Structural Parsing

C. Hoede and L. Zhang*

Faculty of Mathematical Sciences
University of Twente
P.O.Box 217
7500 AE Enschede,The Netherlands

Abstract. In the theory of knowledge graphs, words are represented by word graphs. Sentences are to be represented by sentence graphs. This is called structural parsing. Under consideration of the semantic and syntactic features of natural language, both semantic and syntactic word graphs are formed, the latter expressing the function of word types like nouns, verbs, etc.
Traditional grammar rules can be derived from the syntactic word graphs. However, instead of using traditional parsing, based on the grammar rules, to prepare the construction of a sentence graph, we discuss the relationship with utterance paths. As a result, chunk indicators in sentences are proposed to guide structural parsing.

Key words: Knowledge graphs, word graphs, structural parsing.

AMS Subject Classifications: 05C99, 68F99.

1 Introduction

A natural language processing system always contains a parser, which is a device that has a natural language sentence as input string and that produces a representation of the sentence when it is acceptable. Parsing is the process of structuring a representation of a natural language sentence usually in accordance with a given grammar. There are two important points here; one is that we require a representation as an interlingua (intertransmittal language) that is standing between the natural language sentence accepted and its access structure in a computer, the other is that we require grammars with which the natural language acts in accordance. The former point is independent of the specific language, the latter is dependent on the specific language.

We chose knowledge graphs as the interlingua, due to their advantageous properties in natural language processing. Since based on knowledge graph theory, parsing is more special than traditional parsing methods and is called *structural parsing*. The structural parsing that is introduced in this paper aims at transfering the natural language sentence accepted to a sentence graph, which

* on leave from Northwestern University,Xi'an,P.R. China

stands for the structure (or meaning) of this sentence. A sentence graph is built from word graphs, which just stand for meanings of the words contained in this sentence. This means that word graphs are at the base of parsing. Word graphs were already discussed for prepositions, adwords (including adjectives, adverbs and Chinese quantity words) and logic words in three other papers.

This paper will introduce the theory of structural parsing. In Section 2 some basic notions from knowledge graph theory are recapitulated. In Section 3 semantic and syntactic word graphs are introduced. Section 4 introduces structural parsing and the concept of chunks for sentences as well as for graphs is discussed in Section 5. Section 6 gives an example in which only the main steps of structural parsing are mentioned, due to lack of space.

2 Knowledge graph theory

We refer to the papers of Hoede and Li [3], Hoede and Liu [5] and Hoede and Zhang [6] for an introduction to knowledge graphs as far as needed for this paper. We only recall the following.

Words are considered to be representable by directed labeled graphs. The vertices, or tokens, are indicated by squares and represent *somethings*. The arcs have certain types that are considered to represent the relationship between somethings, as recognizable by the mind. The graphs that we will discuss are therefore considered to be subgraphs of a huge *mind graph*, representing the knowledge of a mind and therefore also called *knowledge graph*. These knowledge graphs are very similar to conceptual graphs, but are restricted as far as the number of types of relationship is concerned.

There are two types of relationships. The binary relationships, the usual arcs, may have the following labels:

- EQU : Identity
- SUB : Inclusional part-ofness
- ALI : Alikeness
- DIS : Disparateness
- CAU : Causality
- ORD : Ordering
- PAR : Attribution
- SKO : Informational dependency.

The SKO-relationship is used as a loop to represent universal quantification. Next to the binary relationship there are the n-ary frame-relations. There are four of these.

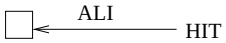
FPAR : Relationship of constituting elements with a concept, being a subgraph of the mind graph.
NEGPAR : Negation of a certain subgraph.
POSPAR : Possibility of a certain subgraph.
NECPAR : Necessity of a certain subgraph.

These four frame relationships generalize the wellknown logical operators. If a certain subgraph of the mind graph is the representation of a wellformed

proposition p, this proposition is represented by the frame, $\neg p$ is represented by the same subgraph framed with the NEGPAR relationship and the modal propositions $\Diamond p$ and $\Box p$ are represented by the same subgraph framed with the POSPAR and the NECPAR relationship respectively. In this way logical systems can be represented by different types of frames of very specific subgraphs. We refer to Van den Berg [2] for a knowledge graph treatment of logical systems.

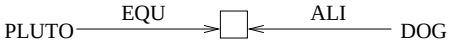
So logic is described by frames of propositions. If a subgraph of the mind graph does not correspond to a proposition the framing, and the representation of the frame by a token, may still take place. Any such frame may be baptized, i.e. labeled with a word. The directed ALI-relationship is used between a word and the token to type the token. Thus

is to be read as "something like a volcano". Note that the token may represent a large subgraph of the mind graph. In particular verbs may have large frame contents. Verbs are represented in the same way. So



is the way the verb HIT is represented.

The directed EQU-relationship is used between a word and a token to valuate or instantiate the token. So



is to be read as " something like a dog equal to Pluto".

The mind graph is considered to be a wordless representation of thought relationships between units of perception. The words come in when certain subgraphs are "framed and named". At the most elementary level the frame contents may just be one relationship. These are the first word graphs to start with. For that reason they formed the first set of word graphs. The frame with contents of frames representing nouns and verbs express the definitions of the concepts

(note that frames do literally take other concepts together). A lexicon of *semantic word graphs* that expresses the meaning of the words is being constructed at the University of Twente. In this paper *syntactic word* graphs will be introduced, as well as parsing by representing a sentence with a knowledge graph, a new field, which we call *structural parsing*.

Definition 1. *Structural parsing is the mapping of a sentence on a semantic sentence graph.*

3 Semantic and syntactic word graphs

We are interested in word graphs in terms of which a sentence will be analyzed. It is fortunate that word graphs were already discussed for prepositions, adwords and logic words as has been mentioned in the introduction. Here, another aspect of word graphs is discussed, namely the semantic and the syntactic representation of a word by word graphs.

3.1 Definitions of syntactic and semantic word graphs

To analyze a sentence to obtain a sentence graph, two pieces of information are necessary; one is the meanings of the words that constitute this sentence, which is called semantic information, the other is the syntax of the words that constitute this sentence, which is called syntactic information. Considering the semantic and the syntactic information of natural language, we develop semantic word graphs for the meanings of words and syntactic word graphs for the syntactic functions of words. Now we give the definitions for semantic and syntactic word graphs.

Definition 2. *A semantic word graph is a word graph, which expresses the meaning of a word.*

The three papers [2], [4] and [5] on word graphs concern semantic word graphs.

Definition 3. *A syntactic word graph is a word graph, which expresses the syntactic functions of a word.*

For example, in Chinese the word "wo3"means "I", and this pronoun has at least the following three usages:

• subject • object • attribute

We can express its functions with three different knowledge graphs as in Figure 1. The word graphs in Figure 1 give the different syntactic functions of the word "wo3". We have chosen to represent the subject and object function by a

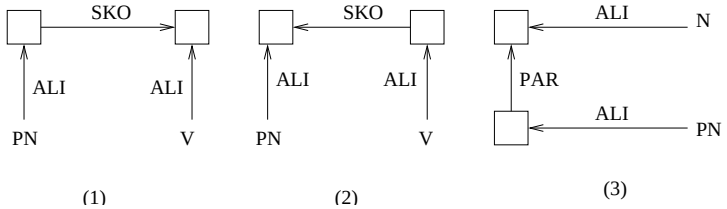


Figure 1. Syntactic word graphs for the pronoun "wo3" corresponding to "I", "me" and "my" in English.

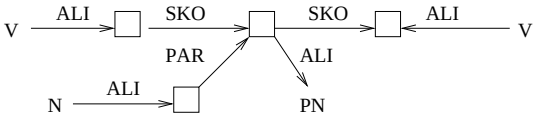


Figure 2. Syntactic word graph for the pronoun "wo3".

SKO-arc, to respectively from a verb V, where on the semantic level we would choose a CAU-arc. The possessive use of "wo3" is expressed by a PAR-arc. We can also express the three syntactic functions with one knowledge graph like in Figure 2, which is called the syntactic word graph of the word "wo3". The meaning of the word "wo3" is expressed by another word graph, which is called the semantic word graph of the word "wo3", see Figure 3.

Consider another Chinese word "ta1", in English "he", "him", or "his". We obtain that the syntactic word graph of the word"ta1" is the same as for the word "wo3". In fact many other words, in this case pronouns, have the same syntactic functions in a sentence, so their functions can be expressed by the same syntactic word graph. The question now is how many different syntactic word graphs there are. This depends on how many word types there are, which will be discussed in the next section. Syntactic word graphs are essentially semantic word graphs for word types. Willems [8] introduced *syntactic graphs*, but took prepositions and functions descriptions like subject or object as labels of arcs. We use the basic ontology of knowledge graphs for all word types, including e.g. prepositions.

3.2 Syntactic word graphs for word types

In Chinese the problem of word types is more complex than in English. There is no Chinese dictionary with word types till now. Therefore, we have chosen to first discuss the types of Chinese words. In English there is no problem, we do not need to reclassify the words.

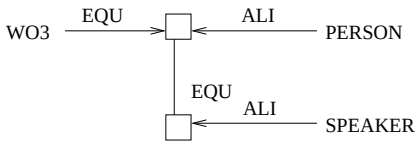


Figure 3. Semantic word graph for "wo3".

Definition 4. *Word types are the types of words, classified in terms of their syntactic functions.*

We chose the main word types as our target to begin structural parsing. In line with our choice to illustrate knowledge graph theory for two rather different languages, for which we chose English and Chinese we classify Chinese words into 8 word types, given in Table I with the terminology in English as well as the symbols that are used in the word graphs.

CHINESE	ENGLISH	SYMBOL
ming2 ci1	noun	N
dong4 ci2	verb	V
xing2 rong2 ci1	adjective	adj
dai4 ci1	pronoun	PN
shu4 ci1	numeral	num
liang4 ci1	classifier	cl
jie4 ci1	preposition	$prep$
fu4 ci1	adverb	adv

Table I: Restricted set of Chinese word types .

In English we also chose 8 word types, but the "classifier"type was replaced by the "determiner"type. We do not give a table nor the word graphs, due to lack of space.

The surface structure of a sentence is to be expressed by its syntactic sentence graph, and the deep structure of a sentence is to be expressed by its semantic sentence graph.

The syntactic word graphs for the 8 word types, given in Table I, can be constructed by expressing the various functions. Noun and verb have the most complicated syntactic word graphs. We refer to an extended preprint of this paper [7] for details. Figure 2 should illustrate the concept of syntactic word graph.

3.3 From syntactic word graphs to traditional grammar

We can derive a traditional grammar, from the syntactic word graphs.

In terms of the word types, rules in a grammar indicate in what order the words can be combined. Such a combination of words should be possible as far as the syntactic word graphs are concerned.

We now shortly describe how grammar rules can be derived from the syntactic word graphs.

The general way to find the rules is to check whether two graphs can be coupled. For example, the PAR-arc from *adj* to N present in the graph for *adj* is found in the graph for N as well. This tells us that the ordered pair *adj* N can occur in a sentence. There should therefore be a rule $X \rightarrow adj\ N$, or, as we prefer because of our aim to develop structural parsing, an inverse rule $adj\ N \rightarrow X$.

X may be chosen to be N as nouns and verbs are the dominant word types in language. Not without reason the first rule of grammar is $S \rightarrow NP\ VP$, where NP can be seen basically as an N and VP as a V, to which various other parts of the sentence graph, that is to be expressed, are added.

Considering all pairs of word types we obtain a set of rules that are such that, when applied in a parsing process, guarantee that the corresponding word graphs, syntactic or semantic , can be coupled.

4 Structural parsing

The goal of structural parsing, the semantic graph of a sentence, is in principle obtained as follows:

- A grammar is used to construct one or more parse trees for the sentence.
- A syntactic sentence graph is derived from syntactic word graphs using a parse tree.
- A semantic sentence graph is derived from the found syntactic sentence graph.

Note that usually many syntactic sentence graphs can be derived by the grammar, but that often only one syntactic graph is suitable semantically, unless there is essential ambiguity.

4.1 A traditional parsing approach

The following procedure could be used:

- In a lexicon for each word a semantic and a syntactic word graph is given for each use of the word.
- The set of grammar rules is used in traditional parsing, which leads to one or more parse trees.
- Syntactic word graphs are combined to a syntactic sentence graph according to bottom-up parsing.
- Each syntactic sentence graph is transformed into a semantic sentence graph by combining corresponding semantic word graphs.

Both traditional bottom-up parsing or top-down parsing can be used to analyze a sentence. One of the most difficult problems for traditional parsing techniques is to get rid of ambiguities. We can produce many syntactic sentence graphs that make no sense, if we use grammars like discussed partly in Section 3.3. We do not like to produce many syntactic sentence graphs for complexity reasons. For this reason, we would like to give our own parsing method, that is adapted to our knowledge graph theory. The key to our new parsing method lies in a discussion of *utterance paths.*

4.2 Utterance paths and chunks

A sentence expresses a sentence graph. The graph is " brought under words ". The speaker chooses an order in which these words are uttered. Corresponding with this order is an ordering of subgraphs of the sentence graph, the word graphs. With such an ordering of subgraphs usually one or more paths can be indicated, depending on whether consecutive words have overlapping word graphs or not. We will make the concept of utterance path clear by an example sentence:

- The volcano, that lies in Alaska, 130 kilometers from Anchorage, erupted in 1992.
- The volcano, that erupted in 1992, lies in Alaska, 130 kilometers from Anchorage.
- 130 kilometers from Anchorage, Alaska, lies the volcano, that erupted in 1992.
- In Alaska, 130 kilometers from Anchorage, lies the volcano, that erupted in 1992.

In all four sentences for one sentence graph we recognize typical paths. "130 kilometers from Anchorage" is one such path. "Erupted in 1992" is another path occurring in all four sentences, and "the volcano, that" also. The ordering "in Alaska" does not occur in the third sentence, but could have been used therein. In the simplified sentence graph in Figure 4 these paths can be read off as texts in the frames.

The remarkable feature is that the six indicated frames, that occur as connected graphs in the non-simplified sentence graph, are expressed as what might be called *chunks* of the sentence, see also the paper of Abney [1] on parsing by chunks. Abney states that people tend to express a sentence in chunks of words, and we see that chunks of the sentence graph are brought under words in some specific order. The four sentences can be described as an ordering of the expressed chunks, 1 to 6:

- $1 \longrightarrow 2 \longrightarrow 4 \longrightarrow 5 \longrightarrow 6 \longrightarrow 3$
- $1 \longrightarrow 2 \longrightarrow 3 \longrightarrow 4 \longrightarrow 5 \longrightarrow 6$

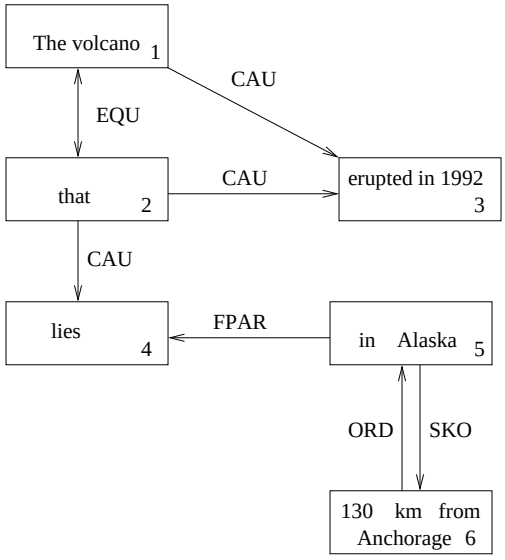


Figure 4. Semantic sentence graph with display of "chunks".

– 6⟶5⟶4⟶1⟶2⟶3
– 5⟶6⟶4⟶1⟶2⟶3 .

Note that "jumps" occur, consecutive chunks, not linked in the sentence graph. In the first sentence there is a jump 6⟶3, in the second sentence a jump 3⟶4, in the third a jump 4⟶1 and in the fourth there are two jumps: 6⟶4 and 4⟶1.

As our goal is to construct the sentence graph from a sentence, the fact that chunks of the graph are expressed as chunks of the sentence leads us to want to read of chunks from the sentence, for which chunks of the sentence graph seem to be easily constructable. A problem for finding chunks of a sentence is that of finding begin point and end point of a chunk. With the interpretation of a sentence, as expressing a sentence graph, as a guide line we will try to find chunk indicators. Note that we only refer to Abney's paper as it presents the idea of chunk. Unlike other papers on parsing by chunks, we do not focus on traditional parsing techniques. In fact, we try to avoid those techniques.

4.3 Chunk indicators

Our reasoning behind the choice of indicators is the following. In terms of knowledge graph theory, frame words, see [6], such as: be, can, may, must, which are auxiliary verbs, and modify the whole sentence, should be a chunk indicator, where the chunk is the whole sentence. For example in the following sentence :

"Can I have a listing of all flights from Amsterdam to Beijing?"

The auxiliary verb "can" modifies the whole sentence. In the sentence graph this is expressed with a POS-frame. We have discussed frame words like BE-frame, NEC-frame, NOT-frame, OR-frame, IF-THEN-frame or POS-frame in [6]. The auxiliary verb " have " is essentially "be with ". The BE-frame can be seen as a chunk indicator too, so that what remains for structural parsing is " I with a listing of all flights from Amsterdam to Beijing ".

Now consider reference words, such as: it, that, the, she, he, her, his, this, ..., etc. They are used to avoid repetition of mentioning something, and hint at a chunk. Consider the sentences

"Every woman thinks she raises children better than her mother", and

"The triangle has a right angle, its sides are 3, 4 and 5 cm, its circumference is 12."

The words "she" and "her" are chunk indicators in the first sentence, the word "its" is a chunk indicator in the second sentence. In Section 4.2, "the volcano" occurred as a chunk. We had the possibility to cut the sentence into two sentences by replacing "that" by "the volcano". Likewise we might replace "it" by "the triangle" and obtain three sentences. These sentences, like all sentences, are clearly chunks.

If two consecutive words can not be combined, they hint at a "jump". Therefore they should belong to different chunks, such as in the following sentence, where the word "up" cannot be combined with "earlier".

"She gets up earlier than John."

Prepositions are very useful in natural language and always link other words. If a preposition is met in a sentence, it hints at a chunk, e.g., in "from Amsterdam to Beijing" or in "in Alaska", see Section 4.2.

Of course comma pairs, in written language, are clearly chunks indicators too, as are pairs of period signs, indicating a whole sentence, or a pair of comma and period sign.

Summing up we list the chunk indicators as follows:

- Indicator 0: Pairs of comma's and/or period signs
- Indicator 1: Auxiliary verbs
- Indicator 2: Reference words
- Indicator 3: "Jumps", with respect to grammar
- Indicator 4: Preposition.

In structural parsing, we do not think complete parse trees are necessary, in fact traditional parse trees may turn out not to be necessary at all. If chunks are recognized, we can give the graphs of these chunks by combining word graphs. After that, we link these *chunk graphs* into a sentence graph.

Of course now there are three problems:

– To what chunks of the sentence do the indicators lead?
– How to make chunk graphs for the found sentence chunks?
– How to link chunk graphs into a sentence graph?

We will not develop a general theory for answering these questions in this paper, but will briefly report on an example. Due to lack of space we have to omit detailed description of the intermediate phases of the process and only give the input and the end result. We refer to an extended preprint of this paper [7] in which details are given.

4.4 An example of structural parsing

We will give input and output of the procedure.

Example sentence

"The volcano, that lies in Alaska, 130 kilometers from Anchorage, erupted in 1992."

The preparatory phase contains two parts.

First, we chunk the sentence by checking indicators, which were discussed in Section 4.3, one by one.

– According to indicator 0, comma's and period signs, we get four chunks directly. Next we cut chunks into sub-chunks according to the other indicators.
– We do not use indicator 1, as there is no auxiliary verb in this sentence.
– The indicator 2 is about reference words. There are two reference words, "the" and "that". A determiner combines with the noun following. "The volcano" is therefore a "complete" chunk, there are no sub-chunks. Other reference words, like pronouns, are separate chunks: "that" is a sub-chunk.
– As for the indicator 3, there are three jumps; between "lies" and "in", "kilometers" and "from", as well as "erupted" and "in". These jumps cut sub-chunks into smaller sub-chunks.
– There are three prepositions, "in" , "from" and "in". Prepositions combine with the noun following. This takes into account indicator 4.
– As there are no further chunk indicators, there is no further chunking.

We get in this way the resulting chunks and sub-chunks:

1. [.[*The volcano*],	CHUNK 1
2. [[*that*][*lies*][*inAlaska*]],	CHUNKS 2, 3 and 4
3. [[130 *kilometer*][*from Anchorage*]],	CHUNKS 5 and 6
4. [[*erupted*][*in* 1992].]	CHUNKS 7 and 8.

Second, for all the words in this sentence, semantic as well as syntactic word graphs should be listed in a lexicon. Since the syntactic word graphs have not been listed, we indicate them with word types abbreviations. The input was the lexicon of Figure 5.

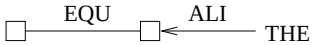
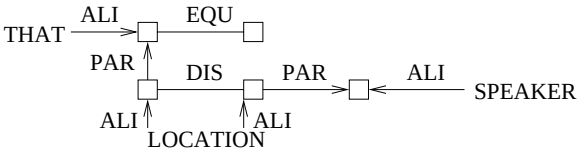
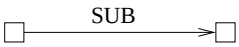
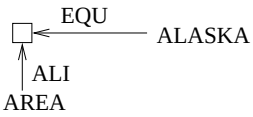
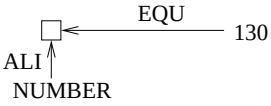
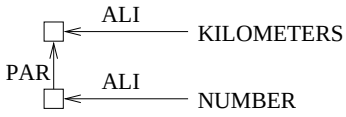
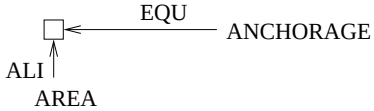
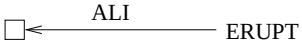
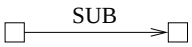
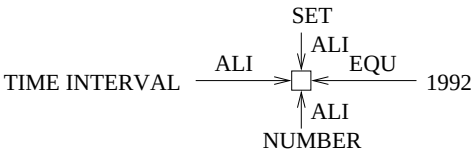
Words	Semantic Word Graphs	Word Types
THE		det
VOLCANO		N
THAT		PN
LY		V
IN		prep
ALASKA		N
130		num
KILOMETERS		N
FROM		prep
ANCHORAGE		N
ERUPT		V
IN		prep
1992		N

Figure 5. Lexicon of the example.

In the main procedure we construct syntactic chunk graphs chunk by chunk. Then we transform them into semantic chunk graphs. The final procedure in structural parsing is to link the semantic chunk graphs and obtain the semantic sentence graph. The output is Figure 6.

To understand the idea of structural parsing the reader should try to recognize the semantic word graphs of the lexicon in the semantic sentence graph.

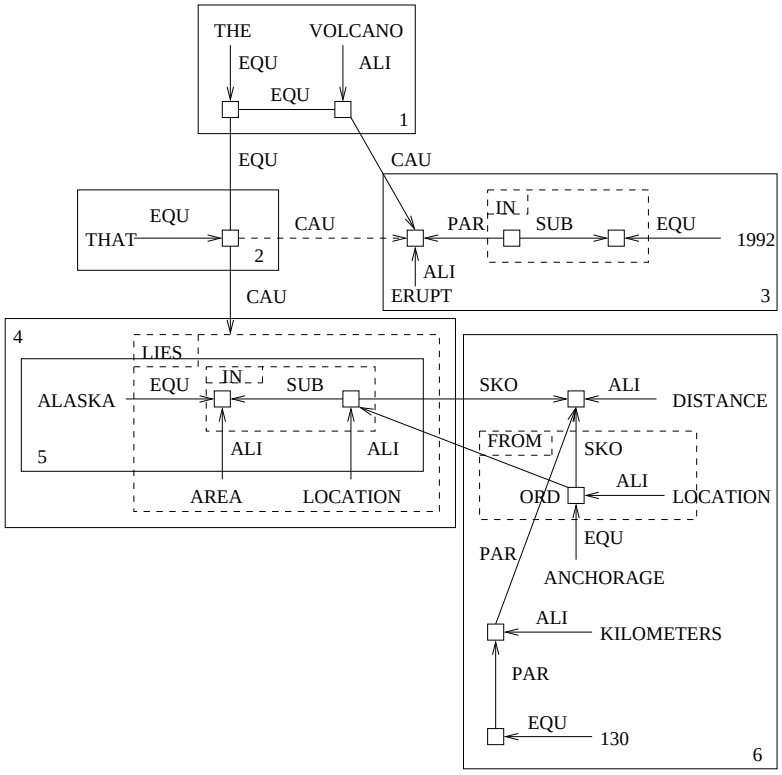


Figure 6. The semantic sentence graph for the example sentence.

5 Conclusion

The standard way of parsing is to find a generation of a sentence by a grammar. Traditional parsing focuses on the syntactic aspects of language and then faces the problem of dealing with semantics. Truth conditional semantics involves a

comparison with a model, i.e. the sentence statement is interpreted within a model.

In knowledge graph theory the approach is from the side of semantics. The meaning of a word or a sentence is considered to be a graph, i.e. the structure is the meaning. Structural parsing then is the mapping of a sentence on a graph. This starts with word graphs for the words, that are to be combined into a sentence graph. The new concept introduced is that of a syntactic word graph for a certain type of word. The word graphs originally considered in the theory are called semantic word graphs. From the syntactic word graphs traditional grammars can be derived.

Sentences are uttered in "chunks", for which a traditionally flavoured theory was designed by Abney [1]. By investigating socalled "utterance paths", we find that parts of the semantic sentence graph are brought under words in such a way that "chunks" of the graph are expressed. This led to the idea that chunks of a sentence have corresponding chunks of the graph. The graph structure suggests certain indicators for chunks in the sentence. With these indicators an example sentence was investigated. The idea turned out to be quite fruitful both for Chinese and English sentences.

References

[1] Abney, S. P. , Parsing by chunks, in *Principle-Based Parsing* (R. Berwick, S. Abney and C. Tenny, eds.), Kluwer Academic Publishers, (1991).

[2] Berg, H. van den, *Knowledge Graphs and Logic: One of Two Kinds*, Dissertation, University of Twente, The Netherlands, ISBN90–9006360–9 (1993).

[3] Hoede, C. and X. Li, Word Graphs: The First Set, in *Conceptual Structures: Knowledge Representation as Interlingua*, Auxiliary Proceedings of the Fourth International Conference on Conceptual Structures, Bondi Beach, Sydney, Australia (P. W. Eklund, G. Ellis and G. Mann, eds.), ICCS'96, (1996) 81–93.

[4] Hoede, C. , X. Li, X. Liu and L. Zhang, *Knowledge Graph Analysis of Some Particular Problems in The Semantics of Chinese,* Memorandum nr. 1516, Faculty of Mathematical Sciences, University of Twente, ISSN 0169-2690 , (February 2000).

[5] Hoede, C. and X. Liu, Word Graphs: The Second Set, in *Conceptual Structures: Theory, Tools and Applications*, Proceedings of the 6th. International Conference on Conceptual Structures, Montpellier, ICCS'98 (M.-L. Mugnier, M. Chein, eds.) Springer Lecture Notes in Artificial Intelligence 1453, (1998) 375–389.

[6] Hoede, C. and L. Zhang, *Word Graphs: The Third Set,* Memorandum nr. 1526, Faculty of Mathematical Sciences, University of Twente, ISSN 0169-2690, (May 2000).

[7] Hoede, C. and L. Zhang, *Structural Parsing*, Memorandum nr. 1527, Faculty of Mathematical Sciences, University of Twente, ISSN 0169-2690, (May 2000).

[8] Willems, M. *Chemistry of Language*, Dissertation, University of Twente, The Netherlands, ISBN 90-9005672-6, (1993).