Modeling syntax of Free Word-Order Languages: Dependency Analysis By Reduction

Markéta Lopatková[1], Martin Plátek[2], Vladislav Kuboň[1]

[1] ÚFAL MFF UK, Praha
{lopatkova,vk}@ufal.mff.cuni.cz
[2] KTIML MFF UK, Praha
martin.platek@mff.cuni.cz

Abstract. This paper explains the principles of dependency analysis by reduction and its correspondence to the notions of dependency and dependency tree. The explanation is illustrated by examples from Czech, a language with a relatively high degree of word-order freedom. The paper sums up the basic features of methods of dependency syntax. The method serves as a basis for the verification (and explanation) of the adequacy of formal and computational models of those methods.

1 Introduction – analysis by reduction

It is common to describe the syntactic structure of sentences of English or other fixed word-order languages by phrase structure grammars. The description of the syntactic structure of Latin, Italian, German, Arabic, Czech, Russian or some other languages is more often based on approaches which are generally called dependency based. Both approaches are based on stepwise simplification of individual sentences, on the so-called analysis by reduction. However, the basic principles of the phrase-structure and dependency based analysis by reduction are substantially different. The phrase-structure based analysis (of fixed word-order languages) can be naturally modeled by the bottom-up analysis using phrase structure (Chomskian) grammars. This paper should help the reader to recognize that it is necessary to model the dependency analysis by reduction of languages with a high degree of word-order freedom differently. We try to explain explicitly the common basis of the methods for obtaining dependencies, presented in [4, 7, 9].

Unlike the artificial (programming) languages, the natural languages allow for an ambiguous interpretation. Instead of a complete formal grammar (of an artificial language), for natural languages we have at our disposal the ability of sentence analysis – we learn it at school, it is described by means of implicit rules in grammars of a given language.

The grammar textbooks are based on the presupposition that a human understands the meaning of a particular sentence before he starts to analyze it (let us cite from the 'Textbook of sentence analysis' (see [10]): "A correct analysis of a sentence is not possible without a precise understanding of that sentence, ... ").

An automatic syntactic analysis (according to a formal grammar), on the other hand, neither does presuppose the sentence understanding, nor has it at its disposal. On the contrary, it is one of the first phases of the computational modeling of a sentence meaning.

What is actually the relationship between the sentence analysis and the analysis by reduction? In simple words, the sentence analysis is based on a more elementary ability to perform the analysis by reduction, i.e. to simplify gradually the analyzed sentences. The following simplified example illustrates the methodology of the dependency analysis by reduction.

Example 1. The sentence *'Studenti dělali těžkou zkoušku.' [Lit.: Students passed difficult exam.]* can be simplified (while preserving its syntactical correctness) in two ways (see also the scheme in Fig. 1) – by the deletion of the word form *studenti* or by the deletion of the word form *těžkou* (but not by the deletion of the word form *zkoušku* – the sentence *'*Studenti dělali těžkou.'* is not acceptable in a neutral context). In the second step we can remove the word form *těžkou* (in the first branch of the analysis) or the word form *studenti*, or even the word form *zkoušku* (in the second branch). In the last step we can delete the word form *zkoušku* (in the first branch), or the word form *studenti.*

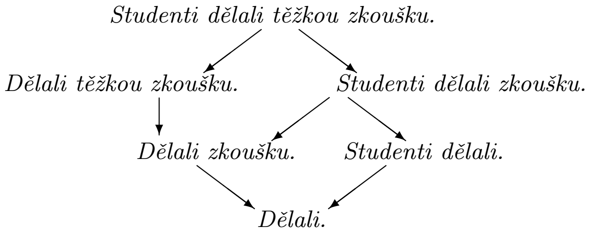

Fig. 1. The DAR scheme for the sentence *'The students passed a difficult exam.'*

The DAR scheme is closely related to a dependency tree, Fig. 2 shows the dependency tree for the sentence *Studenti dělali těžkou zkoušku.*

(i) A particular word depends on (modifies) another word from the sentence if it is possible to remove this modifying word (while the correctness of the sentence is preserved).

(ii) Two words can be removed stepwise in an arbitrary order if and only if they are mutually independent.

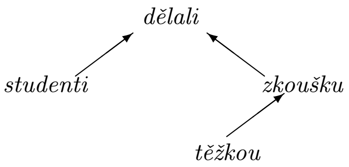

Fig. 2. The dependency tree for the sentence *'Studenti dělali těžkou zkoušku.'*

This example illustrates the way how we can obtain an information about dependencies (relationships between modified and modifying words in a sentence) using DAR. Let us stress the following fact: if taking correct Czech sentences with permuted word order, e.g. *'Těžkou zkoušku studenti dělali.'* or *'Těžkou dělali* *studenti zkoušku.'*, we get totally analogical reduction scheme as for the original sentence (the deleted words are identical in all steps of the reduction). This indicates that the dependency analysis by reduction allows to examine dependencies and word order independently. In other words, it provides a method for studying the degree of independence of the relationship between modified and modifying words in a sentence on its word order.

In this paper we concentrate on the description of rules for a dependency analysis by reduction of Czech, a language with a relatively high degree of word-order freedom, and on clarification of the relation between a dependency analysis by reduction and dependency sentence analysis.

The main reason for studying the analysis by reduction is the endeavor to gain a clear idea about its formal and computational modeling. Note that a formal model of analysis by reduction, restarting automata, is already intensively studied (see e.g. [3, 6]).

2 Dependency analysis by reduction

The **dependency analysis by reduction (DAR)** is based on stepwise simplification of a sentence – each step of DAR is represented by exactly one **reduction operation** which may be executed in two ways:

(i) by deleting at least one word of the input sentence, or

(ii) by replacing an (in general discontinuous) substring of a sentence by a shorter substring.

The possibility to apply certain reduction is restricted by the necessity to preserve some (at least the first one) of the following **DAR principles**:

(a) preservation of syntactical correctness of the sentence;

(b) preservation of lemmas and sets of morphological categories characterizing word forms that are not affected by the reduction operation;

(c) preservation of the meanings of words in the sentence (represented e.g. by valency frame,[3] or by a suitable equivalent in some other language);

(d) preservation of the independence of the meaning of the sentence (the sentence has independent meaning if it does not necessarily invoke any further questions when uttered separately).[4]

With respect to a concrete task (e.g. for grammar checking) it is possible to relax these DAR principles; those which are not relaxed are then called **valid DAR principles** (e.g. in the example 1 we have relaxed the principle of preservation of the independence of sentence meaning).

If it is possible to apply a certain reduction in a certain step of DAR (preserving all valid principles), we talk about **admissible reduction**. By the application of all admissible reductions it is possible to get all **admissible simplifications** of a sentence being reduced.

[3] The valency frame describes syntactic-semantic properties of a word, see e.g. [5].

[4] A sentence with independent meaning consists of a verb, all its semantically 'obligatory' modifications and (recursively) their 'obligatory' modifications, see [7].

We are going to use the term **DAR scheme (reduction scheme)** of a sentence of a given language for an oriented graph, whose nodes represent all admissible simplifications of a given sentence (including the original sentence) and whose edges correspond to all admissible reductions that can be always applied to a starting node of the edge and whose result is the admissible simplification of a sentence in its final node.

Example 2. The reduction scheme of the sentence *'Studenti dělali těžkou zkoušku.'* in Fig 1 illustrates the reductions of the type (i) – we delete at least one word of the input sentence in every step of the DAR whereas the possibility of branching captures the non-deterministic nature of the DAR. The reduction of the type (ii) is illustrated by possible simplification of the sentence *Kursem prošlo patnáct studentů. [Lit.: Course completed fifteen students.]*. Its reduction scheme is presented in Fig 3 (again, the principle (d) of the preservation of independence of meaning is relaxed).

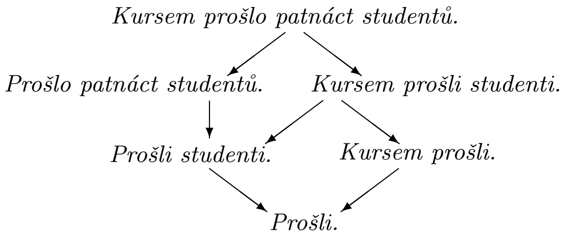

Fig. 3. The reduction scheme for the sentence *'Kursem prošlo patnáct studentů.'*

3 The structure of reduction and a dependency tree

The DAR scheme allows to introduce and classify various types of relationships. On the basis of these relationships we can define a structure of a sentence reduction.

Let us have a language L, a sentence $v \in L$, $v = v_1 v_2 ... v_m$, where $v_1, v_2, ..., v_m$ are the words, and a DAR scheme of the sentence v. We will say that the words v_i $i \in N, N \subseteq \{1, 2, ...m\}$ constitute a **reduction component**, if all words v_i are always removed at the same moment (i.e. in the DAR scheme all words v_i are removed in one step, which corresponds to a single edge in the scheme). We will say that the word v_i is **dependent (in the reduction)** on the word v_j, if the word v_i is deleted earlier than v_j in all branches of the DAR; the word v_j will be called a **governing (in the reduction)** word.

We will say that the words v_i and v_j are **independent on each other (with regard to the reduction)**, if they can be deleted in an arbitrary order (i.e. there is a DAR branch in which the word v_i is deleted earlier than the word v_j, and there is a DAR branch in which the word v_j is deleted earlier than the word v_i).

Based on the terms of dependency and component in the reduction we can define a reduction structure of a sentence, as it is illustrated in the following example.

Example 3. The reduction scheme of the sentence *'Studenti dělali těžkou zkoušku.' [Lit.: Students passed difficult exam.]* which preserves all DAR principles (including the principle (d) preservation of the independence of the meaning of the sentence) can be found on Fig. 4 – the verb *dělat* has two 'obligatory' modifications corresponding to a subject and a direct object, the noun *studenti* does not have obligatory modifications, therefore the sentence with independent meaning has a form *'Studenti dělali zkoušku.' [Lit.: Students passed exam.]*

Studenti dělali těžkou zkoušku.

↓

Studenti dělali zkoušku.

Fig. 4. The DAR scheme for the sentence *'Studenti dělali těžkou zkoušku.'* when applying the principle of preservation the independence of the sentence meaning.

The **reduction structure** can be captured by a diagram in which the nodes represent individual words from the sentence, the horizontal edges connect a reduction component (an edge always connects two neighboring words of a reduction component). The oblique edges reflect reduction dependencies; they are considered to be oriented from the dependent word (or from the whole reduction component) towards the governing word (or, again, towards the whole reduction component, if it is governing that particular word (component)). The linear order of nodes (left to right) captures the word-order (the order of words in the sentence). Fig. 5 shows the reduction structure representing the sentence *Studenti dělali těžkou zkoušku.*

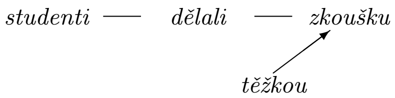

Fig. 5. The reduction structure for the sentence *'Studenti dělali těžkou zkoušku.'*

Traditionally, the structure of a (Czech) sentence is described by a dependency tree. Such a description is transparent and proper for sentences not complicated by coordinations, ellipses and by some marginal phenomena. The **dependency tree** is a structure that is a finite tree in the sense of a graph theory, and it has a root into which all paths are directed and whose nodes are totally (linearly left-to-right) ordered. The nodes represent the occurrences of word forms used in the sentence, the edges represent the relationship between a governing and a governed word (unit) in the sentence.

The only thing left to describe is how to get a dependency tree from a reduction structure. Reduction dependencies are easy, the respective edges characterize the relationship between the modifying and the modified word, the order of words in the sentence is preserved.

For reduction components it is necessary to find out which word from a given component will be considered as governing and which one will be dependent. For this purpose it is necessary to introduce additional rules for individual linguistic phenomena, which are studied in more detail in the following section.

4 Reduction relationships in a natural language

The formal typology of dependencies introduced in the previous section corresponds to a traditional linguistic classification – in this section we will try to describe this correspondence in more detail.

Let us suppose that the reader is familiar with basic linguistic notions such as subordination[5] (relation between modified sentence member and its modifying sentence member), complementation of verb/noun/adjective/adverb, inner participant (argument) and free modification (adjunct), obligatory and optional complementation. Description of these terms can be found e.g. in [9], [7] and [5].

Dependencies (in DAR) allow to model directly the optional free modifications – here it is possible to replace the whole pair by a modified word, a 'head' of the construction (without loosing the independence of meaning, the principle (d) of DAR). Thus we can capture the relationships like *těžká zkouška, jde pomalu, jde domů, přichází včas [Lit.: difficult exam, (she) walks slowly, (he) goes home, (he) comes in time]*. The governing word (in the reduction) corresponds to the modified word in the sentence, the dependent word (in the reduction) corresponds to the word which modifies it (see Fig. 6).

It remains to determine the governing and dependent member in those cases in which the modified or modifying member of this dependency consist of the whole reduction component, rather than of a single word.

(i) If the modifying member consists of the reduction component, then the dependent member is the governing word of this component (the remaining members of the component constitute a subtree with a root in this governing word).

(ii) If the modified sentence member consists of the reduction component, then the whole construction in general has ambiguous meaning (interesting examples for Czech can be found in [2]).

zkouška
těžká
jde
domů

Fig. 6. Dependencies in DAR model free modifications.

Reduction components allow for modeling more complex relationships between word occurrences. These are either (a) morpho-syntactic relationships, or (b) syntactically-semantic relationships.

(a) Reduction components describe so-called **formemes**, the units corresponding to individual sentence members – these are especially prepositional

[5] The term of 'subordination' describes the language relationship, while the term of 'dependency' is reserved here for formal structures, by means of which language relationships are modeled.

groups (as *na stole, vzhledem k okolnostem [Lit.: on table, with respect to circumstances]*) or complex verb forms (*přijel jsem, tiskne se [Lit.: (I) did arrive, (it) is being printed]*).

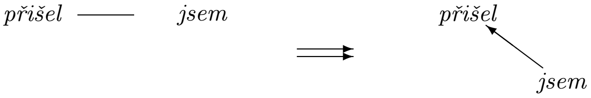

Fig. 7. A possible transformation of formemes into a dependency subtree.

In traditional linguistics each formeme constitutes one node of the diagram, or dependency tree describing syntactic structure of the sentence, see e.g. [10] or [9]. In these theories only the meaningful words (especially meaningful verbs, nouns, adjectives and adverbs) are represented by independent nodes. However, for many practically oriented tasks (e.g. grammar-checking, building of a syntactically annotated corpus) it is appropriate to represent each word of a sentence by its own node. In order to preserve the traditional data type of the dependency tree it is necessary to specify additional rules on the basis of which even the reduction components can be transformed into subtrees, i.e. it is necessary to specify which word from the formeme will be considered governing and which one will be dependent. Such rules are usually of a technical nature and they can differ in individual projects (Fig. 7 shows the solution adopted in [1]).

(b) The second type of relationships modeled by reduction components are syntactically-semantic relationships. These are especially **valency relationships** – the relationships of a verb, noun, adjective or adverb and its obligatory valency complementation(s) (as e.g.*studenti dělali zkoušku, Petr dal Pavlovi dárek, začátek přednášky [Lit.: students passed exam, Petr gave Pavel gift, beginning (of) lecture]*). These constructions cannot be replaced by a single word, the 'head' of the construction, without loosing the independence of meaning, DAR principle (d).

Traditional linguistics captures the valency relationships using dependency tree (see [9] and [10]). The theoretical criterion for the determination of modified and modifying sentence member, the principle of analogy in the layer of word classes is discussed in [9] – the verb is considered as a modified word (as an analogy to verbs without obligatory complementations), the verb complementations are the modifying words; similarly for nouns, adjectives, adverbs and their complementations. This principle of analogy is also adopted for determining the governing word during the transformation of reduction structure to a dependency tree: the verb is considered as a governing word, the verb complementations are its dependent words; similarly for nouns, adjectives, adverbs.

Let us note that the analogy principle can be simply substituted by a relaxation of the condition (d) preserving the independence of meaning of DAR.

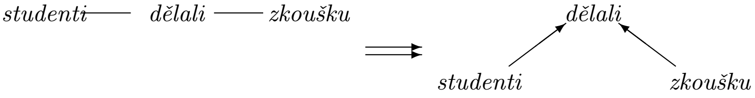

Fig. 8. The transformation of valency relationships into a dependency subtree.

Concluding remarks

The DAR allows to formulate the relationship of basic syntactic phenomena: a dependency and a word order. This approach is indispensable especially for modeling the syntactic structure of languages with a free word-order, where the dependency and word-order are very loosely related and where they are also related in a different manner from language to language (let us compare this situation with English, where the dependencies are determined (mainly) by a very strict word-order).

The paper shows that the dependencies can be derived from two different, not overlapping, simply observable and language independent phenomena: from the reduction dependency and from reduction components. It also points out that the (Czech) traditional linguistic taxonomy of language phenomena corresponds to this division. We have mentioned the formal model of analysis by reduction, restarting automata. We have thus outlined one important step how to pass the observations about dependencies from traditional linguistics into the formal terms suitable for computer linguistics.

Acknowledgements. This paper is a result of the project supported by the grant No. 1ET100300517. We would like to thank an anonymous reviewer for his valuable comments and recommendations.

References

1. Hajič, J.: Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In: Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová (ed. Hajičová, E.), Karolinum, Prague, pp. 106-132 (1998)
2. Holan, T., Kuboň, V., Oliva, K., Plátek, M.: On Complexity of Word Order. In: Les grammaires de dépendance - Traitement automatique des langues (TAL), Vol. 41, No. 1 (q.ed. Kahane, S.), pp. 273-300 (2000)
3. Jančar, P, Mráz, F., Plátek, M., Vogel, J.: On Monotonic Automata with a Restart Operation. Journal of Automata, Languages and Combinatorics, Vol. 4, No. 4, pp. 287-311 (1999)
4. Kunze, J.: Abhängigkeitsgrammatik. Volume XII of Studia Grammatica, Akademie Verlag, Berlin (1975)
5. Lopatková, M.: Valency in the Prague Dependency Treebank: Building the Valency Lexicon. In: PBML 79-80, pp. 37-59 (2003)
6. Otto, F.: Restarting Automata and their Relations to the Chomsky Hierarchy. In: Developments in Language Theory, Proceedings of DLT'2003 (eds. Esik, Z., Fülöp, Z.), LNCS 2710, Springer, Berlin (2003)
7. Panevová, J.: Formy a funkce ve stavbě české věty. Academia, Praha (1980)
8. Plátek, M., Lopatková, M., Oliva, K.: Restarting Automata: Motivations and Applications. In: Proceedings of the workshop "Petrinetze" (ed. Holzer, M.), Technische Universität Műnchen, pp. 90-96 (2003)
9. Sgall, P., Hajičová, E., Panevová, J.: The Meaning of the Sentence in Its Semantic and Pragmatic Aspects (ed. Mey, J.), Dordrecht:Reidel and Prague:Academia (1986)
10. Šmilauer, V.: Učebnice větného rozboru. Skripta FF UK, SPN, Praha (1958)