The NT(M)S Parser: an Efficient Computational Linguistic Tool

Fliedl, Günther
Department of Computational Linguistics
University of Klagenfurt
Carinthia, Austria
guenther.fliedl@uni-klu.ac.at

Mayerthaler, Willi
Department of Computational Linguistics
University of Klagenfurt
Carinthia, Austria
willi.mayerthaler@uni-klu.ac.at

Winkler, Christian
Department of Computational Linguistics
University of Klagenfurt
Carinthia, Austria
christian.winkler@uni-klu.ac.at

Abstract

NT(M)S stands for ***N**atürlichkeits**T**heoretische (**M**orpho)**S**yntax* [6]. This is a grammar model based on generative syntax. The descriptions of grammatical phenomena are represented by trees expressing both constituency and dependency relations. These trees are unfoldings/projections of lexical base-categories. The maximal unfoldings of the respective categories (verb, noun etc.) are markedness theoretically parametrized: the less unmarked, the higher the unfolding. The syntaxtheoretic language of NT(M)S is two-fold. It provides a formal syntactic description and a markedness theoretic evaluation thereof.

Keywords: Natural Language Processing (NLP), Syntax Parsing, Knowledge Engineering, Computational Linguistics.

1. Some Theoretical Features of the Model

We distinguish between dominant and subdominant percolation/heritage. Every tree has just one head and accordingly just one dominant heritage line. The head of a construction usually is a lexical category. So called functional heads normally belong to the area of subdominant heritage.

Given our framework, we do not use abstract theta-grids, but specified predicate-argument-structures filled with concrete semantic rôles as for instance

AGENT/THEME/INSTRUMENT/GOAL/
LOCATION/SOURCE etc.

As in valency or dependency grammar the verb is the head of the whole sentence. Within the noun phrase, the noun is the only permissible head. Functional elements like definite articles, auxiliaries and non-governing affixes are generated under a specifier-node (SPZ).

A markedness-theoretical evaluation of the generated phrase-structures is given according to the principles of naturalness theory [7].

2. Computerlinguistic Aspects of Knowledge Processing

Knowledge processing, as required in conceptual design and predesign, requires a high degree of efficiency from a computerlinguistic model for the analysis of language.

This necessitates both an information base for the language itself and the exact formal definition of linguistic knowledge about the natural language which implies

- a linguistic dictionary,
- a concept lexicon which represents word meanings,
- a computational grammar,
- a process capable of determining the meaning with respect to grammatical values and to the context.

The aim of the NT(M)S approach is to comply with those requirements with regard to praxis orientated

Proceedings of the Workshop on Computer Science and Information Technologies CSIT'99
Moscow, Russia, 1999

computerlinguistic work. However, there are the following additional requirements in the field of NT(M)S:

- the lexicon should be context-sensitive,
- basic concepts of the meaning of words are included in this (restrictive) lexicon,
- determinating parameters as *grammatical values* and *context* should be made explicit by means of a markedness-theoretical system evaluating these parameters.

Computational grammar is considered as the *missing link* between natural language discours and the schemes applied/defined in engineering several conceptual models. In other words, syntactic structures can be translated directly into the respective modeling concepts, the input being any German sentence related to the chosen segment of reality. The analysis is done in some projects, e.g. in NIBA [1], [2], [3], [5] containing the following steps:

- In **phase 1** the words of a text are compiled into a lexicon. This implies automatic categorial, semantic and contextual specification of words.
- In **phase 2** word- and morphosyntactically interpreted microstructures are assigned to wordgroups and words, e.g. the automatic splitting of the german verbal noun *Übersetzer*:

n0_TH_Ri_mask(v([Übersetz_<AGi,TH>]),n_min([er_<Ri> mask]))

This semantically enriched structure consists of Theta-roles (TH = „theme“; AG=agent“; R= referential rôle), the morphosyntactic feature „mask“ for masculine, the categorial nodes n0 = noun, n_min= nominalizing suffix and the word-components *Übersetz* and *-er.* [7]

- **Phase 3** is dedicated to the analysis of (sentence)syntax according to the specific X'-mechanism of NT(M)S [6]. This mechanism is a generator of phrase-structures defining word-phrases as unfoldings/projections from lexical entities; NT(M)S basic lexical categories are:

  V^0 (verb), N^0 (noun), A^0 (adjective), P^0 (preposition), Q^0 (quantifier), ADV^0 (adverb), PT^0 (particles like *sehr*, *mindestens*, *genau*), and SPZ^0 (auxiliaries and determiners). We assume the following maximal unfoldings/ projections :

V^{max}	=	V^4 (main clause)
V^{max}	=	V^3 (subordinate clause)
N^{max}	=	N^3
N[+präd] = N^{max-1}	=	N^2
A^{max}	=	A^2
Q^{max}	=	Q^2
ADV^{max}	=	ADV^2
P^{max}	=	P^2
PT^{max}	=	PT^2
INT^{max}	=	INT^0
ART[-def] = Q^{max-2}	=	Q^0
$SPEZ^{max}$	=	$SPEZ^0$
AUX^{max}	=	AUX^0

$$\boxed{X^n \rightarrow X^n/X^{n-1} \ldots}$$

X^n in case of category recursion and X^{n-1} elsewhere, X^{n-1} = head.

0≤n≤4/[V, main clause]

0≤n≤3/[N, V(subordinate clause)]

0≤n≤2/[A, Q, ADV]

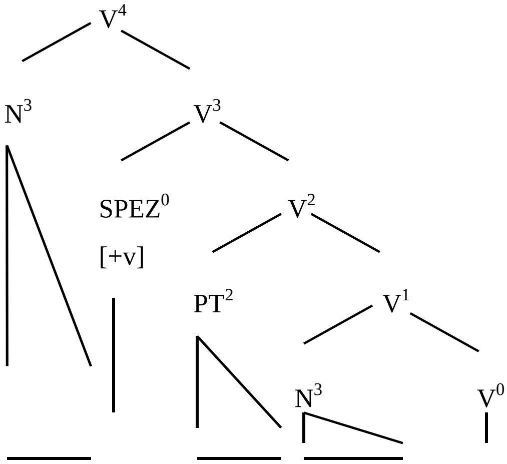
0≤n≤2/[P, PT]

n=0/[SPEZ, INT,PT]

3. Some Examples of Sentence Analysis and Interpretation

3.1 Derivation of perspective determiners

In the following we generate NT(M)S-based syntax trees for the purpose of exemplification of our proposal for automatic derivation of (pre) conceptual schemes out of specific natural language patterns. According to basic assumptions of the NT(M)S, in the default case German *als*-phrases make nouns to *perspective determiners* [FKMMW97]. In the subsequent example *Ein Angestellter betreut als Vertreter ein Gebiet (An employee looks after a district as a representative),* the adverbial phrase *als Vertreter* is analyzed as V^2-adjunct:



Ein Ang. betreut$_i$ *als Vertr. ein Gebiet* AG[TH]$_i$

The potential end-position of the non-finite verb *betreuen (to look after)* is represented in the tree diagram by the rightmost position V^0, in our example filled with features regarding the lexicon. *Betreuen* establishes a relation between the agentive subject *Ein Angestellter* and the object *ein Gebiet (district)* and can thus be called a transitive agent verb. Transitive agent verbs (tVag/2) are part of the class of unmarked bivalent verbs whose subject controls the action encoded by the verb. The object involved, in the default case represented by an accusative N^3, carries the TH(EME)-role. The predicate-argument-structure AG[TH] for unmarked (natural) transitive agent verbs is a result of this definition. Square brackets are intended to distinguish between the external argument AG and the internal argument TH, placed within brackets.

The following characteristics are typical for a*ls*-phrases:

- The particle *als* takes a nounphrase (N^3) as its internal argument,
- *als* is the relational head of a particle-phrase in its maximal unfolding PT^2,
- *als* establishes a semantic relation between the verbal core of the sentence (*betreuen*) and the nounphrase *manager*, dominated by *als,*
- *als* does not involve case-government,
- *als* establishes an agreement relation between the topicalized subject *Angestellter* and the NP *Vertreter*; the nominative case of *Vertreter* is thus assigned by agreement, not by means of government.

There is no agreement between the object-NP *ein Gebiet* and *Vertreter,* so that *Vertreter* has to be perceived of as (adverbial) perspective-determiner of *ein Angesteller betreut.* In this structure two different relations of agreement are involved:

- Case agreement in the nominative between the subject *Ein Angestellter* and the N^3 *Vertreter* within the particlephrase. Although agreement is not unequivocally encoded by declension, it is crucial.
- Semantically motivated agreement of the subject and the N^3 in the particlephrase. Both nounphrases agree at least regarding the semantic notation [+hum], which is a prerequisite for the given coreference relation.

Thus, the interaction of the agreement features [+hum] and [+nom] is the semanto-syntactic condition which had to be considered in the development of the parser, in order to sensitivize automatical syntax analysis for the sub-classification of *als*-phrases. The interplay of agreement features is expressed by the index 'k' in the following matrix:

Ein Angestellter betreut ein Gebiet als Vertreter

$[+hum]_k$	[-hum]	$[+hum]_k$
$[+nom]_k$	[-nom]	$[+nom]_k$

Violating the agreement features leads to a non-acceptable sentence or to the interpretation in the sense of copredication:

**Ein Ang. betreut ein Gebiet als den Vertreter.*

[+nom] [-nom]

**Ein Auto betreut ein Gebiet als Vertreter.*

[-hum] [+hum]

Der Vertreter betreut dieses Gebiet...

[+hum] [-hum]

...als seine wichtigste Provisionsgrundlage.

[-hum]

Relations are, in the default case, lexically rooted (this is the case in the context of the transitive verb *betreuen*). The subject involved carries the agent role and as a consequence the feature [+hum]. An *als*-phrase involved has to comply with certain conditions to be analyzed as *perspective determiner*. The semanto-syntactic interplay of agreement features can be perceived as a set of conditions for the identification of nounphrases as perspective determiners.

4. Derivation of generalization/specialization

Normally, generalizations/specializations are expressed by means of a *is-a*-relation. This corresponds syntactically to a predication with a predicative noun, e.g. *Ein Linguist ist (ein) Wissenschafter*.

NP-external specialization is enabled by the qualifying copula in sentences like *Linguisten sind Wissenschafter* or *Wale sind Säugetiere* (*whales are mammals*). Compare the following structure:

```
v4(n3(n2(q0([ein]),(n0([Linguist_j]))),v3(spz0([ist_v_i]),v2(
n3(n0([j_])),(n2(q0([ein]),n0([Wissenschafter])),v0([<TH,N2
>i])))))
```

With respect to interpretation, the following parameters play a relevant role:

(1) *ist (ein)* (is a)

(2) *Wissenschafter* = N^2 (Predicative noun; *ein* = indefinite article)

(3) *Linguist* = N^0

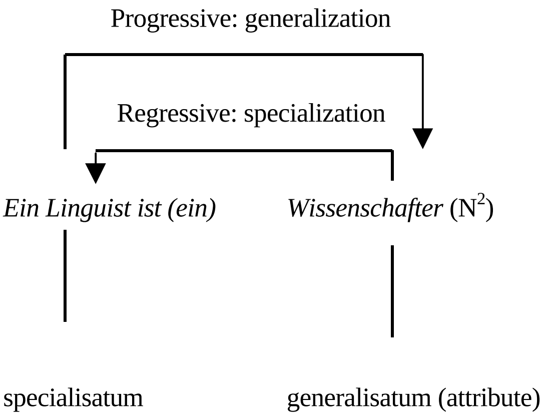
(4) The argument structure assigns the TH-rôle to the subject position and the unspecified maximal projection X^2 to the predicative noun. Compare: ([<TH,X^2_i_>]))))).

The N^2-node is of interest in two respects:

- Given that N has the exponent '2', this term can be interpreted as an attribute[1].
- Given that '*Wissenschafter*' is a noun 'N', it can be classified as generalisatum of *Linguist.*

[1] Typical X^2-phrases in predicative position are A^2-phrases as *Willi ist informiert* (Bill is informed). A^2-phrases are pure attributes.

Inversely, *Linguist* is specialisatum of *Wissenschafter*. The relation between *Linguist* and *Wissenschafter*, that is the relation of a specialization (regressive) vs. generalization (progressive) is exemplified in the following figure:


Progressive: generalization
Regressive: specialization
Ein Linguist ist (ein)
Wissenschafter (N²)
specialisatum
generalisatum (attribute)

5. Conclusion

The NT(M)S as a computational grammar is considered as the „missing link“ between a sector of the *Universe of Discourse* and the schemes in some sorts of conceptual modeling. We propose that semantically enriched syntactic structures can be directly translated into a variety of database notions, the input being any German sentence related to the chosen segments of reality.

References

1. Fliedl G, Kop Ch, Mayerthaler W, Mayr H, Winkler Ch. „Dinge und Zusammenhänge: NTS-gestützte Ableitung konzeptueller Vorentwurf-schemaeinträge aus natürlichsprachlichen Anforderungsdefinitionen.“. In: Ortner E et al. (eds) *Natürlichsprachlicher Entwurf von Informationssystemen.* Universitätsverlag, Konstanz, 1996, pp 260-279 (*Schriften zur Informationswissenschaft 25*)

2. Fliedl G, Kop Ch, Mayerthaler W, Mayr H, Winkler Ch. „NTS-Based Derivation of KCPM Cardinalities: From Natural Language to Conceptual Predesign“. In: van de Riet R et al (eds) *Applications of Natural Language to Information Systems.* IOS Press, Amsterdam-Oxford-Tokyo-Washington, 1996, pp 222-233

3. Fliedl G, Kop Ch, Mayerthaler W, Mayr H, Winkler Ch. „Zur automatischen Generierung von Vorentwurfsmodellen für die Datenbankentwick-lung“. *Papiere zur Linguistik* 1996; 55:153-174

4. Fliedl G, Mayerthaler W, Winkler Ch. „LEXIKON der Natürlichkeitstheoretischen Syntax und Morphosyntax“. *Papiere zur Linguistik* 1996; 55:176-200

5. Fliedl G, Kop Ch, Mayerthaler W, Mayr H, Winkler Ch. „NTS-Based Derivation of KCPM Perspective Determiners“. In: McFetridge P (ed) *Applications of Natural Language to Information Systems.* The Harbour Centre Campus of SFU, Vancouver, 1997, pp 215-226

6. Fliedl G, Mayerthaler W, Winkler Ch. „Lexikon der Natürlichkeitstheoretischen Syntax und Morphosyntax“. Stauffenburg, Tübingen, 1998

7. Fliedl, G. „Natürlichkeitstheoretische Morphosyntax - Aspekte der Theorie und Implementierung“. Narr, Tübingen, 1999