification of all primary bounded constituents. Bounded constituents include verbs, particles, reflexive pronouns, some adverbials and conjunctions coordinating main clauses or primary verb phrases. Primary bounded constituents are identified mainly through removal of other candidates, e.g. licensing by identification of starts of subclauses.[3] This work is based on Stockholm Umeå Corpus 2.0 (Ejerhed *et al*, 1992) and its tagset (unmodified).

2 Identification of Unbounded Constituents and Functional Labelling

Rank-based chunking is a technique for the identification of NPs, PPs and *"som-predikat"* ("as-predicates"), e.g., *Som målvakt var han bra/As a*

[1] The *förfält* does not contain functional constituents of the clause, but typically conjunctions.

[2] Some sentence adverbials (which often are placed here) are recursive, according to examples in Teleman *et al.* (1999). It is, however, a useful fact that many sentence adverbials are bounded.

[3] These processes were described at SLTC 2006.

goalkeeper he was good. (The chunks do not include post-modifying attributes, like PPs or relative clauses.) The procedure starts by assigning a rank number to each word. There are some common forms of NPs which are not properly matched using the ranks. These can mostly be taken care of later.[4]

Wordclass/word	Rank	Wordclass/word	Rank
Som (as a conjunction)	16	Cardinal number, adverb	3
Preposition	15	Participle, adjective, ordinal number	2
Word in genitive case	1/14	Measure nouns[5]	1.5
Determiner	5	Pronoun, proper name, noun	1
Possessive	4		

Table 2: A rank assigning function represents each word with a rank according to this table. The rank is mostly decided by the word-class tag, and no further feature values – thereby avoiding errors which might arise from incorrect tagging of the other features.

In the field areas not occupied by bounded primary constituents, the subsequent chunking procedure uses the following algorithm for chunking by the ranks: Any word with a lower or equal rank number than the previous one is seen as part of the same chunk as the previous word, except for an adjacent pair of words, both of rank 1, which do not produce a personal name.[6] If, however, the rank number is higher than that of the preceding word (or if it is the first word), the assumption is that a new chunk starts by that, and that the previous one is terminated. After this run-through, the default type of the chunk can easily be assigned. If the first word of a chunk is a preposition (rank 15), that chunk will be a PP, similarly rank 16 will make the chunk an "as-predicate". In other cases, the chunk will be a NP/adjective phrase by default, unless the lowest rank number of the chunk is higher than 3 – then it may be an adverb phrase. Special cases include conjunctions that will make the chunk continue regardless of previous and coming rank of words. Proper names will make the chunk include two adjacent ones if the first is part of set of first names and/or the second is part of a set of last names. Words in genitive case will make the current chunk continue, their symbolic rank number 1/14 signals this.[7]

Using valency data for nouns, adjectives, participles and verbs mainly from *Nationalencyklopedins ordbok* (1995-96) and *Lexin – Svenska ord* (1998), chunks are merged into larger ones, encompassing e.g. subclauses.[8] This process includes many manually constructed heuristic rules merging chunks together, often by looking at adjacent words in adjacent chunks. Using the information of Diderichsen's schema and heuristics, functional labeling finalizes the process.

3 Some Preliminary Results

Using (correctly tagged) random test sets from SUC 2.0 not used for training, a manual evaluation has been made for the current state of the implementation. For primary finite verb identification, precision was 99.5 and recall was 97.0 in a set of about 200 sentences. Primary subject identification (taken to mean *exact match* of primary subjects, including all attributes) was correct in 92.6 % of all main clauses in a set of sentences including 405 main clauses. In 2.2 %, subjects were matched partly, and in 0.7 % included in a too long segment marked as primary subject. These results are likely to improve.

References

Paul Diderichsen. 1946. *Elementær Dansk Grammatik.* Gyldendal

Eva Ejerhed, Gunnel Källgren, Ola Wennstedt and Magnus Åström. 1992. *The Linguistic Annotation System of the Stockholm-Umeå Corpus Project.* Department of General Linguistics, Umeå University, Report no. 33

Nationalencyklopedins ordbok, 1995-96, Höganäs

Lexin – svenska ord. 1998, Norstedts Akademiska Förlag

Ulf Teleman, Erik Andersson and Staffan Hellberg. 1999. *Svenska Akademiens grammatik*

[4] Quantifier attributes (which are tagged as adjectives), like *all* in *all/2 fairly/3 new/2 ones/1*, will for example be excluded from the noun phrase, by the basic rank chunking procedure.

[5] *'Measure nouns'* here means nouns taking an NP complement. In Swedish this is frequent and does not use '*of*': *en kopp kaffe/a cup of coffee.*

[6] Lists of more than 21 000 names are used.

[7] The analogy is a deck of cards, where genitive words are "aces", having both the rank 1 and 14 in the continuation of steadily descending numbers that form a chunk.

[8] The lexicon valency information has been turned into lists of prepositions etc. which start post attributes of these words. (A base form functionality is used for the look-up.)