Faculty of Mathematical Sciences

University of Twente

University for Technical and Social Sciences

P.O. Box 217
7500 AE Enschede
The Netherlands
Phone: +31-53-4893400
Fax: +31-53-4893114
Email: memo@math.utwente.nl

Memorandum No. 1527

Structural parsing

C. Hoede and L. Zhang[1]

May 2000

ISSN 0169-2690

[1]Northwestern University, Xi'an, P.R. China

Structural Parsing

C. Hoede and L. Zhang*

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

Abstract

Parsing is an essential part of natural language processing. In this paper, structural parsing, which is based on the theory of knowledge graphs, is introduced. Under consideration of the semantic and syntactic features of natural language, both semantic and syntactic word graphs are formed. Grammar rules are derived from the syntactic word graphs. Due to the distinctions between Chinese and English, the grammar rules are given for the Chinese version and the English version of syntactic word graphs respectively. By traditional parsing a parse tree can then be given for a sentence, that can be used to map the sentence on a sentence graph. This is called structural parsing. The relationship with utterance paths is discussed. As a result, chunk indicators 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 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.

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

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 discuss the theory of structural parsing, and then take Chinese and English sentences as examples to support our theory. In Section 2 some basic notions from knowledge graph theory are recapitulated. In Section 3 semantic and syntactic word graphs are introduced and the grammars for English and Chinese, based on them, are given in Section 4. Section 5 contains some remarks on traditional parsing. Section 6 is the central section and contains the discussion of utterance paths and their relationship with parsing by chunks, which is our approach to structural parsing. Some examples are discussed.

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:

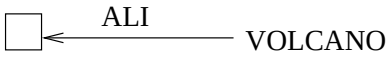
- EQU : Identity
- SUB : Inclusional part-ofness
- ALI : Alikeness
- DIS : Disparateness
- CAU : Causality
- ORD : Ordering
- PAR : Attribution
- SKO : Informational dependency.

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.

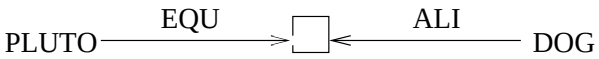
- POSPAR : Possibility of a certain subgraph.
- NECPAR : Necessity of a certain subgraph.

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.

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

The three papers [3], [5] and [6] on word graphs concern semantic word graphs.

Definition 2 *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.

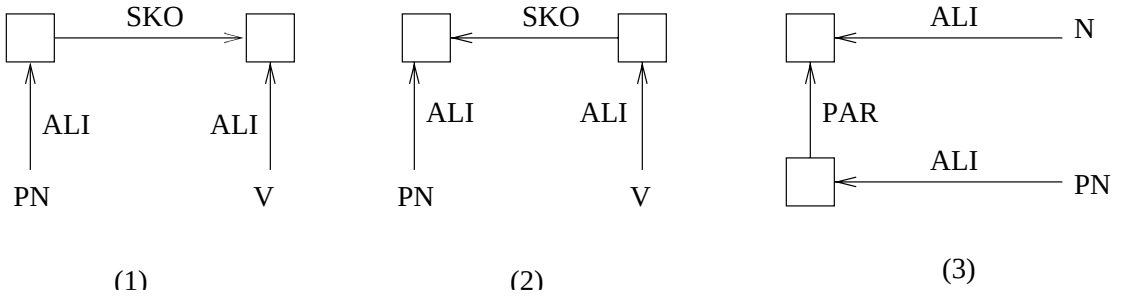


Figure 1: Syntactic word graphs for the pronoun "wo3" corresponding to "I", "me" and "my" in English.

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 SKO-arc, to respectively from

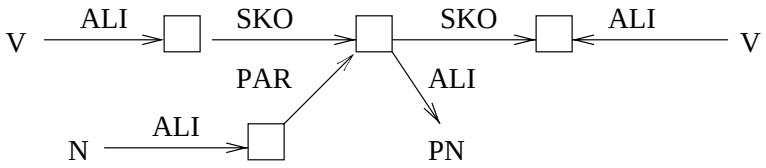


Figure 2: Syntactic word graph for the pronoun "wo3".

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.

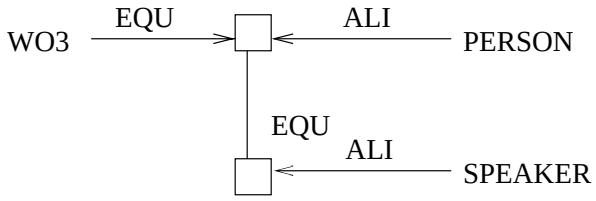


Figure 3: Semantic word graph for "wo3".

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 [7] introduced *syntactic graphs*, but took prepositions and function 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 Word types for Chinese and English

In Chinese the problem of word types is more complex than in English. There is no Chinese dictionary with word types till now. But if there are no word types in our lexicon, the intended structural parsing is impossible. Therefore, it is necessary that we first classify the types of Chinese words. In English there is no problem, we do not need to reclassify the words.

Definition 3 *Word types are the types of words, classified in terms of their syntactic functions.*

It is a problem how to divide Chinese word into types. One of the reasons is that Chinese words show no change of shape feature of words, the type of a word always changes according to the context. One word may have several word types. Let us consider the following two Chinese sentences:

- Ta1 you2 yong3 le.
  (He swam, literally "He swim past time".)
- You2 yong3 you3 yi4 yu2 jian4 kang1.
  (Swimming is good for the health, literally "Swim has good for health".)

The Chinese word underlined, "you2 yong3", is the same in these two sentences, but has different word types. In the sentence"Ta1 you2 yong3 le", the word "you2 yong3"is a verb, the word "le"indicates that the sentence is in past tense. In the sentence"You2 yong3 you3 yi4 yu2 jian4 kang1", the word "you2 yong3"is a noun. There is no change in word shape. What about English?Just look at the same sentences. There is the word "swam"to express the past tense and another word"swimming"to be recognised as a noun.
Although no dictionary mentions word types, there are a lot of views about how to classify Chinese words. According to D.X. Zhu, a Chinese linguist, Chinese words are to be classified into 22 word types[9]. We will list them representing Chinese word types with pin1yin1, spelling, and also give the names of these types in English.

CHINESE	ENGLISH
ming2 ci1	noun
dong4 ci1	verb
xing2 rong2 ci1	adjective
dai4 ci1	pronoun
shu4 ci1	numeral
liang4 ci1	classifier
shi2 jian1 ci1	time
zhu4 ci1	auxiliary
jie4 ci1	preposition
fu4 ci1	adverb
zhuang4 tai4 ci1	state
lian2 ci1	conjunction
fang1 wei4 ci1	direction
yu3 qi4 ci1	tone
tan4 ci1	interjection
chu4 suo3 ci1	location
xiang4 sheng1 ci1	onomatopoeia
qu1 bie2 ci1	comparative
qian2 zhui4	prefix
hou4 zhui4	suffix
biao1 dian3 fu2 hao4	mark
ci4 su4	morpheme

Table I : Classification of Chinese words according to D. X. Zhu

The types listed here are based on the syntactic functions of words. For example, on the one hand, in the sentence " Ta1 zuo2 tian1 lai2 le (He came yesterday.) ", the word " zuo2 tian1 (yesterday) " here modifies a verb " lai2 (come) ", and looks like an adverb syntactically. On the other hand, in the sentence " zuo2 tian1 shi4 qing2 tian1 (Yesterday was sunny) ", the

same word " zuo2 tian1 (yesterday) " looks like a noun syntactically. For this reason it is neither an adverb nor a noun, it is considered by Zhu to belong to the type of " time ".
Based on the 22 word types of Zhu, 71 word types were classified by T. S. Yao [8]. He reclassified each of the above 22 word types into a lot of subtypes, because he has to pay more attention to the word types semantically in order to make natural language processing possible.
In our theory we represent semantic and syntactic features of natural language separately by forming semantic word graphs for semantic features of a word and syntactic word graphs for syntactic features of a word. Typically semantic aspects, like time or location, should not occur on the syntactic level of word types.

When we classify word types, our purpose is to build a grammar that is independent of semantic features and dependent only on syntax. Due to this, we do not need to mention the semantic aspect of a word and can just concentrate on syntax. This is why in principle we agree with the types of Table I that refer to the syntactic view to classify word types. Due to this we can reduce the bigger set of word types into a smaller one.
From our view, as a first reduction, we do not think a "mark", such as "!", "?", "...", etc. should be a type of word, although they play a very important role in expressing meaning. This belongs to another problem area that we call sentence patterns. We will discuss sentence patterns in another paper.
Secondly, we know that prefix, suffix as well as morpheme are very important linguistic concepts, especially in English. However, they belong to the word building problem. We leave this for an independent paper too.
Thirdly, conjunction is always used to describe a connection between sentence and sentence, playing a key role in reasoning. Here we just consider simple sentences, so we leave it to further research.
Fourthly, tone words are very interesting ones. They are very small words and are only used in Chinese. They have at least four usages. Let us give the four usages and example words, because they are unique features in Chinese. We cannot give corresponding English words.
Tone words are used for :

- statement : such as "de", "le", "ni", "ba le", "a"
- question : such as "ma", "ni", "a"
- suggestion : such as "ba", "le", "a"
- exclamation : such as "le", "a" .

Now we give four sentences to explain these distinctive usages. We also try to express each sentence in English, according to the association with the tone word.

- Wo3 men2 qu4 you2 yong3 le. (We went to swim.)
- Wo3 men2 qu4 you2 yong3 ma? (Do we go for a swim?)
- Wo3 men2 qu4 you2 yong3 ba. (Let's go swimming.)
- Wo3 men2 qu4 you2 yong3 la! (It's very nice that we are going to swim!) .

Note that all the tone words happen to appear at the end of a sentence, so that there is no problem in parsing. Tone words can be cut off from the sentence. As for distinguishing the meaning, the semantic word graphs can express this. Also note that e.g. "ma" can both function as mark and as a tone word.
Finally, we want to make structural parsing theory as clear and concise as possible. We therefore start with the simplest and basic situation. The main idea is to first give the main word types and build a grammar based on these word types, and later expand our system by refining the word types and the grammar rules in order to process more complex sentences. Because we would not like our system to be too complicated to work with at the beginning, we just chose the main word types as our target to begin structural parsing.
Therefore we classify Chinese words into 8 word types, given in Table II 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 II : Restricted set of Chinese word types .

In English we also chose 8 word types, but the "classifier"type is replaced by the "determiner"type. We do not give a table.

3.3 Syntactic word graphs for word types

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. Our further analysis will be based on 8 syntactic word graphs for the 8 word types given above. In syntactic analysis, they will be combined to construct the surface structure of a sentence. The Figures 4 and 5 give the syntactic word graphs for the 8 word types given in Table II. These graphs are constructed by expressing the various functions. If, for example, a classifier classifies a noun, see the graph for cl, then in the graph for the noun N the used PAR-arc should also be included. Noun and verb have the most complicated syntactic word graphs. Note that in the syntactic word graph for a preposition the type of arc is indicated by T, as there are several possibilities for the way that arcs link nouns, also see [3].

4 Grammars for Chinese and English

We try to derive the grammar, for English and Chinese respectively, from the syntactic word graphs, given in Section 3.

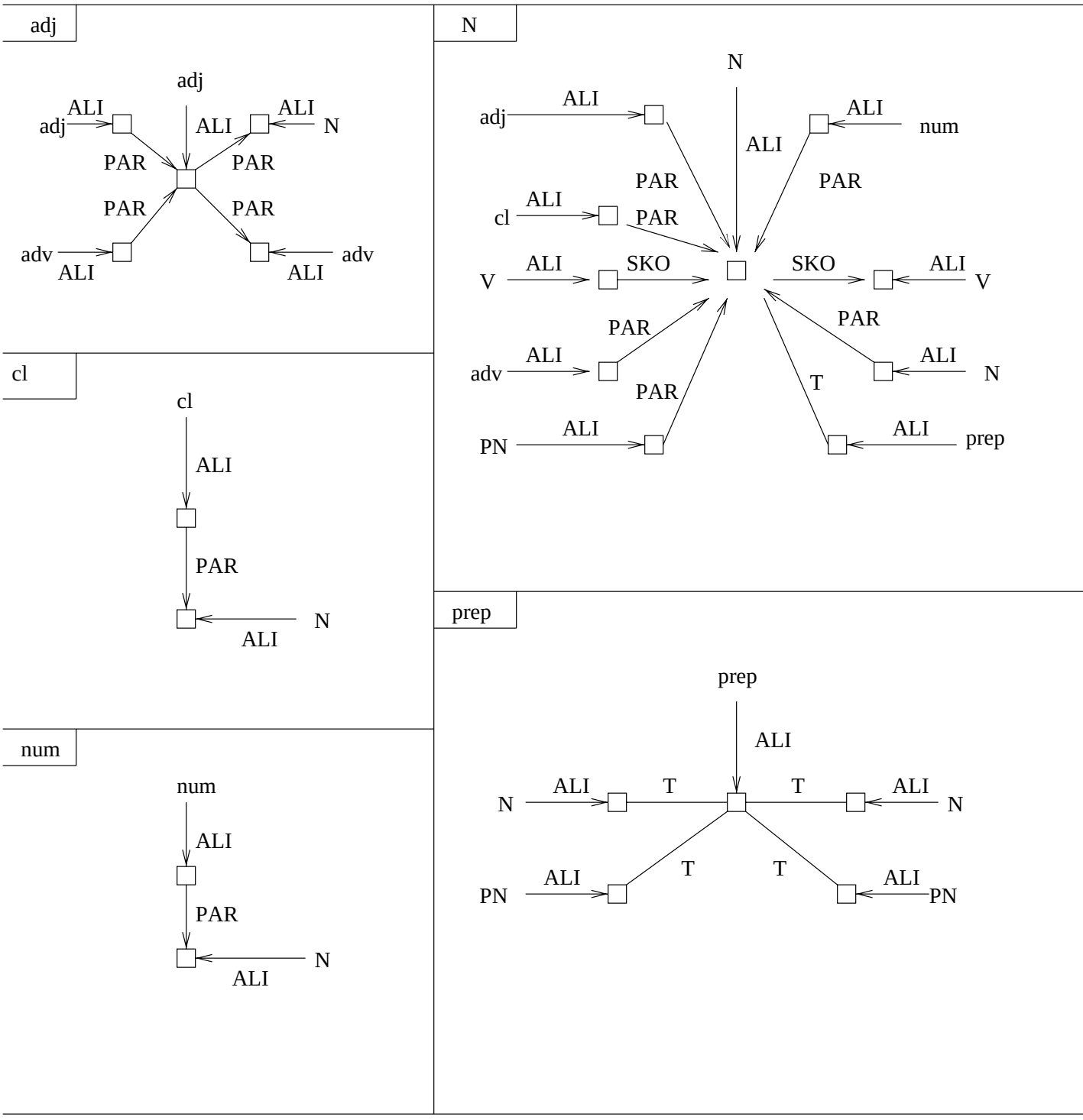


Figure 4: Syntactic word graphs for 5 word types of Chinese.

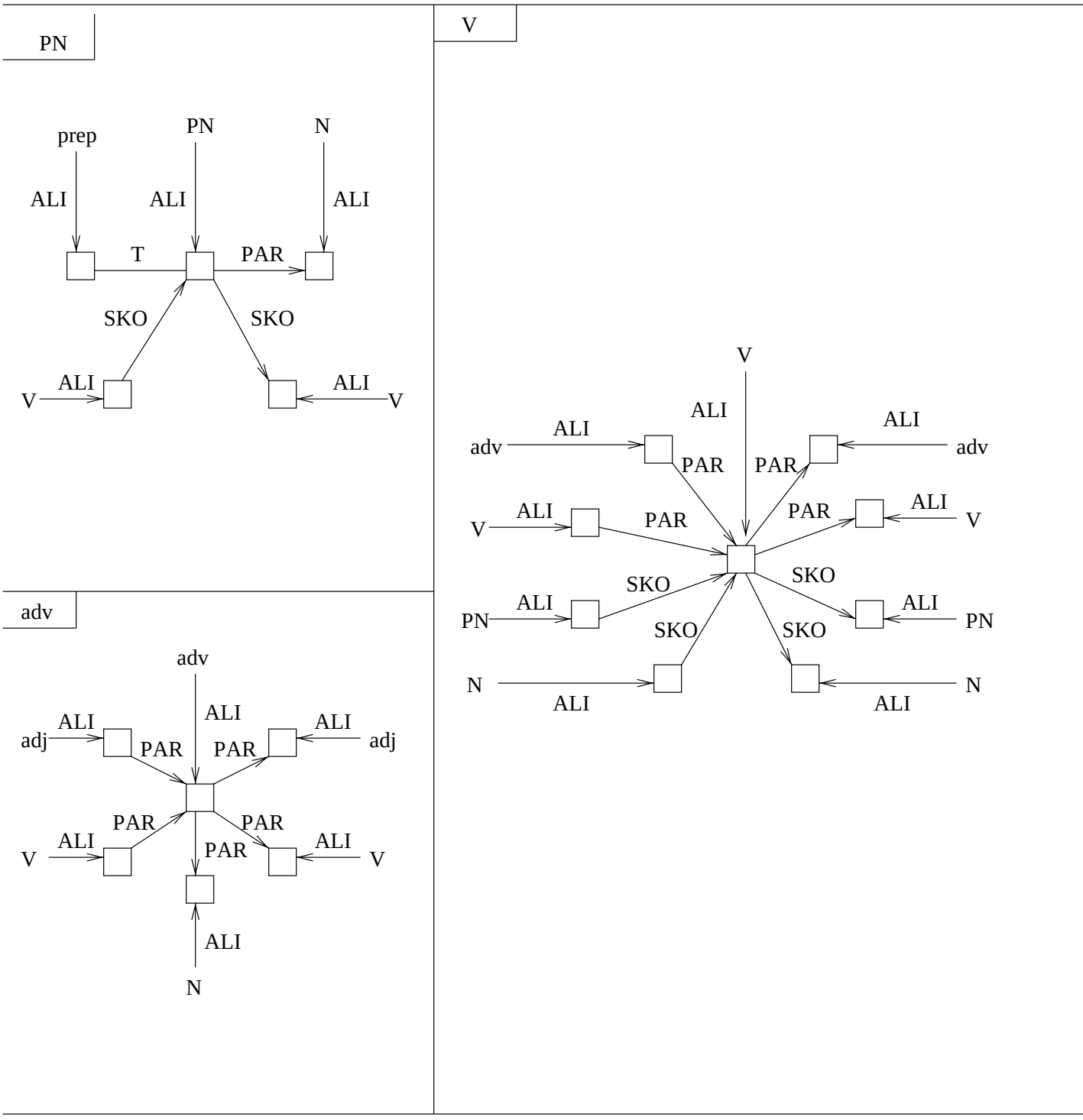


Figure 5: Syntactic word graphs for 3 word types of Chinese.

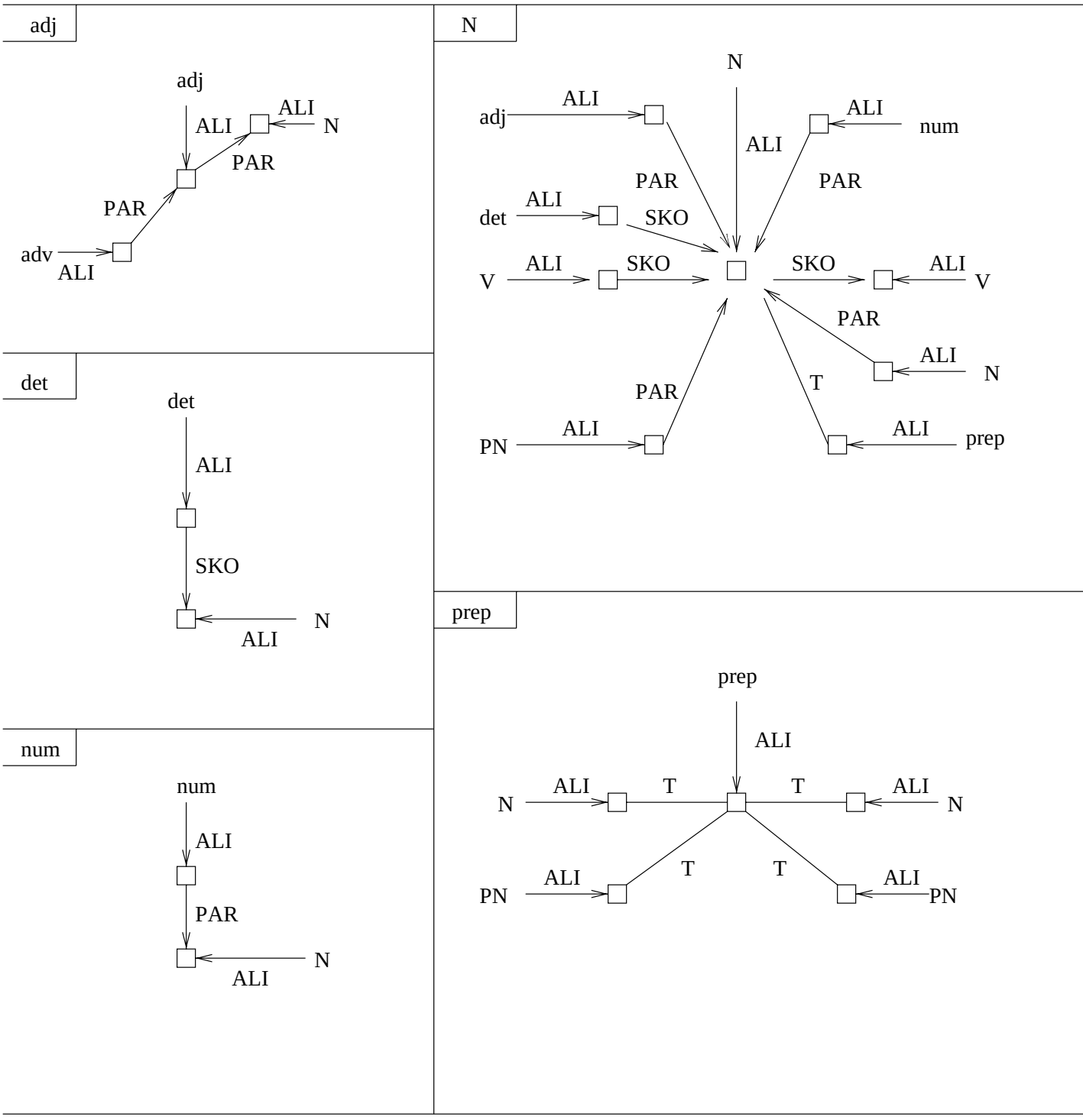


Figure 6: Syntactic word graphs for 5 word types of English.

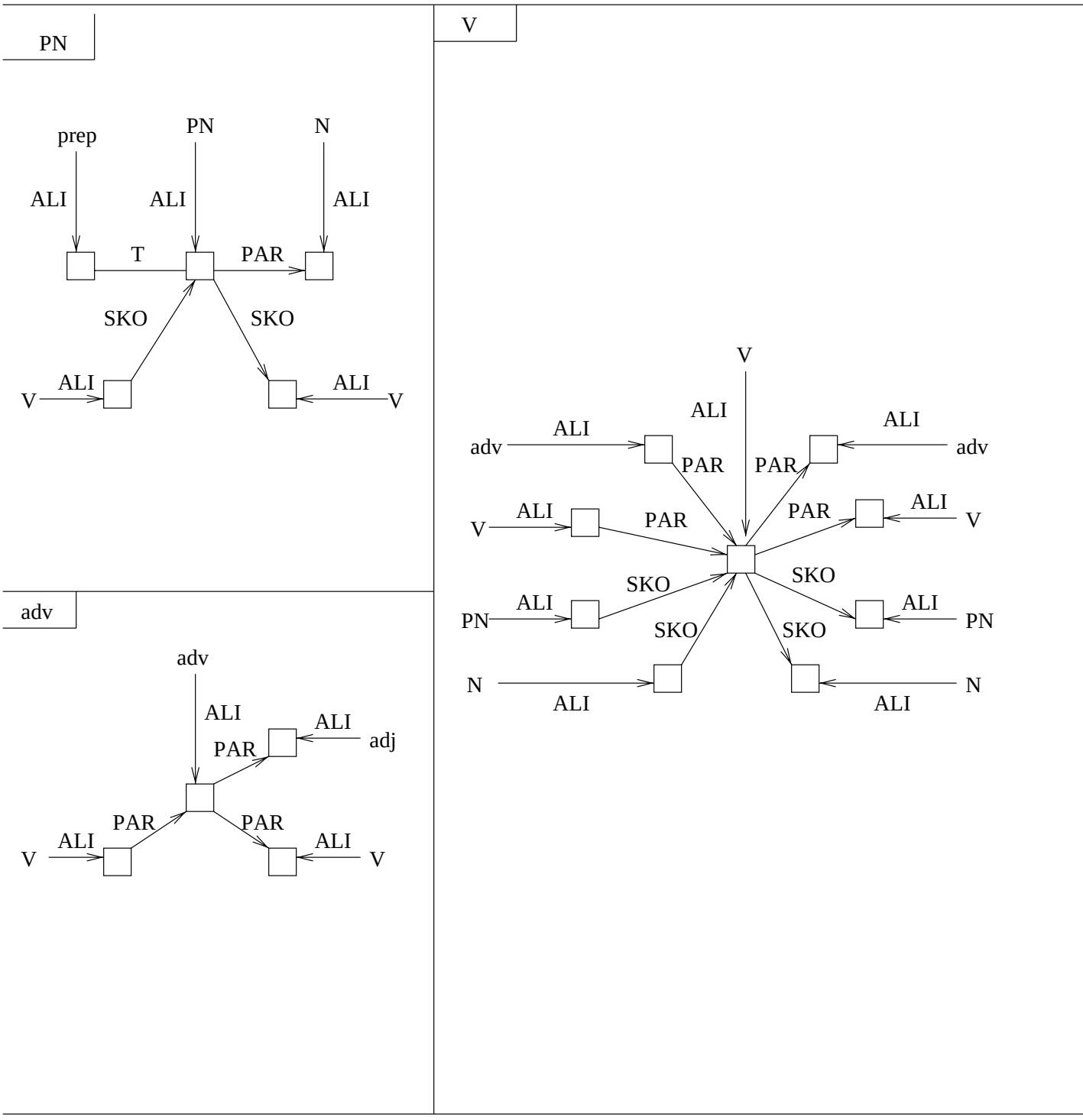


Figure 7: Syntactic word graphs for 3 word types of English.

There were 8 word types in Chinese and 8 word types in English, which were given according to the syntactic functions of words. 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.
First consider Chinese. The above 8 word types have many ways to be combined in a sentence. It is obvious that a noun can be combined with an adjective, a classifier, a numeral or a preposition, such as "hong2 hua2 (red flower)", "yi4 ke1 shu4 (a tree)", in which ke1 is a classifier , "san1 tian1 shi2 jian1 (three days)" or "zai4 jiao4 shi4 li3 (in the classroom)". Also a noun can be combined with a verb as a subject or an object. The following sentences are examples:

Xue2	you2 yong3	hen3	rong2 yi4
V	*N*	*adv*	*adj*
(Learn	swim	very	easy)

Xia4 wu3	lai2	zhun3	xing2	
N	*verb*	*adv*	*adj*	
(Afternoon	come	probably	well)	.

Here the verb"xue2" combines with the noun"you2 yong3" in the first sentence, and the noun "xia4 wu3" combines with the verb "lai2" in the second sentence.
So there may be grammar rules where on the right hand side we read $V \;\; N$ or $N \;\; V$.
Consider a verb, it also has many ways to combine with other words, such as an adverb, an adjective, etc.
Take the following sentence as an example:

Ta1	gang1	zou3.
PN	*adv*	*V*
(He	just	go)

The adv "gang1" combines with the verb "zou3". This example shows the possibility of a rule with $adv \;\; V$ on the right hand side.
Sometimes a noun also can be combined with an adverb. Together they can play the role of a predicate in a sentence. For example,

Ta1	cai2	shi2 liu4 sui4.
PN	*adv*	*N*
(He	only	16 years age)

In the sentence the adverb "cai2" combines with the noun "shi2 liu4 sui4". No verb is mentioned, a rather common feature in Chinese, see [4].
We now describe the grammar rules that we can derive from the syntactic word graphs for Chinese. Note that they are considerably different from the syntactic word graphs for English as we have to take into account very specific forms of sentence building, like in the last example where no verb is mentioned. To enable the generation of such a sentence we need a production rule like $V \rightarrow N$. In "He baker", considered non-wellformed in English, we see a pronoun followed by a noun. From $S \rightarrow NP \;\; VP$ we can obtain this sentence by using $NP \rightarrow PN$, $VP \rightarrow V$ and $V \rightarrow N$ as rules. We decided to include the rule $VP \rightarrow N$.

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 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 is 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.
Special rules are needed to generate NP, VP and also PP, the prepositional phrase. When considering the type N and V in an inverse rule $N\ V \rightarrow Y1$ or $V\ N \rightarrow Y2$, for Chinese, we may have NP, VP or S for $Y1$ and $Y2$.
In [4] it is pleaded to consider adjectives, adverb and classifiers as special cases of *adwords*.
A prepositional phrase can be attached to a verb and should perhaps better then be called an *adwordial phrase*, irrespective of the type of the word or partial sentence to which the phrase is attached. In our grammar we will use the symbol AP.
An AP is sometimes used in a sentence like a noun, and sometimes like a verb. This point is explained in the following sentence shown.

Wang3 xue2 xiao4 men2 kou3	zou3	de na4 ge ren2	shi wo3 peng2 you3 .
AP	V	N	VP

(The man walking to the school gate is my friend.)

The underlined parts " wang3 xue2 xiao4 men2 kou3 (toward the school gate) "and " zou3 (walk) " are the combination of an AP and a V. The whole part is considered to function as a noun in this sentence due to the use of the word"de"(of).
This combination can also be used like a verb phrase :

Ta1	wang3 xue2 xiao4 men2 kou3	zou3 le.
PN	AP	V

(He was walking to the school gate.)

The underlined part that is the combination of an AP and a verb is used like a verb phrase here.
We now list the rules, in standard order:

1. $S \longrightarrow NP\ VP$
2. $NP \longrightarrow PN$
3. $NP \longrightarrow N$
4. $N \longrightarrow adj\ N$
5. $N \longrightarrow cl\ N$

6. $N \longrightarrow num\ N$
7. $N \longrightarrow adv\ N$
8. $N \longrightarrow PN\ N$
9. $N \longrightarrow N\ N$
10. $NP \longrightarrow V\ N$
11. $NP \longrightarrow N\ V$
12. $NP \longrightarrow V\ PN$
13. $NP \longrightarrow PN\ V$
14. $AP \longrightarrow prep\ N$
15. $AP \longrightarrow prep\ PN$
16. $VP \longrightarrow V$
17. $VP \longrightarrow adj$
18. $VP \longrightarrow N$
19. $V \longrightarrow adv\ V$
20. $V \longrightarrow V\ adv$
21. $V \longrightarrow AP\ V$
22. $V \longrightarrow V\ AP$
23. $V \longrightarrow V\ V$
24. $VP \longrightarrow V\ N$
25. $VP \longrightarrow N\ V$
26. $VP \longrightarrow V\ PN$
27. $VP \longrightarrow PN\ V$
28. $adj \longrightarrow adj\ adj$
29. $adj \longrightarrow adv\ adj$
30. $adj \longrightarrow adj\ adv$.

In English there is no word type of "classifier" that is a particular word type in Chinese. Determiners, like "the" and "a", are more often used to modify a noun in an English sentence. So, we replace the classifier type by the determiner type in the set of English word types. Based on the 8 word types that we found, in an analogous way, the following English grammar rules were found:

1. $S \longrightarrow NP\ VP$
2. $NP \longrightarrow PN$
3. $NP \longrightarrow N$
4. $N \longrightarrow N\ N$
5. $N \longrightarrow adj\ N$
6. $N \longrightarrow det\ N$
7. $AP \longrightarrow prep\ N$
8. $N \longrightarrow num\ N$
9. $N \longrightarrow PN\ N$
10. $AP \longrightarrow prep\ PN$
11. $VP \longrightarrow V$
12. $VP \longrightarrow V\ N$
13. $VP \longrightarrow V\ PN$
14. $V \longrightarrow V\ V$
15. $V \longrightarrow adv\ V$
16. $V \longrightarrow V\ adv$
17. $V \longrightarrow V\ AP$
18. $adj \longrightarrow adv\ adj$.

Note that we had to use more rules for Chinese. The main reason is that certain sentences, that would not be considered correct English sentences, had to be parsable in Chinese.
As a final example, an adjective or an adverb can never function as a predicate in English. Consider the following sentences in Chinese :

Tian1	gao1,	lu4	yuan3,
N	*adj*	*N*	*adj*
(Sky	high,	road	far)

Ta1	hen3	gao1.
PN	*adv*	*adj*
(He	very	tall) .

The underlined parts are respectively an adjective " gao1 (high) " or " yuan3 (far) ", or the combination of an adverb " hen3 (very) " and an adjective " gao1 (tall) ". In both cases they are used like a verb.

5 Aspects of parsing

To parse a string, a sentence, according to a grammar, means to reconstruct the parse tree that indicates how the given string can be produced from the given grammar.
The parse tree is a basic connection between a sentence and the grammar from which the sentence can be derived. To reconstruct the parse tree corresponding to a sentence needs a parsing technique. There are dozens of parsing techniques, only two basic types are reviewed in this section, one is bottom-up parsing, the other is top-down parsing.
Also, two search techniques, depth-first search and breadth-first search, are mentioned in this section.

5.1 Top-down parsing

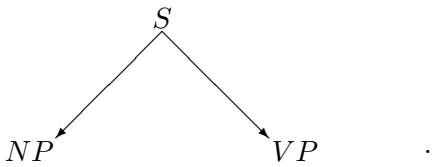
In top-down parsing, we start with the start symbol S and try to deduce the input sentence by constructing the parse tree, which describes how the grammar was used to produce the sentence.
Suppose we have the following simple grammar for natural language, and suppose the sentence is "He hits the dog".

$$
\begin{array}{l}
S \longrightarrow NP\ VP \\
NP \longrightarrow the\ N \\
NP \longrightarrow PN \\
VP \longrightarrow V \\
VP \longrightarrow V\ NP \\
N \longrightarrow dog \\
PN \longrightarrow he \\
V \longrightarrow hit \qquad .
\end{array}
$$

First we try the top-down parsing method.
We know that the production tree must start with the start symbol:

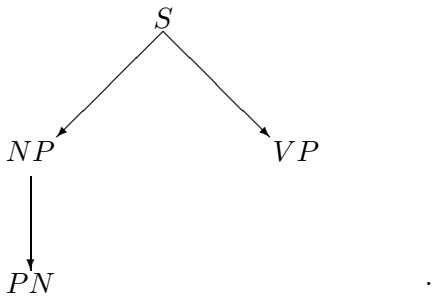
S

We only have one rule for S : $S \longrightarrow NP\ VP$:

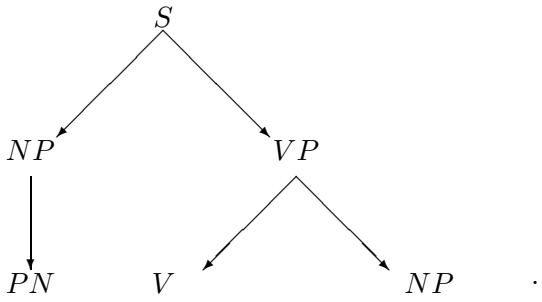


.

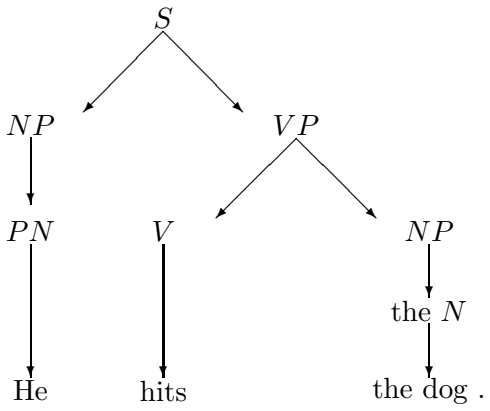
We have two rules for NP: $NP \longrightarrow the\ N$ and $NP \longrightarrow PN$. The first rule would require "the noun" for some noun, the second rule would require a "pronoun"; this leads to the choice of applying the second rule and we obtain:



Again two rules may be applied for VP: $VP \longrightarrow V$ and $VP \longrightarrow verb\ NP$. The second one is fit for this sentence:



We continue this process by applying the first rule for NP and the sentence is deduced by substituting the actual words:



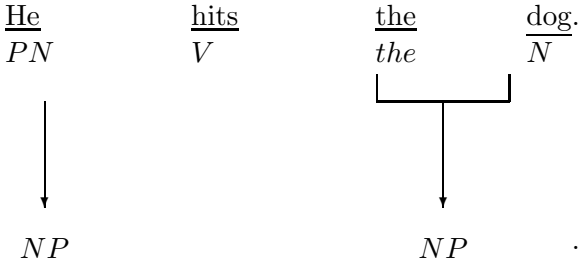
Top-down parsing tends to identify the production rules in prefix order, in which a sentence is deduced by using production rules from the left-hand side to the right-hand side. Note that we have to choose the proper rules to reach our goal. There is a search problem.

5.2 Bottom-up parsing

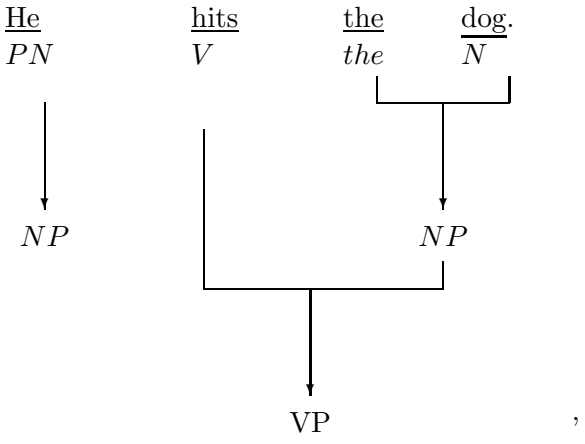
In bottom-up parsing, we start with the sentence as the input string and try to reduce it to the start symbol that is usually expressed by symbol S. Here the keyword is *reduce*. We reduce the input (sentence) to the substring (segment) that is the result of the last step by applying an inverse rule of grammar in postfix order. When we find that the right-hand side of a rule can match with a segment, we replace the segment with the left-hand side of the rule and repeat the process, until only the start symbol is left.
Suppose we have the same grammar as above and suppose the sentence is also " He hits the dog". Now we try the bottom-up parsing method.
The first step is to recognize the word type for each word as follows.

He	hits	the	dog.
PN	V	*the*	N

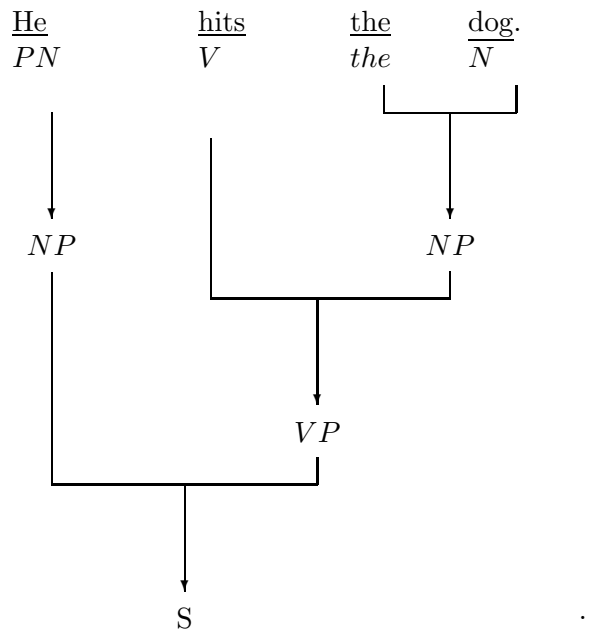
Then we recognize the word "He "as derived by $NP \longrightarrow PN$, the word "the "and "dog "as derived by $NP \longrightarrow the\ N$. Hence we try



Again we find only one recognizable substring, namely "$V \ \ NP$ " that can been derived by $VP \longrightarrow V\ NP$. So here we are forced to



and also our last reduction step leaves us no choice:



Both parsing techniques can be used in structural parsing. Note that also in bottom-up parsing we may have more possibilities to choose from.

5.3 Search techniques

Search techniques are used to guide the parsing process through all its possibilities to find one or all parsings.
There are in general two methods for searching, which are depth-first search and breadth-first search in the production tree of partially generated solutions.
Suppose there are several alternatives for further processing a half-solved problem. In depth-first search we concentrate on one alternative, and continue with that alternative until we reach a dead end. Then we go back in the production tree to chose another alternative. In breadth-first search we keep all the alternatives for each half-solved problem, unless we reach a dead end, then we chose one alternative to repeat the same process.
The distinction between breadth-first search and depth-first search is rather evident. Both of them are valid, we can just chose one of both.

6 Structural parsing

Parsing in natural language processing is defined as transforming a sentence or text into a representation of that sentence or text. This representation can therefore also be a knowledge graph, which includes the concept of a word graph and that of a sentence graph. A word graph expresses the meaning of a word with a *structure*, and a sentence graph expresses the meaning of a sentence with a *structure*. A word graph is then a basic unit in natural language processing; a sentence graph is to be formed from the word graphs of the words that appear in this sentence.
Knowledge graphs for words can be both syntactic and semantic, as we have seen in Section 4. A syntactic word or sentence graph expresses the syntax of a word or sentence; a semantic word or sentence graph is to express the semantics of a word or sentence, see Section 2.

Parsing by representing a sentence with a knowledge graph is a new field, which we call *structural parsing.*

Definition 4 *Structural parsing is the mapping of a sentence on a semantic sentence graph.*

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.

6.1 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, discussed in Section 4, 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 the grammars listed in Section 4. We do not like to produce many syntactic sentence graphs for complexity reasons. For this reason, we would like to give our own parsing, that is adapted to our knowledge graph theory. The key to our parsing method lies in a discussion of *utterance paths.*

6.2 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.

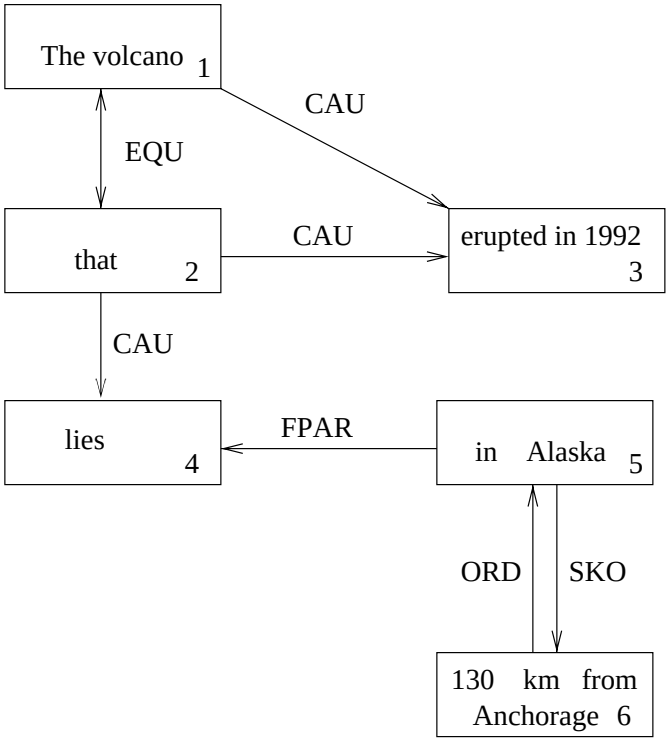


Figure 8: Semantic sentence graph with display of “chunks”.

- 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 6 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:

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

Note that " jumps " occur, consecutive chunks, not linked in the sentence graph. In the first sentence there is a jump 6⟶3, in the 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.

6.3 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 6.2, " the volcano", that 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 6.2.

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: Prepositions.

In structural parsing, we do not think complete parse trees are necessary. 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 consider a few examples as an experiment, in which also other points discussed in this paper should become clear.

6.4 Examples of structural parsing

In this section there are two example sentences : one is in English, the other is in Chinese. We will give a detailed analysis, indicating the different phases.

Example 1

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

Phase 0

The preparatory phase contains two parts.
First, we chunk the sentence by checking indicators, which were discussed in Section 6.2, 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 been listed in Section 3, here we indicate them with word types abbreviations, see Figure 9.
We construct syntactic chunk graphs chunk by chunk, like in Figure 10. The syntactic word graphs used are only represented with relevant arcs. The other arcs are indicated by dotted lines.

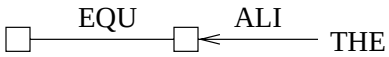
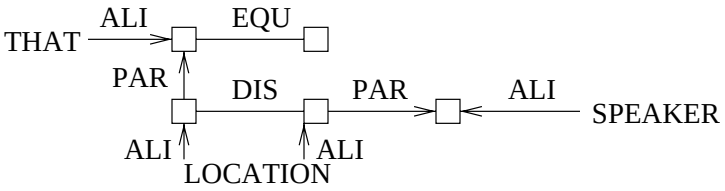
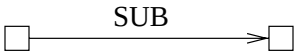
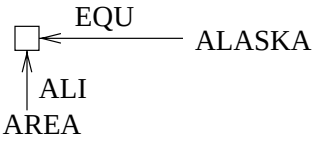
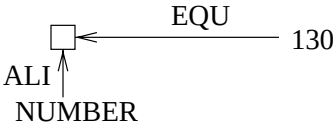
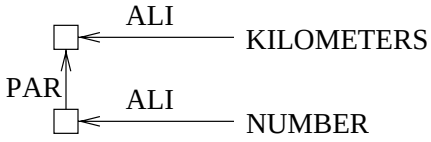
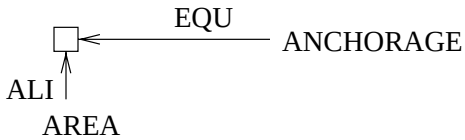
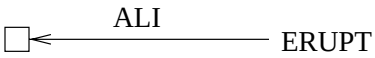
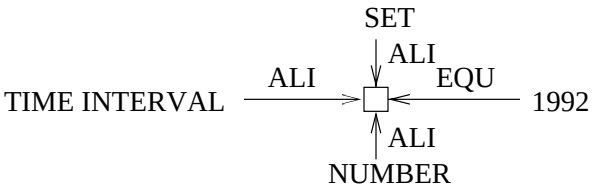
Words	Semantic Word Graphs	Word Types
THE		det
VOLCANO		N
THAT		PN
LY		V
IN		prep
ALASKA		N
130		num
KILOMETERS		N
FROM		prep
ANCHORAGE		N
ERUPT		V
IN		prep
1992		N

Figure 9: Lexicon of example 1.

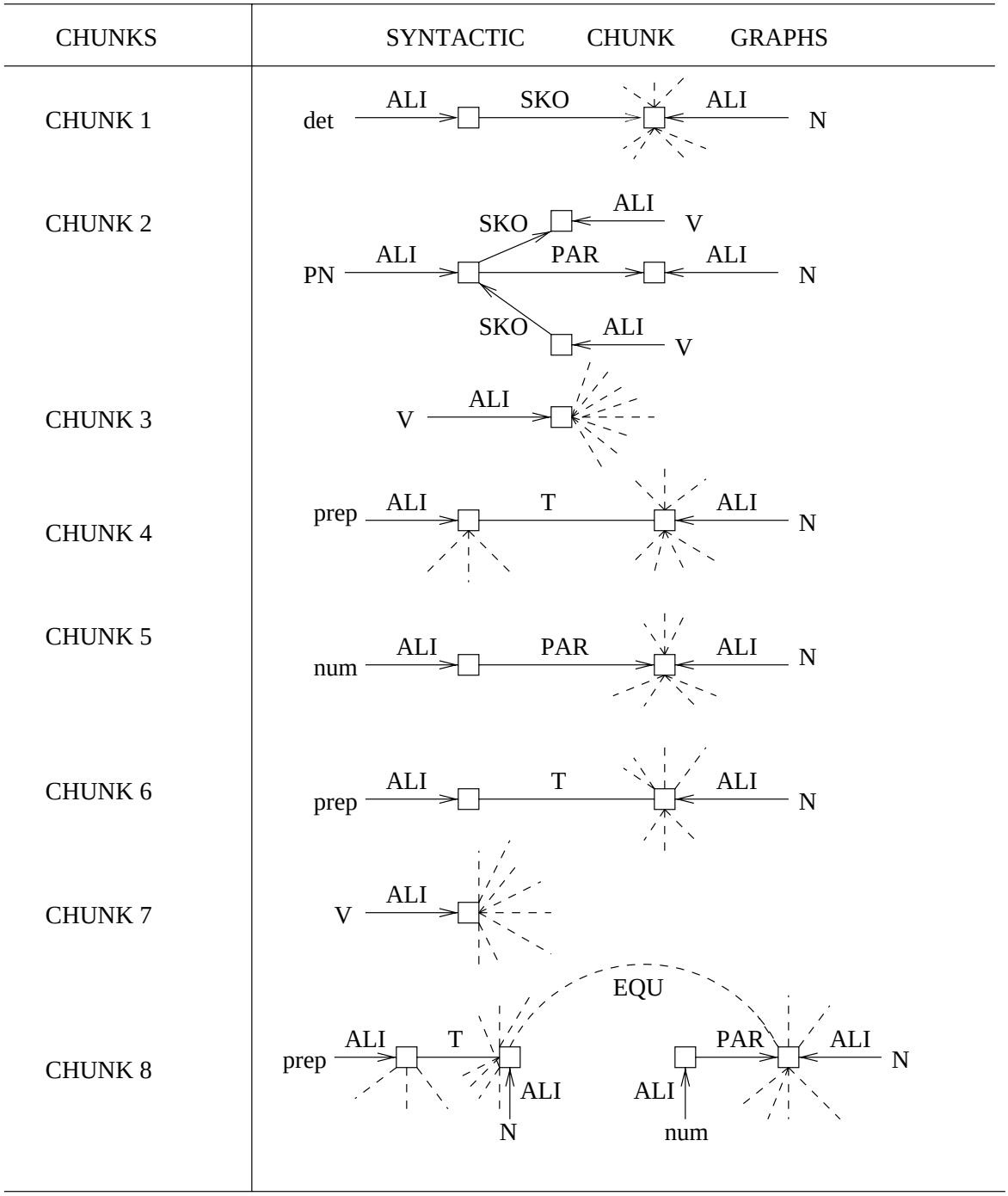


Figure 10: Syntactic chunk graphs of example 1.

So now the sentence is expressed like ". CH1, CH2 CH3 CH4, CH5 CH6, CH7 CH8. " We will combine syntactic chunk graphs into a bigger one when they can be linked syntactically. We use a new number to indicate a chunk, which may be a combination of sub-chunks.

$CH9 \quad = CH1$
$CH10 = CH2\ CH3\ CH4$
$CH11 = CH5\ CH6$
$CH12 = CH7\ CH8$.

Checking these chunks, from CH9 to CH12, we found that only CH2 and CH3 can be combined into one chunk, others allow no linking syntactically in this phase. The following figure shows the combination of CH2 and CH3.

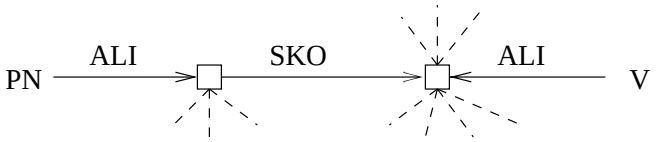


Figure 11: The syntactic graph of CH2+CH3.

Since there are no further linkings syntactically, we now will construct semantic chunk graphs by using the simple semantic word graphs given in Figure 9. Note that we might have given expanded versions of these semantic word graphs.

Phase 2

In order to make things clear, we renumber all the chunks as follows :

$CH9 \quad = CH1$
$CH10a = CH2\ CH3$
$CH10b = CH4$
$CH11a = CH5$
$CH11b = CH6$
$CH12a = CH7$
$CH12b = CH8$.

Now we give the formation of semantic chunk graphs in Figure 12 and describe this in more detail. We do not consider word changes like " lies " instead of " ly ".

- For CH9, which is " the volcano ", the semantic chunk graph can be given directly.
- With simple semantic word graphs, the verb " ly " cannot be combined with "in Alaska" directly, so we have semantic chunk graphs of CH10a and CH10b separately.
- The semantic chunk graph of " 130 kilometers " can be obtained from the simple semantic word graphs of " 130 " and " kilometers ".

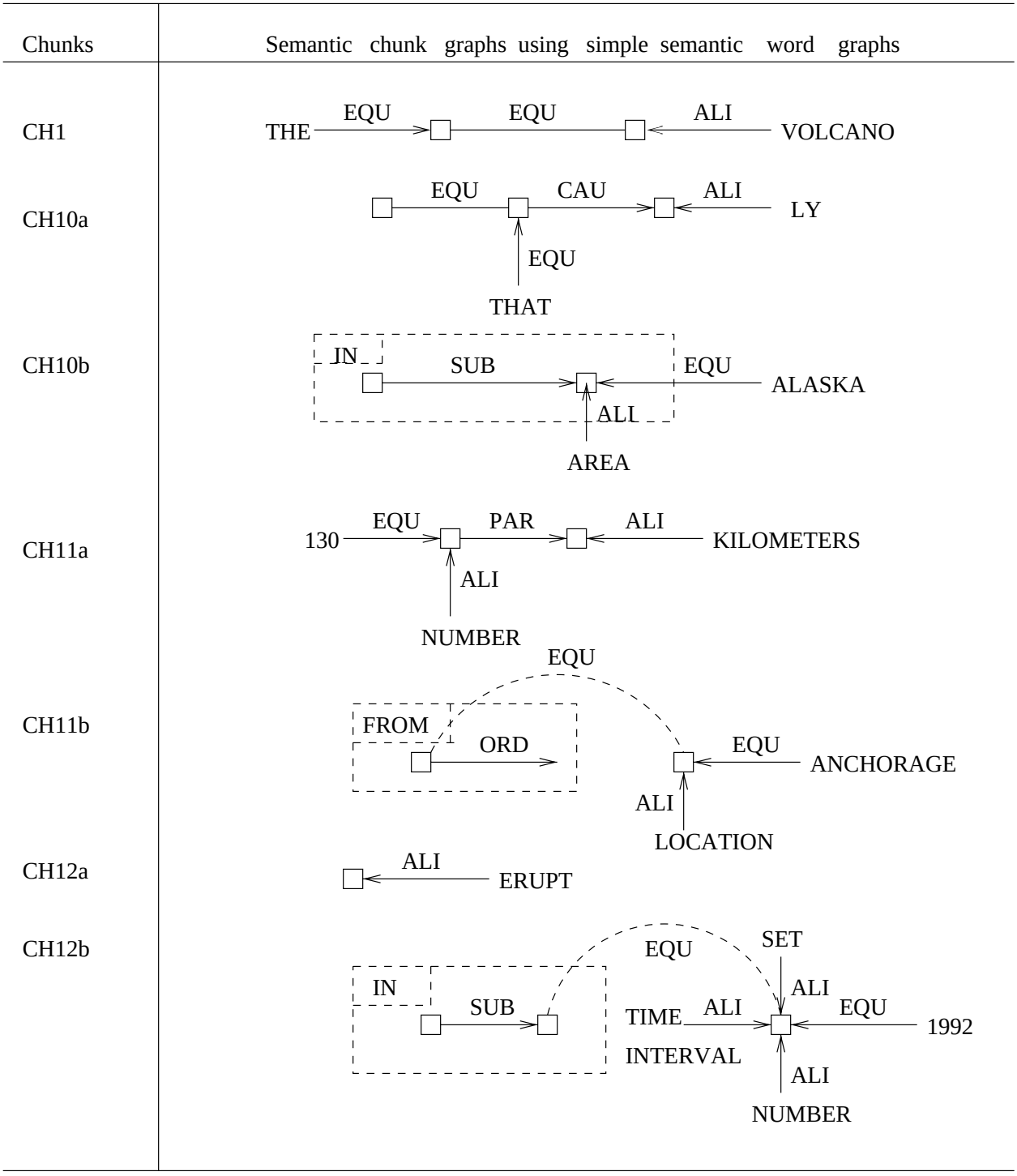


Figure 12: Semantic chunk graphs using simple semantic word graphs from Figure 9.

- The semantic chunk graph of " from Anchorage " can be obtained by identification of tokens as indicated by the EQU-link from the token in the preposition " from " to that in the noun " Anchorage ".
- The semantic chunk graph of " erupt " is very simple.
- In the semantic word graph of " in ", the tokens are not specified, so the right hand token can be identified with another token, like that for the number 1992. Note that we expressed " 1992 " with an ALI-arc to indicate the set nature of the time interval.

Due to the fact that the used semantic word graphs are so simple, we also have some sub-chunks which cannot be combined in this phase. To achieve this we need some background knowledge. This is introduced by expanding the simple semantic word graphs into more complex ones.

Phase 3

First we expand "ly" and "in". For "ly" we add two FPAR-arcs from tokens of type "area" . For "in" the tokens are given type "area". We do this, because semantically "ly" and "in" are both related to areas. Then CH10a and CH10b can be combined into CH10 as in Figure 13.

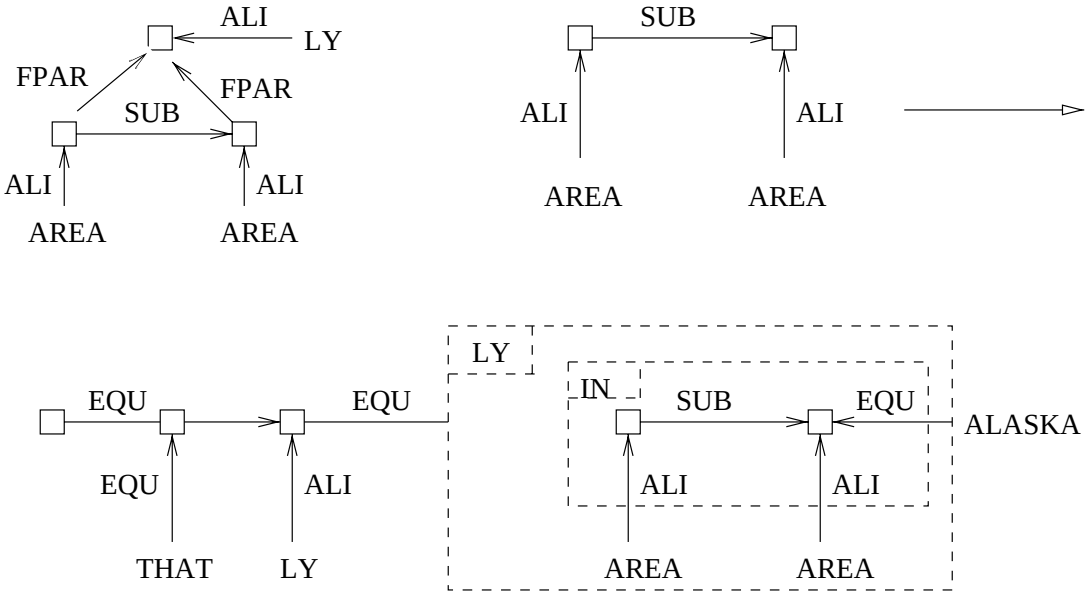


Figure 13: The semantic chunk graph of CH10.

We expand "from" with an ALI-arc linking to "location" , and expand "130 kilometers" with a PAR-arc linking to "distance" that has two SKO-arcs both to "location". So, CH11a and CH11b can be combined into CH11 as in Figure 14:

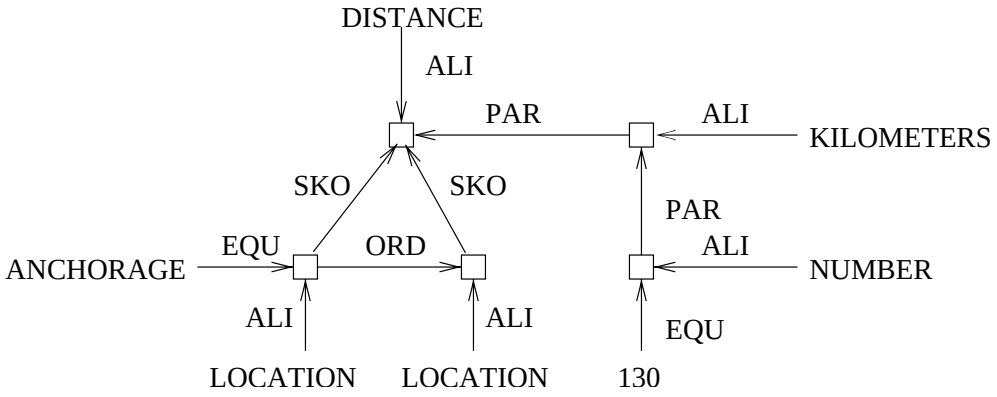


Figure 14: The semantic chunk graph of CH11

We expand " erupt " with a PAR-arc and a CAU-arc. There are two points: "erupt" is a verb, it should have a CAU-arc coming in; semantically "erupt" is related to a "location". Linking "in" to "erupt" with a PAR-arc is the only possibility to combine them. CH12a and CH12b should be combined into CH12 like in Figure 15.

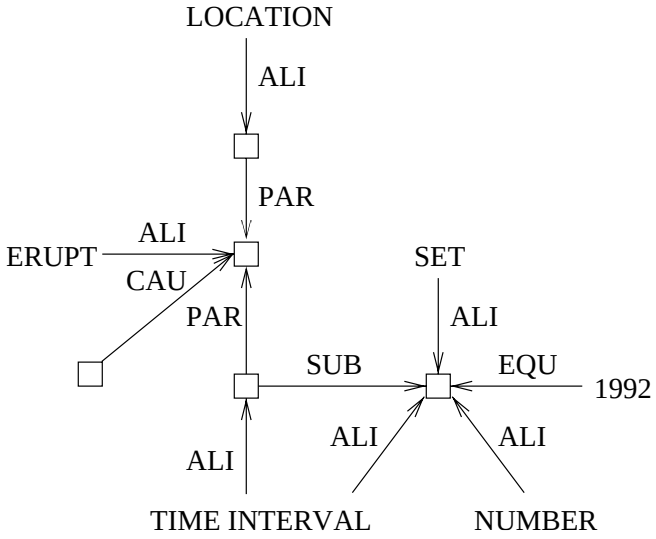


Figure 15: The semantic chunk graph of CH12

Now we get the result:

$S = CH1, CH10, CH11, CH12.$

Here the sentence has 4 chunks, which have corresponding semantic chunk graphs.

Finally, we link the four chunk graphs from the left hand to the right hand side, unless there is a jump, and we obtain the semantic sentence graph, see Figure 16, where some arcs used in the analysis, like the ALI-arcs to "set" or "number", have been omitted for reason of clarity.

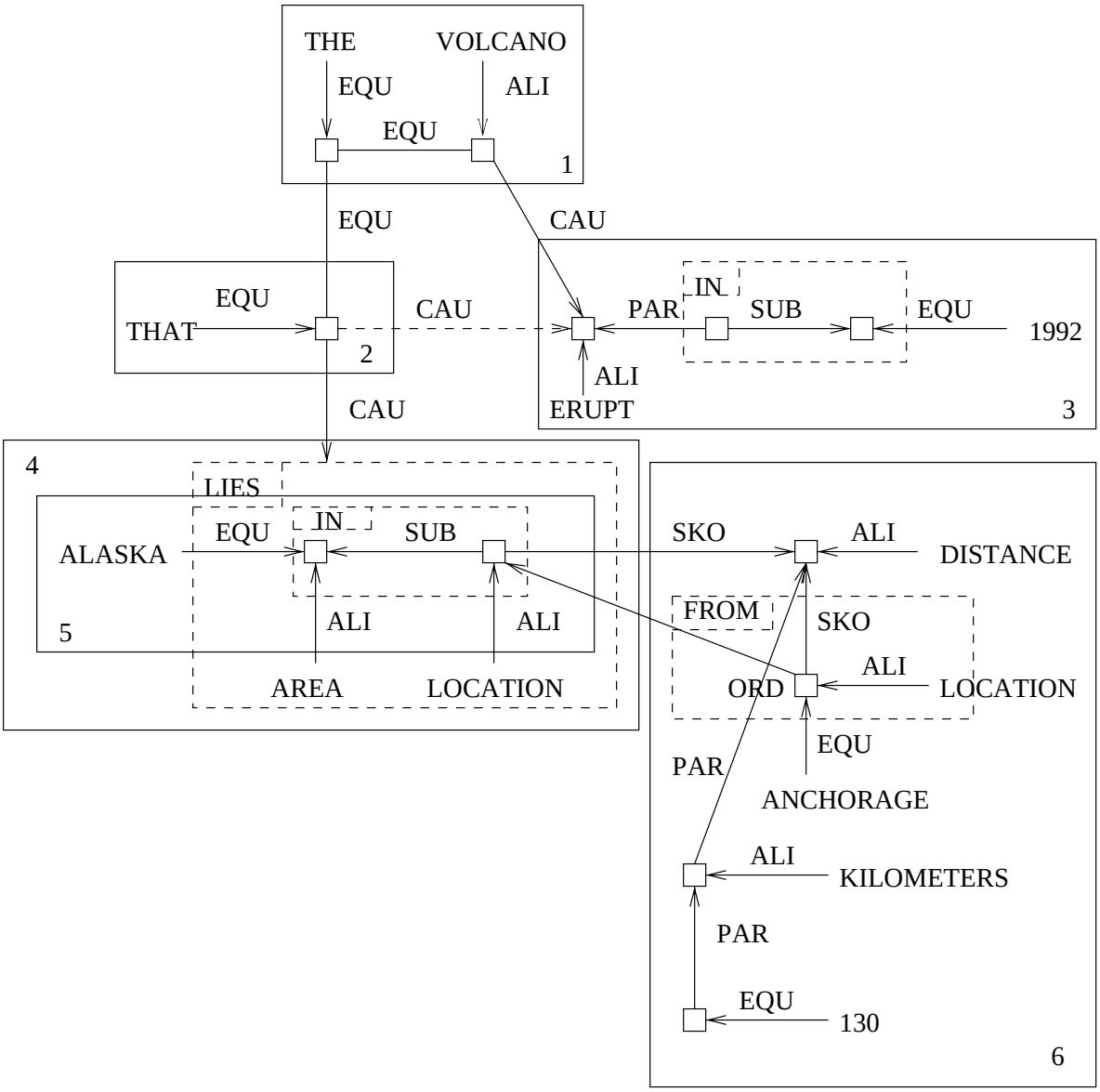


Figure 16: The semantic sentence graph for the English example sentence.

Now let us take a Chinese sentence as an experiment to test the 5 phases of structural parsing.

Example 2

Wo3 de xiao3 di4 di4 zai XIAN de xi1 bei3 da4 xue2 shang4 xue2.

(I of small brother in XIAN of west north university study.)

" My small brother studies in Northwest University of XIAN. "

In Phase 0, a lexicon of this sentence is given in Figure 17.

In Phase 1, the sentence is chunked as follows, according to chunk indicators:

1. [$Wo3$]
2. [de]
3. [$xiao3$ $di4$ $di4$]
4. [$zai4$ $XIAN$ de]
5. [$Xi1$ $Bei3$ $Da4$ $Xue2$ $shang4$ $xue2$] .

In Phase 2, using syntactic word graphs, we obtain syntactic chunk graphs like in Figure 18.

In Phase 3, syntactic chunk graphs are linked into bigger ones, see the upper part of Figure 19.

In Phase 4, CH6 and CH7 are expressed by semantic word graphs, like in the lower part of Figure 19.

Note that expansion is needed for linking " zai4 " with " XIAN ", and "ren2" has been added for CH6.
There are two problems: " zai4 " has an unspecified token; in the chunk graphs " ren2 " is not linked.
Both problems are solved by expanding the semantic word graphs. For "di4 di4" and for "shang4 xue2" , expansions are chosen as in Figure 20.

Then the " wrong " CAU-arc from " XBDX " to " shang4 xue2 " is cut, because " ren2 " does not occur in the expansion of " XBDX ". The CAU-arc looks for " ren2 " and therefore links with " di4 di4 ". The " wei4 zhi4 " of " shang4 xue2 " " fills " the unspecified token in " zai4 ".
In Phase 5, the semantic sentence graph is completed, see Figure 21.

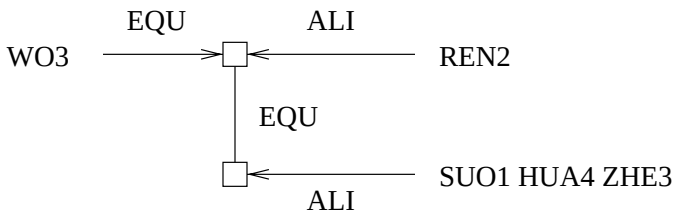
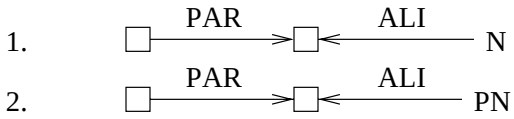
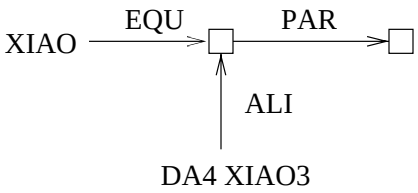
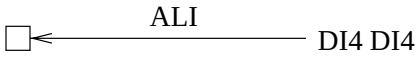
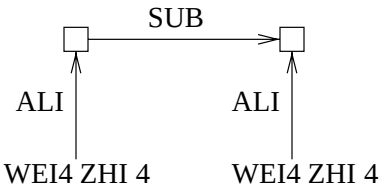
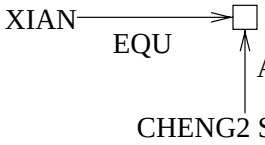
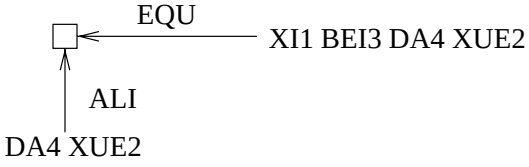
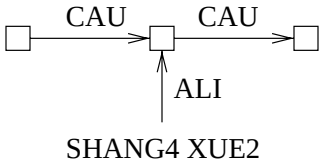
Words	Semantic Word Graphs	Word Types
WO3		PN
de		prep
XIAO		adj
DI4 DI4		N
ZAI4		prep
XIAN		
XI1 BEI3 DA4 XUE2		N
SHANG4 XUE2		V

Figure 17: Lexicon of example 2.

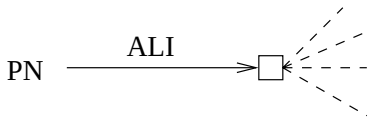
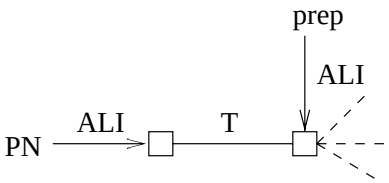
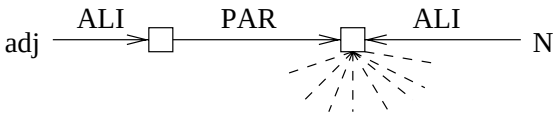
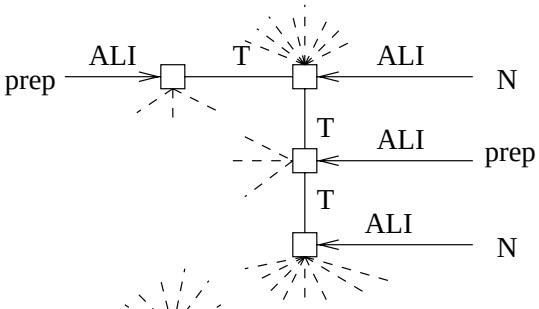
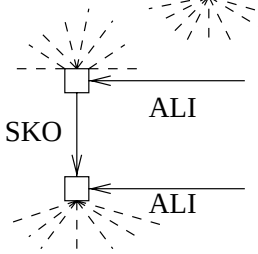
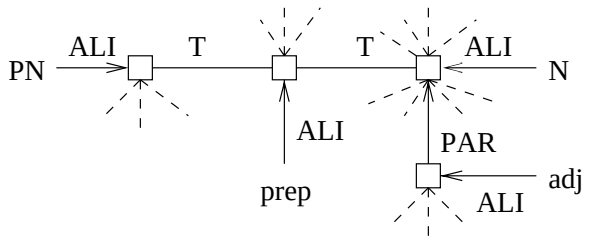
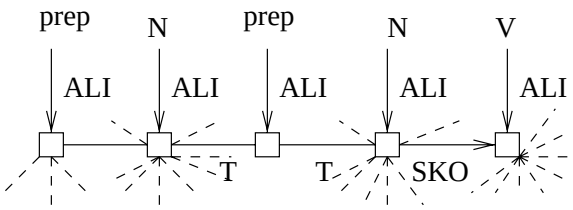
CHUNKS	SYNTACTIC CHUNK GRAPHS
CHUNK 1	
CHUNK 2	
CHUNK 3	
CHUNK 4	
CHUNK 5	

Figure 18: Syntactic chunk graphs of example 2.

CH6: CH1+CH2+CH3	
CH7: CH4+CH5	

Bigger syntactic chunk graphs

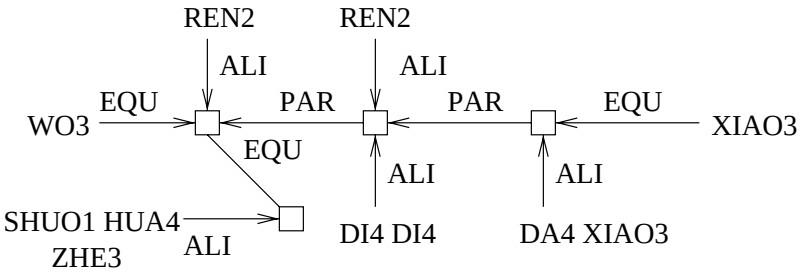
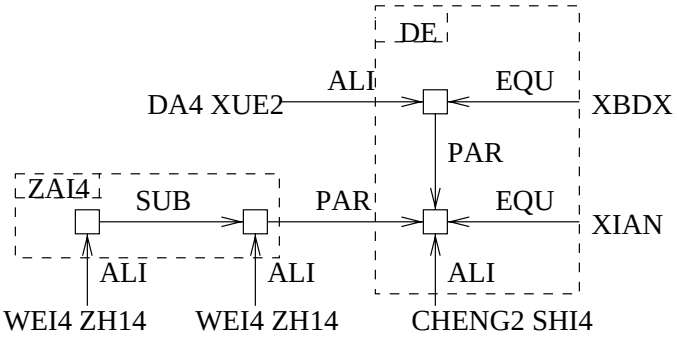
CH6	
CH7	

Figure 19: Semantic chunk graphs of CH6 and CH7.

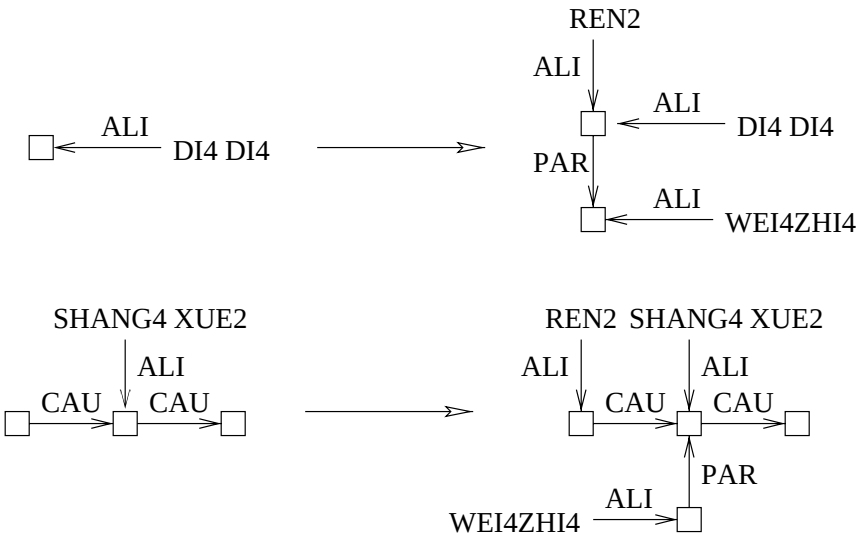


Figure 20: Expansions for "di4 di4" and "shang4 xue2".

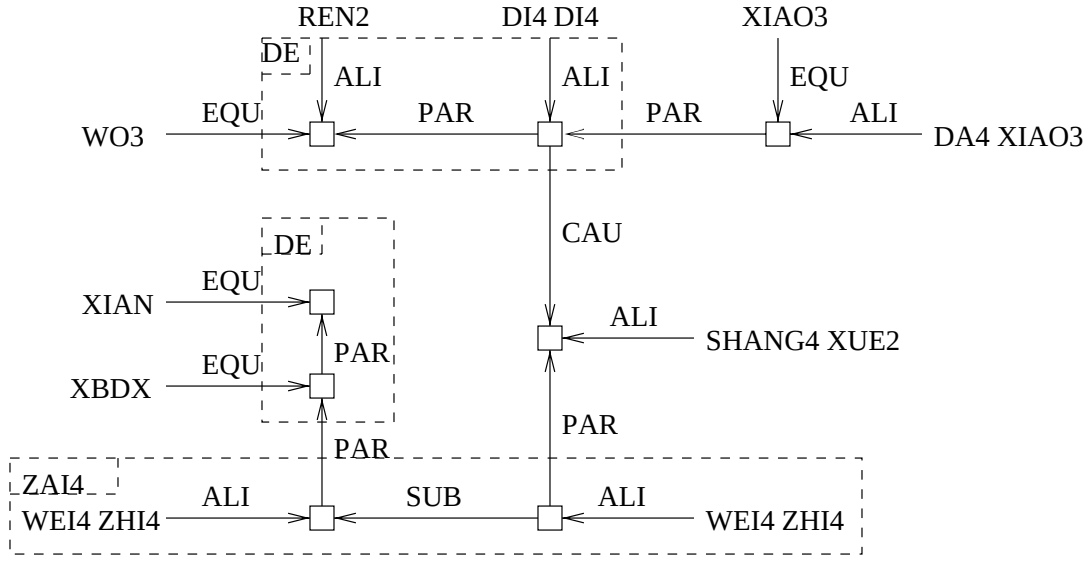


Figure 21: The semantic sentence graph for the Chinese example sentence.

7 Conclusion

The standard way of parsing is to find a generation of a sentence by a grammar. The most elementary aspects of mapping a sentence on a socalled parse tree were mentioned in Section 5. 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 structure 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 in this paper is that of a syntactic word graphs for a certain type of word. The word graphs originally considered in the theory are called semantic word graphs. From the syntactic word graphs grammars can be derived, which is done for both Chinese and English.
As syntactic word graphs express the way words function with respect to each other, which has corresponding linking of semantic word graphs, a traditional parse tree obtained with the derived grammars will allow corresponding linking of semantic word graphs to obtain a semantic sentence graph.
It turns out, however, that the preparatory step of obtaining a syntactic sentence graph need not be carried out according to the traditional approach. 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 a Chinese sentence and an English sentence were investigated. The idea turned out to be quite fruitful. Semantic sentence graphs were obtained, be it that background knowledge concerning the words had to be included by expanding the semantic word graphs. That this is possible, however, is considered to be a point of strength of knowledge graph theory.

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

[8] Yao, T. S. , *Natural Language Processing*, Qinghua University Press, Beijing, ISBN 7-302-01911-8, (1995).

[9] Zhu, D. X. , *Grammar Tutorial*, Commerce Printing House Press, Beijing, (1984).