# Ontology-based semantic construction, underspecification and disambiguation

**Philipp Cimiano**
Institute AIFB
University of Karlsruhe
cimiano@aifb.uni-karlsruhe.de

**Uwe Reyle**
Institut für Maschinelle Sprachverarbeitung
University of Stuttgart
uwe@ims.uni-stuttgart.de

## 1 Introduction

Ambiguity is ubiquitous and one of the major problems in NLP. It occurs at the lexical level, in certain syntactic constructions, meaning assignment and also pragmatically in determining the purpose of a particular sentence. Each component of an NLP system thus faces the problem of ambiguity control.

At the levels of tokenizing, part-of-speech tagging and morphological analysis, the progress that has been made within linguistics and within computational linguistics over the past years has brought systems into reach that are able to cope with the ambiguity problem at their own level. This success is mainly due to the development of statistical disambiguation techniques and inductive methodologies to derive linguistic knowledge from large data samples. To cope with the ambiguity problem at the level of syntax, packed parse forests are used to represent sets of grammatical representations and statistical algorithms allow in addition to put probabilistic weights on analyzes.

The experimental and inductive methodologies as well as the amount of linguistic knowledge needed for the development of a natural language interpretation system varies, however, between its different levels of analysis. The derivation of detailed lexical knowledge that is relevant to the construction of semantic representations is a significantly more complex task than the derivation of knowledge needed for a computational grammar or morphology. And so is the task of dealing with ambiguity control within semantics and pragmatics.

Within semantics the problem of ambiguities has been approached by underdetermination of meaning. Main emphasis has been put on the underspecified representation of lexical ambiguities (van Deemter, 1991; van Deemter, 1995; Reyle, 1996; Buitelaar, 1998; Buitelaar, 1996; Pustejovsky, 1998), quantifier scope ambiguities (Reyle, 1993; Bos, 1995; Pinkal, 1996), and ambiguities arising from the distributive/collective distinction in connection with plural NPs (Reyle, 1996). Problems of text coherence are dealt with in (van Deemter, 1991; Reyle, 1995; Schiehlen, 1999; Roßdeutscher and Reyle, 2000; Reyle and Roßdeutscher, 2001).

To interleave syntactic and semantic ambiguities (Schiehlen, 1999; Schiehlen, 1996; Dörre, 1997) construct underspecified semantic representations on the basis of syntactic forests; and (Pinkal, 1996) tries to derive semantic representations from syntactically ambiguous or incomplete input by means of functional application. (Muskens, 2001) moves from syntactic structures and truth conditions to descriptions thereof and this way offers a uniform way to underspecify syntax and semantics.

In this paper we extend Musken's Logical Description Grammar (LDG) approach (Muskens, 2001) such that ambiguity may in addition to the mutual syntactic and semantic constraints LDG provides be restricted by (i) lexical semantic information and (ii) an underlying ontology associated with the lexical items. In this paper we apply our approach to the underspecification and resolution of lexical ambiguities. In general, the approach is also applicable to the resolution of PP-attachment, coordination as well as thematic role ambiguities (compare (Bunt, 2003)). We will show this in a further paper. Interestingly, Musken's LDG approach allows a straightforward combination with statistical methods such as mentioned above. In particular it seems possible to consider independently derived part-of-speech tags (by a statistically trained POS tagger as in (Schmid, 1994) for example) for the input sentences and thus have an additional and external source of disambiguation besides the lexical and ontological knowledge. We do not further investigate this possibility in this paper.

## 2 Logical Description Grammar

Muskens' Logical Description Grammar is to a great extent based on LTAG (Joshi and Schabes, 1997) and D-Tree grammars (Rambow and Vijay-Shanker, 1995), but offers a declarative account of the tree operations used in these formalisms. Muskens distinguishes three kinds of descriptions: (i) *general descriptions*, (ii) *input descriptions*, and (iii) *lexical descriptions*.

**ad(i):** General descriptions are nothing but a set of axioms defining linguistic tree structures by means of the two binary relations proper dominance $\lhd^+$ and linear precedence $\prec$[1] In addition each node $k$ of a tree is (a) labeled by its part-of-speech type, e.g. $l(k) = dp$, where $l$ is the labeling function, (b) assigned a positive and negative anchor, $\alpha^{+/-}$. The positive anchor of a node $k$ is required to be lexical by the axiom $\forall k \; lex(\alpha^+(k))$; the negative anchor requires the same for all nodes except for the root node (i.e. $\forall k(k = r \vee lex(\alpha^-(k))))$ which is negatively anchored to itself by the axiom $\alpha^-(r) = r$. The role of these anchoring functions is to enforce a pairing of nodes that are only positively or negatively marked in the elementary tree descriptions of lexical entries such that each node of the resulting tree is marked both positive and negative.

**ad(ii):** Input descriptions are constructed on the basis of the sentence to be analyzed. The input description of sentence (1) is given in (2). It states that there are exactly 2 lexical nodes carrying the lexemes of the sentence which are linearly ordered by $\prec$.

(1) Ginger croaked.

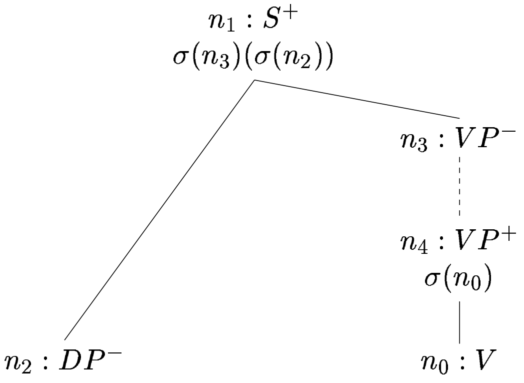
(2) $\exists n_1 \exists n_2 (n_1 \prec n_2 \wedge Ginger(n_1) \wedge croaked(n_2)$
$\wedge\, \forall n(lex(n) \leftrightarrow n = n_1 \vee n = n_2))$

**ad(iii):** Lexical descriptions come in two types. The *classifying descriptions* associate with each lexical item $\gamma$ its part-of-speech tag and its semantic value $\gamma'_\tau$ with type $\tau$. $croak'_{\langle e,t\rangle}$ and $croak''_{\langle e,t\rangle}$ correspond to the *caw* and *die* interpretations of *croak*.

(3)
$$\begin{aligned}
&\forall n\,[Ginger(n) \rightarrow \\
&\quad ((l(n){=}pn \wedge \sigma(n){=}Ginger'_e) \\
&\quad \vee\; (l(n){=}cn \wedge \sigma(n){=}ginger'_{\langle e,t\rangle}) \\
&\quad \vee\; (l(n){=}adj \wedge \sigma(n){=}ginger'_{\langle\langle e,t\rangle,\langle e,t\rangle\rangle}))] \\
&\forall n\,[croaked(n) \rightarrow (l(n) = iv \wedge \\
&\quad (\sigma(n) = croak'_{\langle e,t\rangle} \vee \sigma(n) = croak''_{\langle e,t\rangle}))]
\end{aligned}$$

*Elementary tree descriptions* associate substructures of trees with lexical items. For intransitive verbs they require the configuration in (4). The lines represent immediate dominance $\lhd$ and the dotted line its transitive closure, $\lhd^*$.

(4)



Formally (4) is expressed by (5), where $\Gamma(n_0, n_1, n_2, n_3, n_4)$ abbreviates a set of relationships expressed in terms of immediate dominance $\lhd$, dominance, $\lhd^*$, $\alpha^+$ and $\alpha^-$ that mirrors the structure in (4).

(5)
$$\begin{aligned}
&\forall n_0\,[l(n_0) = iv \rightarrow \exists n_1...n_4(l(n_1){=}s \wedge \\
&l(n_2){=}dp \wedge l(n_3){=}vp \wedge l(n_4){=}vp \wedge \\
&\sigma(n_1) = \sigma(n_3)(\sigma(n_2)) \wedge \sigma(n_4) = \\
&\sigma(n_0) \wedge \Gamma(n_0, n_1, n_2, n_3, n_4))]
\end{aligned}$$

Further classifying and elementary tree descriptions that are relevant in the context of (1) together with their corresponding formulas are given in (6),[2]

[1] For the exact formulation of these axioms we refer the reader to (Muskens, 2001). Note that infinite trees are not ruled out yet. Finiteness of the models of the descriptions will only be guaranteed in conjunction with the input descriptions.

[2] We oversimplify syntactic representations here. The point is that the subject DP either dominates a proper name or a (bare) common noun.

(6) $$\begin{array}{ccc} n_6\text{:DP}^+ & n_8\text{:DP}^+ & n_8\text{:DP}^+ \\ | & | & | \\ n_7\text{:PN}^- & n_9\text{:CN}^- & n_9\text{:ADJ}^- \end{array}$$

$$\forall n\ [l(n){=}pn{\rightarrow}\exists n'(l(n'){=}dp \wedge \Gamma(n,n'))]$$
$$\forall n\ [l(n){=}cn{\rightarrow}\exists n'(l(n'){=}dp \wedge \Gamma(n,n'))]$$
$$\forall n\ [l(n){=}adj{\rightarrow}\exists n'(l(n'){=}dp \wedge \Gamma(n,n'))]$$

The general axioms, the input description, the descriptions of the elementary trees and the ontological axioms then form a logical theory and the models of this theory correspond to all the possibilities of successfully selecting the appropriate elementary trees for each word in the input description as well as combining these trees to yield a syntactic and semantic representation of the sentence under consideration. Parsing thus boils down to identifying positively anchored nodes with negatively anchored ones such that category, tree and order information is respected. This will, however, lead to 6 semantically different readings of (1). The reason is that the theory does not take ontological nor lexical semantical information into account.

## 3 The interaction of lexical meanings

In order to restrict the set of readings the information that both $croak_{die}$ and $croak_{caw}$ require living beings for their subject must be brought into play and compared with the ontological classes of the three interpretations of the lexem *Ginger*. We will thus assume an ontology the backbone of which is a concept hierarchy in which the relevant concepts for a certain domain are partially ordered with regard to their generality/specifity. Furthermore, the ontology specifies semantic relations between concepts, e.g., that *roots* are some sort of *plants*, that the *agent* of a *croaking* state is a *animate*, etc. Formally, our ontological model is basically the one described in (E. Bozsak et al., 2002):

**Definition 1 (Ontology)**
*An ontology is a structure* $O := (C, \leq_C, R, \sigma_O)$ *consisting of (i) two disjoint sets* $C$ *and* $R$ *called concept identifiers and relation identifiers respectively, (ii) a partial order* $\leq_C$ *on* $C$ *called concept hierarchy or taxonomy, (iii) a function* $\sigma_O : R \rightarrow C^+$ *called signature. We denote the immediate subordination wrt.* $\leq_C$ *by* $\leq_C^+$

A corresponding ontology reflecting the conceptualization relevant for the interpretation of example 1 is depicted in figure 1. The classifying descriptions are now enriched such that they are located within the structure of the underlying ontology:

(7) $$\begin{aligned}&\forall n\,[Ginger(n) \rightarrow \\ &((l(n){=}pn \wedge \sigma(n){=}Ginger'_e \wedge c(n) = \textsc{person}) \\ &\vee\ (l(n){=}cn \wedge \sigma(n){=}ginger'_{\langle e,t\rangle} \\ &\wedge\ c(n) \leq_C^+ \textsc{root} \leq_C^+ \textsc{plant}) \\ &\vee\ (l(n){=}adj \wedge \sigma(n){=}ginger'_{\langle\langle e,t\rangle,\langle e,t\rangle\rangle} \wedge \\ &c(n) = \textsc{colour}))]\end{aligned}$$

The function $c$ maps nodes to concepts of the ontology. To rule out that two concepts corefer we must extend the axioms of the general descriptions with 8.

(8) $c_1 \neq c_2$, if $c_1$ and $c_2$ are distinct concept names.

It is important to note that we assume that $\leq_C$ is not only defined on concepts but also on instances or individuals, i.e. entities of type $e$. Furthermore, it does not seem reasonable to assume the existence of a uniquely named concept for each sense of every noun or verb we consider. In fact, different nouns and verbs share meanings and can be grouped into classes (Levin, 1993). For this reason, in some cases we will assume the existence of a certain concept $c(n)$ without naming it but merely stating that it is directly subsumed by some other concept (compare 9 below). In this sense we will gloss over the details concerning ontological modeling and the appropriate naming of concepts as this issues are out of the scope of this paper.

The requirement that the verb *croaked* needs an animate subject will be expressed by extending the elementary tree description in (5) to the effect that the concept associated with the subject node must be animate, as shown in 9.

(9) $$\begin{aligned}&\forall n_0\,[l(n_0) = iv \wedge croaked(n_0) \rightarrow \\ &\exists n_1 ... n_4 \exists c\,(l(n_1){=}s \wedge l(n_2){=}dp \wedge l(n_3){=}vp \wedge \\ &l(n_4){=}vp \wedge \sigma(n_1) = \sigma(n_3)(\sigma(n_2)) \wedge \\ &\sigma(n_4) = \sigma(n_0) \wedge \Gamma(n_0,n_1,n_2,n_3,n_4) \wedge \\ &c(n_2) \leq_C \textsc{animate} \wedge \\ &((\sigma(n_0) = croak'_{\langle e,t\rangle} \wedge\ c(n_0) \leq_C^+ \textsc{cawing}) \\ &\vee(\sigma(n_0) = croak''_{\langle e,t\rangle} \wedge\ c(n_0) \leq_C^+ \textsc{dying})))]]\end{aligned}$$

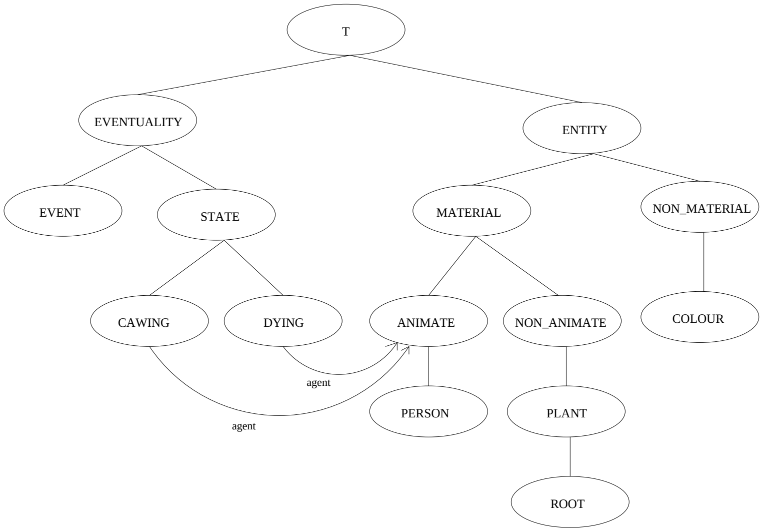


Figure 1: Example Ontology

Thus, when including ontological information about the selectional restrictions of a certain verb, it is not possible anymore to express the general structure of intransitive verbs as in 5. Instead, we need one elementary tree description for each verb or class of verbs.

Figure 2 depicts a compact representation of the two different senses of *croaked* by means of an elementary tree enriched with ontological information. In particular the set notation {CAWING,DYING} is interpreted as a disjunction of concepts the corresponding node can represent.

Furthermore, we also have to pass the ontological information to the nodes dominating a certain lexical element, i.e.

(10)
$$\begin{array}{ccc} n_6\text{:DP}^+\text{:c} & n_8\text{:DP}^+\text{:c} & n_8\text{:DP}^+\text{:c} \\ | & | & | \\ n_7\text{:PN}^-\text{:c} & n_9\text{:CN}^-\text{:c} & n_9\text{:ADJ}^-\text{:c} \end{array}$$

$$\forall n\, [l(n){=}pn{\rightarrow}\exists n'\ (l(n'){=}dp \wedge c(n) = c(n') \wedge \Gamma(n,n'))]$$
$$\forall n\, [l(n){=}cn{\rightarrow}\exists n'\ (l(n'){=}dp \wedge c(n) = c(n') \wedge \Gamma(n,n'))]$$
$$\forall n\, [l(n){=}adj{\rightarrow}\exists n'\ (l(n'){=}dp {\wedge} c(n) = c(n') \wedge \Gamma(n,n'))]$$

When a sentence is parsed, the ontological structure will be added to the axiom set. Together with ontologically enriched descriptions the models for a sentence will be restricted to also meet the ontological requirements that come with the different meanings of its lexical items. In our example 1, as a net result we have reduced the number of readings from 6 to 2, i.e. the ones corresponding to the readings in which an animate being cawed or died. Both readings are compactly represented in figure 3.

## 4 Conclusion

Musken's (Muskens, 2001) Logical Description Grammar (LDG) approach allows to underspecify syntax and semantics in a uniform way in form of logical descriptions. The models of these logical descriptions correspond to possible completions or disambiguations of the underspecified representations. We have extended this approach by including lexical/ontological information and shown how this extension can be applied to the underspecification and resolution of lexical ambiguities. Furthermore, we have described how an appropriate ontological model can be used to constrain the resolution of the types of ambiguities we focus on. In a further paper we shall also show that the approach is applicable to PP-attachment, thematic-role as well as coordination ambiguities. In addition, we also plan to include morphological information in our logical descriptions. Certainly the models yielded are not always unique. This is in line with the fact that the sentential context is not always enough to disambiguate a certain expression and allows to pass on these ambi-

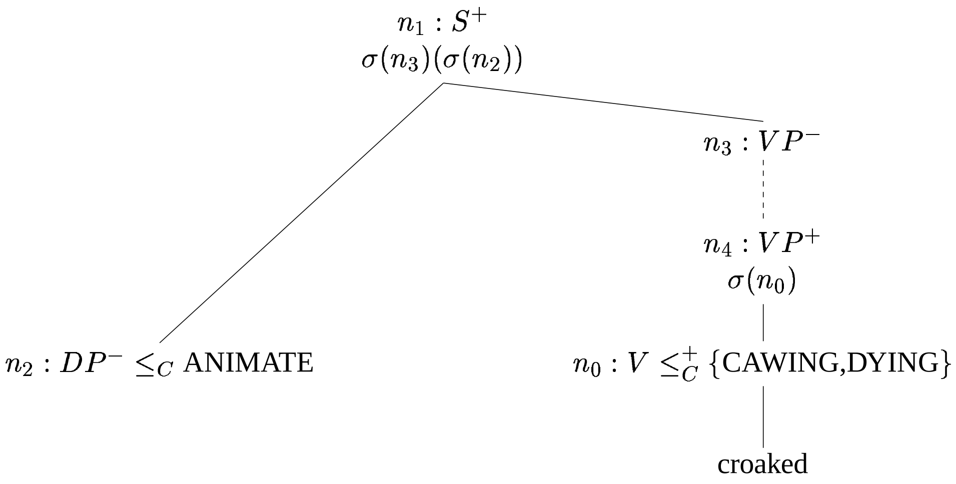

$n_1 : S^+$
$\sigma(n_3)(\sigma(n_2))$
$n_3 : VP^-$
$n_4 : VP^+$
$\sigma(n_0)$
$n_2 : DP^- \leq_C$ ANIMATE
$n_0 : V \leq_C^+$ {CAWING,DYING}
croaked

Figure 2: Compact representation of the two elementary trees of *croaked*.

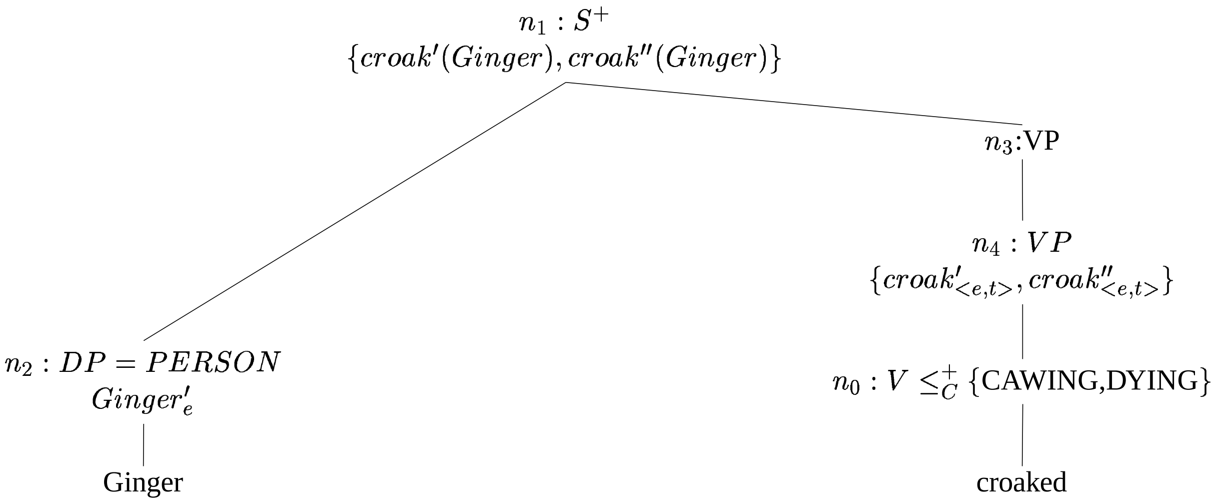

$n_1 : S^+$
$\{croak'(Ginger), croak''(Ginger)\}$
$n_3$:VP
$n_4 : VP$
$\{croak'_{<e,t>}, croak''_{<e,t>}\}$
$n_2 : DP = PERSON$
$Ginger'_e$
$n_0 : V \leq_C^+$ {CAWING,DYING}
Ginger
croaked

Figure 3: Compact representation of the two readings of *Ginger croaked*.

guities to other components where they can be actually resolved. In fact, as already mentioned in the introduction, the LDG approach allows a smooth and straightforward integration of statistical methods as external sources of disambiguation. In this line, it would be an interesting option to consider a statistically trained POS-tagger as for example in (Schmid, 1994) as such an additional and external source of disambiguation and integrate the independently derived POS tags into the logical input description for each sentence.

## References

Johan Bos. 1995. Predicate logic unplugged. In *Proceedings of the Tenth Amsterdam Colloquium*.

P. Buitelaar. 1996. A lexicon for underspecified semantic tagging. In *Proceedings of ANLP96 SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What and How?*

P. Buitelaar. 1998. *CoreLex: Systematic Polysemy and Underspecification*. Ph.D. thesis, Computer Science, Brandeis University.

H. Bunt. 2003. Underspecification in semantic representations: Which technique for what purpose? In Harry Bunt, Ielka van der Sluis, and Roser Morante, editors, *Proceedings of the 5th International Workshop on Computational Semantics*, pages 37–54.

J. Dörre. 1997. Efficient construction of underspecified semantics under massive ambiguity. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*.

E. Bozsak et al. 2002. KAON - Towards a large scale Semantic Web. In *Proceedings of the Third International Conference on E-Commerce and Web Technologies (EC-Web)*. Springer Lecture Notes in Computer Science.

A.K. Joshi and Y. Schabes. 1997. Tree-adjoining grammars. In *Handbook of Formal Languages*, volume 3, pages 69–124. Springer.

B. Levin. 1993. *English Verb Classes and Alternations*. University of Chicago Press.

R. Muskens. 2001. Talking about trees and truth-conditions. *Journal of Logic, Language and Information*, 10(4):417–455.

M. Pinkal. 1996. Radical underspecification. In P. Dekker and M. Stokhof, editors, *Proceedings of the 10th Amsterdam Colloquium*, pages 587–606.

J. Pustejovsky. 1998. The semantics of lexical underspecification. *Folia Linguistica*.

O. Rambow and K. Vijay-Shanker. 1995. D-tree grammars. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*.

U. Reyle and A. Roßdeutscher. 2001. Temporal underspecification in discourse. In C. Rohrer, A. Roßdeutscher, and H. Kamp, editors, *Linguistic Form and its Computation*, chapter 8, pages 255–283. CLSI.

U. Reyle. 1993. Dealing with ambiguities by underspecification: Construction, representation and deduction. *Journal of Semantics*, 10(2):123–179.

U. Reyle. 1995. On reasoning with ambiguities. In *Proceedings of the 6th Meeting of the Association for Computational Linguistics, European Chapter*, pages 1–8.

U. Reyle. 1996. Co-indexing labelled DRSs to represent and reason with ambiguities. In K. van Deemter and S. Peters, editors, *Semantic Ambiguity and Underspecification*. CSLI Publications.

A. Roßdeutscher and U. Reyle. 2000. Constraint-based bottom up discourse interpretation. In U. Reyle, editor, *Semantic Ambiguity and Underspecification*. Arbeitsberichte des Sonderforschungsbereichs 340, Stuttgart/Tübingen, Nr. 164.

M. Schiehlen. 1996. Semantic construction from parse forests. In *Proceedings of the 16th International Conference on Computational Linguistics*.

M. Schiehlen. 1999. *Semantikkonstruktion*. Ph.D. thesis, Universität Stuttgart.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

K. van Deemter. 1991. *On the Composition of Meaning*. Ph.D. thesis, University of Amsterdam.

K. van Deemter. 1995. Towards a logic of ambiguous expressions. In K. van Deemter and S. Peters, editors, *Semantic Ambiguity and Underspecification*, pages 203–237. CSLI Publications.