# Reducing Complexity in Parsing Scientific Medical Data, *a Diabetes Case Study*

**Dimitrios Kokkinakis** 

Center for Language Technology and Språkbanken Department of Swedish University of Gothenburg, Sweden dimitrios.kokkinakis@svenska.gu.se

#### Abstract

The aim of this study is to assemble and deploy various NLP components and resources in order to parse scientific medical text data and evaluate the degree in which these resources contribute to the overall parsing performance. With parsing we limit our efforts to the identification of unrestricted noun phrases with full phrase structure and investigate the effects of using layers of semantic annotations prior to parsing. Scientific medical texts exhibit complex linguistic structure but also regularities that can be captured by pre-processing the texts with specialized semantically-aware tools. Our results show evidence of improved performance while the complexity of parsing is reduced. Parsed scientific texts and inferred syntactic information can be leveraged to improve the accuracy of higher-level tasks such as information extraction and enhance the acquisition of semantic relations and events.

### 1 Introduction

Linguistic annotation of textual corpora in any field, and in specialized fields in particular, is a demanding and complex task, absolute necessary for data-driven language processing, human language technologies and knowledge mining. One such type of processing is at the syntactic level, i.e. syntactic parsing. The aim of this study is to develop and evaluate a method of identifying unrestricted noun phrases with full phrase structure from a scientific medical corpus. To ease the evaluation, in lack of an appropriate gold standard, we selected random sentences from the available corpus with mentions of the word diabetes. This subset then was automatically annotated and manually inspected and corrected. Furthermore, we tried to be minimalistic by assembling and deploying various existing NLP components and resources in order to evaluate the degree in which these resources contribute to the

overall parser's performance. Analysis of scientific texts is a challenging task caused by deviant and idiosyncratic uses of vocabulary and syntax and complex linguistic structure. However, we believe that there are also regularities that can be captured by pre-processing the texts with specialized semantic tools. This way the complexity of parsing in scientific discourse, e.g., ambiguities, can be reduced, while the technical vocabulary increases the lexical coverage. For parsing we use finite-state cascades and sequential finitestate transducers. The focus of the current work is on the extraction of complete noun phrases, an important step that upon succession paves the way for the extraction of more complex structures and functional syntactic relations. Parsing is important for in-depth semantic interpretation; inferred syntactic and semantic information can be used to improve the accuracy of higher-level tasks such as information extraction and enhance the acquisition of relations and events.

## 2 Background

Parsing technology has seen a dramatic improvement over the last decade and a number of fairly robust parsers are available for a growing number of languages and application domains. This is a trend that has been accelerated by the appearance of wide coverage grammars and statistical parsing modules, both based on the availability of various treebanks such as the Penn or the GENIA treebanks (cf. Rimell & Clark, 2009). The commonest strategies to parsing are constituency/phrase structure or dependency parsing; for a review of parsing strategies, cf. Ljunglöf & Wirén (2010). In the first, words combine into phrases which repeatedly combine to form the sentence; while in the second syntactic analysis take the form of binary relations, that hold between words; Pyysalo (2008); Nivre (2005). Phrase structure grammars yield fast and reliable results without the need of large (annotated) corpora while dependency parsing is essentially

very similar to the concept of valency extended to all word classes. In the medical field, there have been a number of approaches to syntactic parsing (Leroy et al., 2003; Ohta et al., 2005; Lease & Charniak, 2005). The goal for most of these approaches was with the extraction of various types of relations between phrases with named entities, e.g. proteins, since good precision and recall figures for extracting such relations requires a reliable syntactic analysis of the text. A large body of work to dependency parsing in the (bio)medical domain is based on the GENIA corpus Kim et al. (2003); see for instance Rinaldi et al. (2008) and Pyysalo (2008). Nevertheless, the issue of domain adaptation of existing grammars is still an open issue. It has been discussed that adaptation efforts should be on lower, local levels of representation (domain specific part-of-speech, dictionary collocations, named entities, terminology) not on full parse trees (Leash & Charniak, 2005; Huang et al., 2005; Aubin et al., 2005; Grover et al., 2005; Hogan et al., 2011). For example, Leash & Charniak (2005) showed clear improvements of the parsing accuracy considering a combination part-of-speech/named entities. Accuracy increased from 81.5% to 82.9% of a Penn Treebank-trained parser applied on biomedical literature. For a survey of comparing and combining six state-of-the-art chunkers for the biomedical domain see Kang et al., (2010).

# 3 Materials and Method

The ever-increasing amount of biomedical (molecular biology, genetics, proteomics) and clinical data repositories increase in a dramatic manner. Such data appropriately annotated with event-level information are a valuable source of evidence-based research and text mining activities, such as information extraction, semantic search, question&answering and knowledge discovery. Syntactic parsing is considered an important ingredient for event-based information extraction from medical free text. Extracting pieces of information pertaining to specific events requires the extraction of argument mentions, often syntactic, that play a specific role within the event. In order to support the automated extraction of events, annotated corpora with eventlevel information is a necessary requirement; cf. Wattarujeekrit et al. (2004) and Thompson et al. (2009).

For our study we have selected a random sample of 120 sentences<sup>1</sup> with the mention of the word diabetes from a large corpus of scientific medical Swedish (Kokkinakis & Gerdin, 2010). The average length of a tokenized sentence in the sample is 23,8 tokens. Despite the small size of the sample, we can still find characteristics, typical of the medical scientific language such as terminology overload and coordinative constructions. For instance, it is rather common with coordinative phrases such as: Man har visat att serumnivåerna av CRP, IL-6, fibrinogen, PAI-1, amyloid A och sialinsyra är förhöjda vid typ 2diabetes 'It has been shown that serum levels of CRP, IL-6, fibrinogen, PAI-1, amyloid A and sialic acid are elevated in type 2 diabetes'. These characteristics are expressed by a syntactic and vocabulary variability which compared to "ordinary" language requires adapted parsing strategies in order to be able to effectively capture the peculiarities of the genre. Terminology and named entity recognition can contribute to reliably resolve some of these problems.

# 3.1 Parsing Method

For the identification and labeling of noun phrases we apply an easy-first parsing deterministic approach using finite-state cascades. A finitestate cascade consists of a sequence of levels; phrases at one level are built on phrases at the previous level. Levels consist of rules, alias groups of patterns, ordered according to their internal complexity and length. A pattern consists of a category and a regular expression and parsing consists of a series of finite transductions. Spans of input elements are reduced to a single element in each transduction; i.e. regular expressions are translated into finite-state automata, the union of which yields a single, deterministic, finite-state, level recognizer; for further details of the approach cf. Abney (1997). We use an existing generic grammar for modern Swedish which has been evaluated for both basic noun phrases and functional labels in general language corpora, for the noun phrases the precision reported was 97.82% and recall 94.5% (but without resolved attachments); details are reported in Kokkinakis & Johansson Kokkinakis (1999). The workflow of all the various steps are shown in Figure 1; here "attachment" refers to nouns and adjectives while the "term & entity aware rules" are integrated in the parser.

<sup>&</sup>lt;sup>1</sup>The sample and some of the resources, e.g. multiwords, can be found at: <a href="http://demo.spraakdata.gu.se/svedk/parse/">http://demo.spraakdata.gu.se/svedk/parse/</a>>.



Figure 1. The workflow of the parsing process.

### 3.2 Parsing Adaptation, Preliminaries

Manual annotation of large amount of data with complex linguistic information, such as syntactic trees, is a costly enterprise, and means to remedy for this should be exploited. Moreover, since many parsers rely on several layers of representation there are various possible ways to enhance their performance even in different domains than the one they have been designed for. Our methodology is motivated by the fact that parsing performance can be gained by applying and improving on a number of pre-parsing stages. The idea is that various morphosyntactic and semantic representation layers can pave the way of substantial text complexity reduction as long as the parser can be made aware of these layers. Therefore, by putting effort on various levels of representation (pre-processing) we can hypothesize, and actually show, that performance can be improved. Thus, we help the parser in such a way that it can avoid some hard decisions, e.g. bracketing and structural ambiguities. We follow, and to a certain degree, extend the idea of Lease & Charniak (2005) discussed earlier. Our strategy is primarily based on four (domain) adaptations:

- recognition of multiword expressions
- recognition of medical terminology
- recognition of named entities
- attachment for nouns & adjectives.

In an indirect way, the recognition of terminology and named entities implies that multiword expressions are also recognized and the number of unknown words is reduced, while the lexical coverage increases. Consider for instance the examples: Drottning Silvias Barn- och Ungdomssjukhus 'Queen Silvia Children's Hospital' in which 5 tokens constitute a *coherent* entity and the more complex: DIGAMI 1-studien (Diabetes mellitus insulin glucose infusion in acute myocardial infarction 1) [3] visade att [...]. 'The DIGAMI-1 study [...] showed that [...]' in which 9 tokens constitute a single entity. Apart from the terminology recognition (and the manual addition of domain vocabulary at the lexical resources we use) the rest are domain-independent adaptations.

## 3.3 Parsing Adaptation, Steps

Adaptations deal with the resolvement of at least *some* of the possible types of problems that can arise during parsing. For instance, we manage to dramatically increase the lexical coverage by efficiently dealing with unknown words, e.g., genre specific vocabulary. The majority of unknown words can be captured by the use of domain terminologies. A number of individual terms from such terminologies have been incorporated into the part-of-speech tagger's lexicon, for that purpose we use the TnT tagger (Brants, 2000). Similarly, for various types of multiword tokens, we have manually added a large number of common multiword function words (e.g., adverbs, preposition, determiners) in the part-ofspeech tagger's lexicon<sup>2</sup>. While for the majority of other types of multiword expressions (i.e., terms and named entities) which are identified during terminology and named entity recognition, possibly erroneous part-of-speech annotation does not have impact during parsing. For instance the part-of-speech annotation of the segment: en latent diabetes mellitus 'a latent diabetes mellitus' becomes: en/DI@US@S latent/AQPUSNIS diabetes/NCUSN@IS mellitus/XF (XF stands here as a tag for foreign

<sup>&</sup>lt;sup>2</sup> Nivre & Nilsson (2004) have showed that significant improvement in parsing accuracy for Swedish could be achieved if multiword function words are taken under consideration.

words) while the parsing of this segment (in a simplified form) becomes *np*: <*en latent di*abetes> np: <mellitus>, that is two separate noun phrases. However, if we apply terminology recognition and then combine (in some suitable way) that information with the part-of-speech (e.g., by adding a feature to the part-of-speech) then we end with the following annotation: en/DI(a)US(a)Slatent/AQPUSNIS diabetes/ NCUSN@IS-TRM-B mellitus/XF-TRM-I. The parser, being aware of these new features, will treat diabetes and mellitus as a unit (e.g., with a rule such as ART? ADJ\* TERM+) and favorize the term annotation, since term and entity label features are given higher precedence compared to part-of-speech tags by the parser. In this case, the parser produces a correct phrasal constituent, one noun phrase, namely *np-mdcn:* <en latent *diabetes mellitus>.* 

The following example will be used to illustrate some of these steps in sections 3.3.1-3.3.3:

Lit: "Malmö 22 January 2008 - New data published today, confirm the effectiveness of basal insulin Levemir® (insulin detemir) as a treatment once a day for people with type 2 diabetes.".

### 3.3.1 Medical Terminology Recognition

We use the Swedish Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT) for terminology recognition. Terminology is actually used for two reasons: (i) to improve the performance of the generic part-of-speech tagger and (ii) to actually aid the recognition of the terminology and consequently also the annotation of terms in text in which the parser has been modified to be aware of. First, we extracted one-word terms (ca 30k) and semi-automatically added those with their full morphosyntactic description, to the part-of-speech tagger's backup lexicon. Using regular expressions over the suffixes of the terms we automatically added appropriate morphosyntactic descriptions and manually reviewed a number of unmatched cases, usually Latin terms, which we added with the label for foreign words. Secondly, we performed terminology recognition and then merged the output to the representation format required by the parser, thus the parser becomes aware of the terminology in a single, simple step. The previously mentioned example follows below after terminology annotation (annotations are given between the XML tag *snomed* with attributes c concept, h idnumber and o original form). Note that for simplicity reasons *qualifier values* have been filtered away:

Malmö 22 januari 2008 - Nya data som publiceras idag styrker effektiviteten hos <snomed c="substance" h="25305005" o="långtidsverkande insulin">basinsulinet</snomed> Levemir® (<snomed c= "substance" h="414515005" o="detemir insulin">insulin detemir</snomed>) som behandling en gång om dagen för personer med <snomed c="disorder" h="44054006" o="diabetes mellitus typ 2">typ 2diabetes</snomed>.

### 3.3.2 Named Entity Recognition

In exactly the same manner, as previously, we apply the generic named entity recognizer which also serves two important purposes. Firstly, to aid the recognition and annotation of single and multiword named entities and secondly, in an indirect way, to aid the appropriate recognition of (unknown) multiword expressions/tokens. The annotation of the previous example shown this time below illustrates how this type of annotation looks like. After named entity recognition the example sentence takes the following form:

<ENAMEX TYPE="LOC" SBT="PPL">Malmö </ENAMEX> <TIMEX TYPE="TME" SBT="DAT"> 22 januari 2008</TIMEX> - Nya data som publiceras <TIMEX TYPE="TME" SBT="DAT"> idag</TIMEX> styrker effektiviteten hos basinsulinet <ENAMEX TYPE= "OBJ" SBT= "MDC">Levemir®</ENAMEX> (insulin detemir) som behandling <NUMEX TYPE="MSR" SBT="FRQ">en gång om dagen</NUMEX> för <ENAMEX *TYPE="PRS"* SBT="CLC">personer </ENAMEX> med typ 2-diabetes.

In the above annotations, *ENAMEX* stands for a named entity, *TIMEX* for a time entity and *NU-MEX* for a measure entity. All annotations produce also two attributes (not used in the current study) namely main *TYPE* and *SuBType*; details are provided in Kokkinakis (2004).

## 3.3.3 Structural Ambiguity / Attachment

For the structural ambiguity/preposition attachment disambiguation we use a generic Swedish valency/subcategorization lexicon which has been manually enhanced for genre-specific nouns (such as *ulceration*), which all take a contextually optional prepositional phrase as complement; e.g., *ulceration av tumören* 'ulceration

Malmö 22 januari 2008 - Nya data som publiceras idag styrker effektiviteten hos basinsulinet Levemir® (insulin detemir) som behandling en gång om dagen för personer med typ 2-diabetes.

of the tumor') and adjectives (such as *resistent*) which also take a contextually optional prepositional phrase as complement, e.g. resistent mot autokrint insulin 'resistant to autocrine insulin'. There is also a small number of nouns that show a semantic preference for two arguments, such as övergång 'transition, as in övergång från blodglukos till plasma-glukos 'transition from bloodglucose to plasma glucose'. This type of lexical information is applied after part-of-speech tagging using a contextually-driven filter that determines whether a suitable feature can be added to nouns' or adjective's part-of-speech annotation. This is a naïve but reliable way to capture lexical semantic preferences without a lot of effort. Thus, in the example from section 3.3 there are two such tokens identified and annotated with the feature -VAL, for valency (attached to the appropriate nominal or adjectival heads); the descriptions morphosyntactic NCUSN@DS. NCUSN@IS stand for common nouns and SPS stands for a preposition, the tags for the rest of the words have been omitted for simplicity. The tagset we use is an extended version of the Swedish MULTEXT tagset<sup>3</sup>.

Malmö 22 januari 2008 - Nya data som publiceras idag styrker effektiviteten/NCUSN@DS-VAL hos/SPS basinsulinet Levemir® (insulin detemir) som behandling/NCUSN@IS-VAL en gång om dagen för/SPS personer med typ 2-diabetes.

#### 3.3.2 Merging and Parsing Awareness

All results are merged into a uniform representation. In order to make the parser aware of all the annotations we have added two new levels of manually written rules into the parser's original sequence of levels. Recall that each level contains a handful of rules, and phrases recognized by the rules of one level are built on phrases at the previous level. The two new levels, at the very beginning of the parser's level set, are used to only recognize and process the sequences of terminology and named entity annotations, labeling them as either noun phrases, e.g., nplocation, np-disorder, or adverbial phrases, as in the case of time expressions. Examples of rules for these new levels and an example of parsing output are given in Appendix A&B. The parser provides several possible ways to produce output results. For instance, all features can be explicitly generated and in the example we can see *lemma*,

the base form of each token and *sem* the semantics (term or entity labels) or N/A (non-applicable) otherwise.

Note that each level is proceeded by abbreviated bundles of enhanced part-of-speech tags, e.g., one mnemonic name for all adjectives (*ADJ*) or one for all possible location annotations (*LOC-B* and *LOC-I*). By this technique the actual rules become simpler, flexible and human readable; examples are also given in Appendix B. This convention does not exclude the possibility to actually use a particular part-of-speech tag or even word in a rule, which implies that rules can be *lexicalized* i.e. use words in the production rules.

### 4 Results and Evaluation

We performed a four-stage evaluation in order to measure the contribution of the adaptations to the overall performance of the results.

Model	Pr	R	F-m	F-
				impr.
baseline comp. to the gold	39.53%	47.22%	43.04%	
-backup+mwe	56.76%	58.33%	57.53%	14,4%
-backup+mwe+val	53.66%	61.11%	57.14%	14.1%
-backup+mwe+val+NER	64.10%	69.44%	66.67%	23,6%
-backup+mwe+val+NER+term	89.19%	91.67%	90.41%	47,3%

### Table 1: Evaluation results

We manually corrected the output of the automatic parsing with *all* adaptations and used it as a gold standard for the evaluation, using the 2008-version of the *evalb* software by Sekine & Collins with default values. The only change we made was to convert the parsing output in order to run *evalb*. Brackets, that the parser returns as delimiters, were converted to parentheses, that *evalb* requires. The results in table 1 stand for: *Pr* bracketing precision, *R* bracketing recall, *F* bracketing f-measure and *F-impr*. for f-measure improvement.

We also measured the number of unknown words in the sample before and after the vocabulary enhancement of the part-of-speech tagging for the sentences examined showed only a small improvement, mainly because the majority of unknown words were actually annotated correctly by the tagger; 388 tokens (13,6%) were unknown (5,4% of those were wrongly annotated), reduced to 215 unknown tokens (7,5%) with the use of the enhanced vocabulary. The number of multiword function words was 27, e.g., *t ex* 'for example'.

<sup>&</sup>lt;sup>3</sup> <http://spraakbanken.gu.se/parole/tags.phtml> (last visited in August 2011).

#### **5** Conclusions

To our knowledge there hasn't been any other report to use of parsing Swedish medical corpus of any type. Therefore any direct comparisons are difficult to make. With simple modifications, as shown in Table 1, the parser becomes aware of the different (shallow) semantic annotations produced during pre-processing. We believe of course that any type of parsing strategy can benefit from the integration of this type of annotations. The results show a substantial improvement of accuracy which can be attributed to a number of factors such as structural ambiguity reduction, increasing lexical coverage, enhanced processing of coordinative structures. For the future we intend to extend the subcategorizations to verbs, particularly since relevant Swedish lexical resources are available for that purpose. We also plan to adapt the rest of the original, generic parser in order to be able to support event-based information extraction from Swedish medical corpora.

#### Acknowledgments

This work is partially supported by the Centre for Language Technology (CLT) <a href="http://clt.gu.se/">http://clt.gu.se/</a>.

#### References

- Abney S. 1997. Part-of-Speech Tagging and Partial Parsing. In *Corpus-Based Methods in Language* and Speech Processing. Young & Bloothooft (eds). 4:118-136. Kluwer.
- Aubin S., Nazarenko A. and Nédellec C. 2005. Adapting a general parser to a sublanguage. Recent Advances in Natural Language Processing (RANLP) Pp. 89-93. Bulgaria.
- Brants T. 2000. TnT: a statistical part-of-speech tagger. Sixth Conference on Applied NLP. Seattle, USA.
- Grover C., Lapata M. and Lascarides A. 2005. Comparison of Parsing Technologies for the Biomedical Domain. *Natural Language Engineering* 11(1), 27-65, CUP.
- Hogan D., Foster J. and van Genabith J. 2011. Decreasing Lexical Data Sparsity in Statistical Syntactic Parsing - Experiments with Named Entities. Proceedings of the ACL Workshop: Multiword Expressions: from Parsing and Generation to the Real World (MWE). Portland, USA
- Huang Y., Lowe HJ., Klein D. and Cucina RJ. 2005. Improved Identification of Noun Phrases in Clinical Radiology Reports.. J Am Med Inform Assoc. 12(3): 275–285.

- Kang N., van Mulligen EM. and Kors JA. 2010. Comparing and combining chunkers of biomedical text. *J Biomed Inform.* 44(2):354-60. Epub Nov 4.
- Kim J-D., Ohta T. Tateisi Y. and Tsujii, J. 2003. GE-NIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*. 19 Suppl 1:i180-2.
- Kokkinakis D. and Johansson Kokkinakis S. 1999. A Cascaded Finite-State Parser for Syntactic Analysis of Swedish. 9th European Chapter of the Association of Computational Linguistics (EACL). 245-248. Norway.
- Kokkinakis D. 2004. *Reducing the Effect of Name Explosion*. Workshop: Beyond Named Entity Recognition, Semantic Labelling for NLP tasks. In conjunction with the 4th Language Resources and Evaluation Conference (LREC). Lisbon, Portugal.
- Kokkinakis D. and Gerdin U. 2010. A Swedish Scientific Medical Corpus for Terminology Management and Linguistic Exploration. Seventh Language Resources and Evaluation Conference (LREC). Malta.
- Lease M. and Charniak E. 2005. Parsing Biomedical Literature. Second International Joint Conference on Natural Language Processing. 58-69. Korea.
- Leroy G., Chen H. and Martinez J.D. 2003. A shallow parser based on closed-class words to capture relations in biomedical text. *J Biomed Inform*. 36(3):145-58.
- Ljunglöf P. and Wirén M. 2010. Syntactic Parsing. In *Handbook of Natural Language Processing*, 2nd edition. Indurkhya N. and Damerau FJ. (eds). Pp. 59-91. CRC Press.
- Nivre J. 2005. Dependency Grammar and Dependency Parsing. Technical Report. Växjö University.
- Nivre J. and Nilsson, J. 2004, Multiword Units in Syntactic Parsing. MEMURA 2004 workshop. pp. 39-46. Lisbon
- Ohta T., Tateisi Y. and Tsujii J. 2005. Syntax Annotation for the GENIA corpus. IJCNLP, pp. 222-227. Korea.
- Pyysalo S. 2008. A Dependency Parsing Approach to Biomedical Text Mining. Turku Centre for Computer Science. TUCS Dissertations: 105, U of Turku, Finland.
- Rimell L. and Clark S. 2009. Porting a lexicalizedgrammar parser to the biomedical domain. *J of Biomed Inf.* 42:5.
- Rinaldi F., Schneider G., Kaljurand K. and Hess, M. 2008. Dependency-Based Relation Mining for Biomedical Literature. Sixth Language Resources and Evaluation Conference (LREC). Morocco.
- Sekine S. and Collins MJ. 2008. EVALB. <a href="http://nlp.cs.nyu.edu/evalb/EVALB20080701.tgz">http://nlp.cs.nyu.edu/evalb/EVALB20080701.tgz</a>

- Thompson P., Iqbal SA. McNaught J. and Ananiadou S. 2009. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, 10:349.
- Wattarujeekrit T, Shah PK. and Collier N. 2004. PASBio: predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics.* 19;5:155.

### **APPENDIX A**

ı

```
<s id="955">
2
       [np-location
з
         [NPOON@OS-LOC-B <t id="955 1"/> Malmö lem=malmö]]
4
5
       [rp-time
         [MCOONOS-TME-B <t id="955 2"/> 22 lem=22]
б
         [NCUSN@IS-TME-I <t id="955_3"/> januari lem=januari]
7
         [MCOONOS-TME-I <t id="955 4"/> 2008 lem=2008]]
       [FI <t id="955 5"/> - lem=-]
       [np
10
         [AQPOPNOS <t id="955 6"/> Nya lem=ny]
11
         [NCNPN@IS <t id="955 7"/> data lem=data]]
12
       [np-som
13
         [PH@0000gs <t id="955 8"/> som lem=som]]
14
       [vg p f
15
         [V@IPSS <t id="955 9"/> publiceras lem=publicera]]
16
17
       [rp-time
         [RGOS-TME-B <t id="955 10"/> idaq lem=idaq]]
18
       [vq a f
19
         [V@IPAS <t id="955 11"/> styrker lem=styrka]]
20
       [np_attach_pp
21
         [np-val
22
           [NCUSN@DS-VAL <t id="955 12"/> effektiviteten lem=effektivitet]]
23
         [pp
24
           [SPS <t id="955 13"/> hos lem=hos]
25
26
           [np-medical
              [NCNSN@DS-MDC-B <t id="955 14"/> basinsulinet lem=basinsulin]
27
              [NPOON@OS-MDC-B <t id="955 15"/> Levemir@ lem=levemir@]
28
              [FP <t id="955_16"/> ( lem=(]
29
              [NCNSN@IS-MDC-B <t id="955_17"/> insulin lem=insulin]
30
              [XF-MDC-I <t id="955 18"/> detemir lem=detemir]
31
              [FP <t id="955 19"/>) lem=)]]]
32
       [CCS <t id="955 20"/> som lem=som]
33
       [np attach pp
34
         [np-val
35
            [NCUSN@IS-VAL <t id="955 21"/> behandling lem=behandling]
36
             [np-mesr-frq
37
               [DI@US@S-MSR-B <t id="955 22"/> en lem=en]
38
               [NCUSN@IS-MSR-I <t id="955 23"/> gång lem=gång]
39
               [SPS-MSR-I <t id="955_24"/> om lem=om]
40
               [NCUSN@DS-MSR-I <t id="955 25"/> dagen lem=dag]]]
41
42
         [pp
            [SPS <t id="955 26"/> för lem=för]
43
            [np-person
44
              [NCUPN@IS-PRS-B-VAL <t id="955 27"/> personer lem=person]]]
45
         [pp
45
            [SPS <t id="955 28"/> med lem=med]
47
48
            [np-medical
              [NCUSN@IS-MDC-B <t id="955 29"/> typ lem=typ]
49
              [NCUSN@IS-MDC-I <t id="955 30"/> 2-diabetes lem=2-diabetes]]]]
50
       [FE <t id="955 31"/> . lem=.]
51
52
     </s>
```

Example parsing with basic constituent annotations including the part-of-speech and the feature *lem*[ma].

### **APPENDIX B**

```
1 #Level 'medical entities'
 2 :mdcn-entity
   article = D000P08|D00US08|D000P08|D00NS08|DF00P08|DF00S08|DF0US08|DF0US08|...;
 3
   adjective = AF00PG0S|AF00PN0S|AF00SNDS|AF0MSNDS|AF0NSNIS|AF0USNIS|AP000N0S|...;
 4
   adverb = RGOA | RGOC | RGOS | RGCS | RGPS | RGSS | RHOS;
 5
   numerical = MC0000C|MC00G0S|MC00N0S;
 6
   MDCN-B = AF00000A-MDC-B|AF000N0S-MDC-B|AF00PG0S-MDC-B|AF00PN0S-MDC-B|...;
 7
   MDCN-I = AF00000A-MDC-I|AP000NOS-MDC-I|AF00PG0S-MDC-I|AF00PNOS-MDC-I|...;
 8
 9
10
   np-medical -> article? (numerical|adjective)* MDCN-B MDCN-I*
11
                 # Rules for Special Cases:
                       e.g. 'vita "leverfläckar" i.e., 'while "liver spots"'
12
                 #
                 | article (numerical|adjective|adverb)* FP MDCN-B MDCN-I* FP
13
                 | article? FP (numerical|adjective|adverb)+ FP MDCN-B MDCN-I*
14
15
                 1 ...
16
  #Level 'location entities'
17
18 :loc-entity
19 article = DOGOPGS|DOGUSGS|DOGOPGS|DOGNSGS|DFGOPGS|DFGOSGS|DFGNSGS|DFGUSGS|...;
20 adjective = AFOOPGOS | AFOOPNOS | AFOOSNDS | AFOMSNDS | AFONSNIS | AFOUSNIS | APOOONOS | ...;
21 adverb = RGOA | RGOC | RGOS | RGCS | RGPS | RGSS | RHOS;
22 numerical = MCOOOOC | MCOOGOS | MCOONOS;
23
   LOC-B = AF00000A-LOC-B|AF000N0S-LOC-B|AF00PG0S-LOC-B|AF00PN0S-LOC-B|...;
24 LOC-I = AF00000A-LOC-I|AF000N0S-LOC-I|AF00PG0S-LOC-I|AF00PN0S-LOC-I|...;
25
26 np-location -> article? (numerical|adjective)* LOC-B LOC-I*
27
                 # Rules for Special Cases, e.g. 'USA, Canada och England'
                 | LOC-B (FI LOC-B) + CCS LOC-B
28
                 | article (numerical|adjective|adverb)* FP LOC-B LOC-I* FP
29
                 | article? FP (numerical|adjective|adverb)+ FP LOC-B LOC-I*
30
31
                 1 . . .
32 ;
33 #Rest of the grammar ...
34 . . .
35 #Level 'attachment'
36 :attachment
37 np = np-medical|np-location|...;
38 np-val = NCUSN@DS-VAL | NCUSN@IS-VAL | NCNPN@IS-VAL | ...;
39
40 # Attach a PP using the 'VAL' feature that is added to certain tokens after
41 # part-of-speech tagging
   np attach pp ->
42
   # Here, 'SPS' is the part-of-speech for any preposition
43
          np-val [pp = SPS np]
44
45
```

Part of the grammar that shows several available levels. The first two (and new to the existing generic grammar) deal with medical and location entities followed by other entity specific and then by general rules (not shown here). At the end of the figure there is an example of an attachment rule, applied after the recognition of basic phrase constituents, e.g. noun phrases and verbal groups (i.e., an obligatory lexical head plus optional auxilaries or even adverbs if those intervene between an auxiliary and a head verb).