A Case Study of a Free Word Order

Vladislav Kuboň and **Markéta Lopatková** and **Jiří Mírovský**
Charles University in Prague
Faculty of Mathematics and Physics
Czech Republic
{vk,lopatkova,mirovsky}@ufal.mff.cuni.cz

Abstract

The paper aims at the investigation of free word order. It concentrates on the relationship between (formal) dependencies and word order. The investigation is performed by means of a semi-automatic application of a method of analysis by reduction to Czech syntactically annotated data.

The paper also presents the analysis of introspectively created Czech sentences demonstrating complex phenomena which are not sufficiently represented in the corpus. The focus is on non-projective structures, esp. those connected with the position of clitics in Czech. The freedom of word order is expressed by means of a number of necessary shifts in the process of analysis by reduction.

The paper shows that this measure provides a new view of the problem, it is orthogonal to measures reflecting the word order freedom based on a number of non-projective constructions or clitics in a sentence. It also helps to identify language phenomena that generally pose a problem for dependency-based formalisms.

1 Introduction

The phenomenon of word order freedom plays an important role in natural language processing. The less rigid the word order, the bigger challenge for all kinds of parsing algorithms it constitutes.

In this paper we are going to study the relationship between *(formal) dependencies* – defined through analysis by reduction, a stepwise simplification of a sentence preserving its correctness (Lopatková et al., 2005)– and *word order*. We want to gain better insight into the problem by means of the application of a semi-automatic procedure to syntactically annotated data. This method can verify the concept of the analysis by reduction (introduced in Section 2) against real data and, at the same time, it can shed more light on the relationships between these two syntactic phenomena, dependency and word order.

Our goal is twofold:

First, we focus on typical, 'core' (projective) word order. We are going to quantify how many sentences can be completely processed by a simple analysis by reduction (i.e., sentences are (correctly) reduced until only the predicate is left in the sentence). In order to be able to perform this task for a large volume of data, namely syntactically annotated data from the Prague Dependency Treebank (PDT) (Hajič et al., 2006)), we have developed an automatic procedure (requiring, of course, a subsequent manual checking) on a relatively large subset of the PDT data. The results are presented in Section 3.

Second, we present an analysis of more 'peripheral' word order. When it is impossible to perform simple reduction (without violating the correctness constraint), we strengthen the analysis by reduction by the concept of 'shifts' – word order modifications which help to preserve the correctness constraint. Such data provide a very interesting material for the analysis of individual linguistic phenomena involved in complicated sentences (Section 4). We are primarily concentrating on the analysis of sentences which have been discovered as problematic in previous research, see esp. (Holan et al., 2000).

1.1 The Background

In the world of dependency representation, there are three essential (and substantially different) syntactic relationships, namely 1. *dependencies* (the relationship between a governing and a modifying sentence member, as e.g. a verb and its ob-

27th Pacific Asia Conference on Language, Information, and Computation pages 222−231

ject, or a noun and its attribute), 2. *'multiplication'* of two or more sentence members or clauses (esp. coordination), and 3. *word order* (i.e., the linear sequence of words in a sentence).[1]

In this paper we are concentrating on the phenomena 1 and 3, i.e., on the relationships of dependency and word order. The interplay between these two basic syntactic relationships is relatively complex especially in languages with a higher degree of word order freedom. The reason is simple – while the dependency relations are indicated primarily by morphological means (as morphological cases and agreement (Daneš et al., 1987)), the word order expresses primarily phenomena like communicative dynamism and topic-focus articulation (Hajičová and Sgall, 2004).

Within dependency linguistics, these relationships have been previously studied especially within the Meaning-Text Theory: the approach aiming at the determination of dependency relations and their formal description is summed up esp. in (Mel'čuk, 2011). An alternative formal description of dependency syntax can be found in (Gerdes and Kahane, 2011). Our approach is based on the Czech linguistic tradition represented mainly in (Sgall et al., 1986).

Let us now formulate the basic principle underlying the *analysis by reduction*: roughly speaking, if one of the words creating a possible governor-modifier pair can be deleted without changing the distribution properties of the pair (i.e., the ability to appear in the same syntactic context) then it is considered as a modifying one (dependent on the latter one). This is applicable on so called endocentric constructions (as, e.g. *small table, Go home!*); for exocentric constructions (as *Peter met Mary*), the principle of analogy on the part-of-speech level is applied, see (Sgall et al., 1986; Lopatková et al., 2005).

The reason for exploiting the analysis by reduction is obvious: it allows for examining dependencies and word order independently. The method has been described in detail in (Lopatková et al., 2005),its formal modeling by means of restarting automata can be found in (Jančar et al., 1999; Otto, 2006; Plátek et al., 2010). A brief description of its basic principles follows in Section 2.

There is a number of approaches aiming at formalization of word order complexity – let us mention especially the notions of non-projectivity (Marcus, 1965; Holan et al., 2000), (multi-)planarity, gap-degree and well-nestedness – a thorough overview is provided in (Kuhlmann and Nivre, 2006). All these approaches are based on the interplay between the ordering introduced by edges in a tree and the linear ordering of tree nodes. All these approaches look at the problem from the point of view of *complexity of word order*.

An alternative approach to the problem of measuring the word-order freedom has been introduced in (Kuboň et al., 2012). This approach is based on a number of word order shifts[2] necessary for correct analysis by reduction. Contrary to tree-based measures, the number of shifts can somehow express the degree of word order freedom (or the number of *strict word-order constraints* applied). It was shown that number of shifts is 'orthogonal' to the non-projectivity, see (Kuboň et al., 2012).

The experiments in (Kuboň et al., 2012) showed that – with only basic constraints on the analysis by reduction carried out on the limited set of sentences from the PDT – the minimal number of shifts enforced did not exceed one. Here we present a special construction in Czech requiring at least two shifts (Section 4.1), which disproves the hypothesis.

2 Methodology – Analysis by Reduction

Let us first describe the main ideas behind the method used for sentence analysis. *Analysis by reduction (AR)* is based on a stepwise simplification of an analyzed sentence. It defines possible sequences of reductions (deletions) in the sentence – each step of AR is represented by *deleting* at least one word of the input sentence; in specific cases, deleting is accompanied by a *shift* of a word form to another word order position.

Let us stress the basic constraints imposed on the analysis by reduction, namely:
(i) the obvious constraint on preserving individual word forms, their morphological characteristics and/or their surface dependency relations, and
(ii) the constraint on preserving the correctness (a

[1] (Tesnière, 1959) considers linear order vs. structural order and also divides the structural relationships between connexion (now dependency) and junction (coordination).

[2] The shift operation should not be confused with (syntactic) movement in transformational or derivational theories as it is not limited to discontinuous constituents or displacement.

grammatically correct sentence must remain correct after its simplification).

Note that the possible order(s) of reductions reflect dependency relations between individual sentence members, as it is described in (Plátek et al., 2010). The basic principles of AR can be illustrated on the following Czech sentence (1).

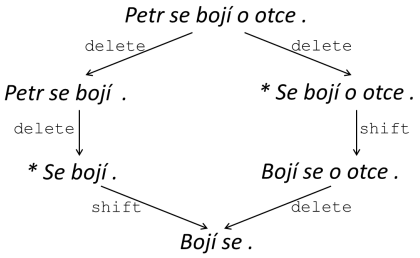
Example 1.

Petr se bojí o otce.

'Peter - *refl* - fears - for - father'

'Peter fears for his father.'

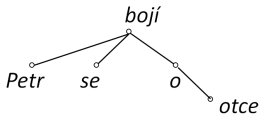
The analysis by reduction can be summarized in the following scheme:



Sentence (1) can be simplified in two ways:
(i) Either by simple deletion of the prepositional group *o otce* 'for father' (following the constraint on correctness of the simplified sentence, the pair of word forms must be deleted in a single step; see the left branch of the scheme).
(ii) Or by deleting the subject *Petr* (the right part of the scheme).[3] However, this simplification results in an incorrect word order variant starting with a clitic[4] **Se bojí o otce*; thus the change of word order (the shift operation) is enforced $\rightarrow_{shift}$ *Bojí se o otce*.

Now, we can proceed in a similar way until we get the minimal correct simplified sentence *Bojí se.*

We can notice that the order of reductions reflects the dependency relations in the corresponding dependency tree. Informally, the words are 'cut from the bottom of the tree'; i.e., a governing node must be preserved in a simplified sentence until all its dependent words are deleted, see (Lopatková et al., 2005).[5] In other words, AR corresponds to the dependency tree for sentence (1):



2.1 Exploiting AR in Experiments

Let us now briefly list the conditions which we have applied in our experiments with AR:

1. *Data selection.*
As we have already mentioned, we focus on the interplay between dependency relations in a sentence (i.e., binary relations between modified and modifying sentence members) and its (linear) word order. Thus, in the initial phase of our investigations, we concentrate on sentences which do not contain phenomena of obviously non-dependent character (esp. coordination, apposition, and parentheses). We also focus only on sentences with a single finite verb (and thus typically consisting of a single clause only).

2. *Shift limitations – the application of the shift operation is limited to cases where it is enforced by the correctness preserving principle of AR.*
In other words, shift operation can be applied only in those cases where a simple deletion would result in a sentence with erroneous word order and a shift (word order modification) can correct it, as in sentence (1).

3. *Optimality – we presuppose a choice of an 'optimal' shift.*
Although we are working with a single syntactic structure for a sentence, there are typically several possibilities how to perform AR (as in sentence (1) with two possible branches of AR). We focus on those branches of AR that show a minimal number of shifts. However, the condition of optimality may sometimes be difficult to achieve, the optimal shifts are not obvious in complicated sentences combining more linguistic phenomena, as it is discussed in Section 4.2.2.

4. *Projectivity – we allow only for projective reductions.*
Reduction of non-projective dependencies is not

[3]Note that Czech is a pro-drop (null-subject) language. Thus it is possible to reduce a sentence subject (if present at all) at any moment – provided that all words depending on the subject have already been reduced; the sentence remains syntactically correct.

[4]Czech has strict grammatical rules for clitics – roughly speaking, they are usually located on the sentence second (Wackernagel's) position, see esp. (Avgustinova and Oliva, 1997; Hana, 2007).

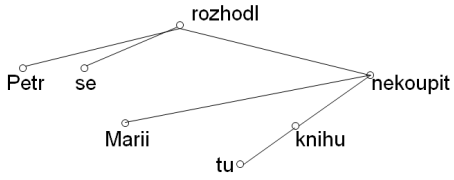
[5]As described in the cited article, the relations between the preposition and its 'head' noun as well as between the verb and his clitic is rather technical as both words involved in the relation must be reduced within a single step. Here we adhere to the practice used for the PDT annotation.

allowed.[6] In other words, a dependent word in a distant position cannot be deleted (with the only exception of limited technical non-projectivities caused, e.g., by prepositions).

The constraint allowing only projective reductions makes it possible to describe a core projective word order. It shows that – even within projective constructions – certain constraints on word order exist, esp. in connection with the position of clitics. Thus a measure based on the number of necessary shifts does not correlate with non-projectivity, see also (Kuboň et al., 2012).

Let us demonstrate the processing of non-projective reductions on the following example (2) (based on (Holan et al., 2000), modified).

Example 2.
Petr se Marii rozhodl tu knihu nekoupit.
'Peter - *refl* - Mary - decided - the book - not-to-buy'
'Peter decided not to buy the book to Mary.'



The word *Marii* (indirect object of the verb *nekoupit* 'not-to-buy') cannot be reduced as it is 'separated' from its governing verb by the main predicate *rozhodl* 'decided' (i.e., by the root of the dependency tree) and thus the relation *Marii – nekoupit* 'to-Mary – not-to buy' is represented by the non-projective edge in the dependency tree. Thus within projective AR, the shift must be performed to make the reduction possible: $\rightarrow_{shift}$ *Petr se rozhodl Marii tu knihu nekoupit.* $\rightarrow_{delete}$ *Petr se rozhodl tu knihu nekoupit.*

3 A Semi-Automatic Application of AR on the PDT Data

3.1 The Data

For humans, especially for native speakers of a particular natural language, it is easy to apply the analysis by reduction, at least when simple sentences are concerned. However, this application exploits the fact that the human understands the sentence and that (s)he is naturally able to reduce it step by step. When we are aiming at applying AR automatically, we have to 'substitute' (at least to some extent) the understanding using the syntactically annotated data (with subsequent manual correctness checking).

For our experiments, we make use of the data from the Prague Dependency Treebank 2.0 (PDT, see (Hajič et al., 2006)).[7] The syntactic structure – a single dependency tree for a single sentence – actually guided the process of AR.

The PDT contains very detailed annotation of almost 49,500 Czech sentences. The annotation is performed at multiple layers, out of which the analytical (surface syntactic) layer is the most relevant for our experiments; we are taking into account only training data (38,727 sentences).

3.2 Searching the Data

For obtaining a suitable set of test sentences for AR as well as for searching the data, we exploit a PML-TQ search tool. PML-TQ is a query language and search engine designed for querying annotated linguistic data (Pajas and Štěpánek, 2009), based on the TrEd toolkit (Pajas and Štěpánek, 2008). TrEd with the PML-TQ extension (primarily designed for processing the PDT data) allows users to formulate complex queries on richly annotated linguistic data. Let us exemplify some types of queries used for obtaining the subset of the PDT data for automatic analysis by reduction.

The output of the first (simplified) example query provides a set of test sentences with the required properties (line 3 - sentence length is limited to 10-25 tokens; 4,5 - no coordination and apposition nodes; 6,7 - no parentheses; 8,9 - just one finite verb; 10,11 - no numerals in test sentences):

```
1 t-root
2 [atree.rf a-root $r :=
3   [descendants()≥ 10, descendants()≤ 25,
4     0x descendant a-node
5     [afun in {"Coord", "Apos"}],
6     0x descendant a-node
7     [is_parenthesis_root="1"],
8     1x descendant a-node
9     [m/tag ~ "^V[Bipqt]"],
10    0x descendant a-node
11    [m/tag ~ "^C" ] ] ] ;
```

Out of the 38,727 sentences of the training data of PDT, only 2,453 sentences remained after the application of this preprocessing filter. Although this number constitutes only 6.33% of the training set, it is still too big for manual testing. This fact

[6] Informally, projective constructions meet the following constraint: having two words n_{gov} and n_{dep}, the second one being dependent on the first one – then all words between these two words must also (transitively) depend on n_{gov}.

[7] http://ufal.mff.cuni.cz/pdt2.0/

clearly shows the necessity of a semi-automatic method of applying AR to the data.

The second query gives a set of non-projective sentences from PDT where the non-projectivity is not caused by a preposition (AuxP) or by emphasizing words (AuxZ) or by some particles (AuzY) (line 5). Note that the output must be filtered in order to merge possible multiple results (line 9).

```
1 t-root
2 [atree.rf a-root $r :=
3   [descendant a-node $p :=
4    [1+x same-tree-as a-node
5     [ afun !~ "Aux[PYZ]$", !ancestor $p,
6     ((ord < $c.ord & ord > $p.ord)
7     \/ (ord > $c.ord & ord < $p.ord)) ],
8     a-node $c := [ ] ] ] ] ;
9 >> for file() & "#" & $r.id
      give $1 sort by $1
```

The second query gives 6,357 non-projective sentences (out of 38,727 sentences in the training data, i.e. 16.41%), in which non-projectivity is not caused by the 'technical' decisions how to annotate prepositions, particles and emphasizing words.

3.3 The Automatic AR Procedure

The automatization of the AR requires a very careful approach. It is necessary to guarantee the correctness of the analyzed sentences in each step of the AR. Let us briefly sketch individual rules guiding the automatized AR:

1. *Reduction rules.* The process is oriented bottom-up, it starts with the leaves of the dependency tree and it removes all nodes marked by analytical functions for attributes, adverbials, objects and subjects, whenever possible. One very important word-order condition is preserved, namely the one guaranteeing that the neighboring nodes are removed first, followed by those which are connected by projective edges.

2. *Preserving non-projectivity.* A node cannot be reduced if this reduction would result in some non-projective edge becoming projective.

3. *Prepositions.* If a node is governed by a preposition, it is necessary to reduce both nodes at once, in a single step. This also has a consequence for the relationship of immediate neighbourhood – prepositions are ignored in this relationship. Prepositions are also ignored when projectivity is tested – i.e., if the only source of a non-projective edge is a preposition, the sentence is treated as projective (this is justified by rather technical annotation of prepositions in PDT).

4. *Clitics.* Clitics may be reduced only together with their governing word. There is also one very important constraint preventing ungrammatical constructions – no reduction may be performed which would leave a clitic on the first sentence position.

5. *Comparison.* Pairwise constructions *čím – tím* 'the – the' cannot be reduced. Other types of comparisons *jako, než* 'as, than' are being reduced together with their last children.

6. *Particles.* Particles are in principle being reduced with regard to the word order constraint, unless they belong to a set of special cases – *coby, jako, jakoby, jakožto* 'as, like' are being reduced together with their parent, similarly as in the case of comparison.

7. *Emphasizing expressions.* If the word order permits it, they can be reduced in the same way as, e.g., adverbials. If a prepositional group is involved, it is reduced as a single unit.

8. *Punctuation and graphical symbols.* Reduction can be applied when the governing word is being reduced.

9. *Full stop.* Sentence ending punctuation is reduced as a final step of AR.

Note that in some cases, we do not insist on a complete reduction (with only the predicate left at the end). Even with the set of test sentences mentioned above and the incomplete reductions, the automatic AR gives us interesting results – see the resulting tables in the following section. Apart from the numerical results, this approach also helped to identify other minor phenomena which do not have a dependency nature.

3.4 Analysis of the Results of the Automatic Procedure

Here we quantify and analyse the results of the automatic AR applied on the test sentences from the PDT. First of all, the following table provides numbers of sentences where specific problematic phenomena appear (from the complete set of the training data from PDT, i.e., from 38,727 sentences).

	phenomenon
12,345	sentences containing clitic(s)
	out of which 3,244 non-projective (26.3%)
850	with the comparison or complement introduced by *coby, jako, jakoby, jakožto*
	out of which 451 non-projective (53.1%)
895	with the comparison expressed by *než*
	out of which 323 non-projective (36.1%)
844	with the comparison with ellipsis
	out of which 302 non-projective (35.8%)
32	with the comparison expressed by *čím-tím*
	out of which 17 non-projective (53.1%)

Let us mention the reasons why we consider these phenomena problematic from the point of view of AR. First, clitics have a strictly specified position in a Czech sentence; thus they may cause a complex word order (including number of non-projective edges, see Section 4). Second, a comparison (frequently accompanied by ellipses) has also complex and non-dependency character.

Let us now look at the results of simple (projective) reductions as described in the previous subsection. The first column describes the number of nodes (= word forms) to which the processed sentences were reduced; the second column gives the number of corresponding sentences and the third column gives their proportion with respect to the whole test set of 2,453 sentences:

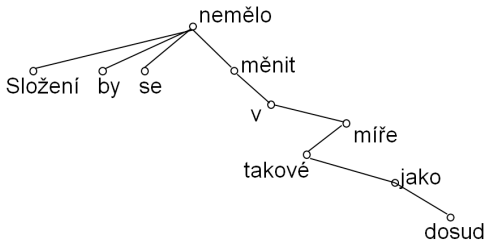
nodes	sentences	%	cumulative coverage
1	1,640	66.86	
2	29	1.18	68.04
3	354	14.43	82.47
4	235	9.58	92.05
5	113	4.61	96.66
6	44	1.79	98.45
7	21	0.86	99.31
8	10	0.41	99.72
9	5	0.20	99.92
10	2	0.08	100.00

We can see that our 'careful' automatic model of simple AR (projective AR without shifts) can process almost 67% of the test set (plus 15.6% sentences are reduced into simple structures with 2 or 3 nodes). Note that 282 (out of 2,453 test sentences) 308 sentences were non-projective (i.e., 11.50% sentences cannot be fully reduced in the course of projective AR).

After a manual analysis of the sentences that were reduced automatically to two nodes (29 in total), we can see that 23 sentences contain a clitic (dependent on the predicate) that prevents the full reduction, or an auxiliary verb (6 cases) or punctuation (1 case) (both auxiliary verbs and punctuation are represented as separate nodes in PDT). Further, 5 sentences which start with subordinating conjunction complete the list (as, e.g., → *Že rozeznáte* 'That (you) recognize').

resulting in 2 nodes	phenomenon	resulting in 3 nodes
29	total sentences	354
23	clitic(s)	310
6/1	aux. verb / punctuation	74
n/a	non-projectivity	37
5	others	0

In order to illustrate the most complicated cases, let us look at one sentence from the 'bottom' part of the table. → *Složení by se nemělo měnit v takové míře jako dosud.* 'The composition should not keep changing in such a degree as so far.' (10 nodes remain as a result of the simple AR).



The first word *Složení* must be preserved in order to preserve correct position of the clitics *by* and *se* (an auxiliary verb and a reflexive particle); further, the non-projective edge *takové – jako dosud* 'such – as so far' in the comparison ('separated' by the governing node *míře* 'degree') stops the process of AR.

The results presented in the previous tables actually support the claim that the automatic procedure works surprisingly well given the complexity of the task. It is able to reduce more than 92% of input sentences to trees with 4 or less nodes. On top of that, it fails to reduce the tree (by a failure we understand the reduction to 7 or more nodes) in 1.55% of cases only.

4 Manual Analysis of Sentences Requiring a Shift within AR

Let us focus on sentences that cannot be reduced (in the course of AR) by simple step-by-step deletion: such attempt would result in a sentence with incorrect word order, see sentence (1). In order to deepen our understanding of the phenomena under

investigation, we decided to analyze selected sentences manually. A further automatization might be attempted in the subsequent phases of our investigation.

As it was shown in the previous sections, the role of shifts during the analysis by reduction is twofold:
1. *To keep the correctness preserving constraint*, which concerns primarily the cases when an input sentence contains a clitic (as in sentence (1)); this issue is addressed in Section 4.1.
2. *To enable projective AR of non-projective sentences*, as it was exemplified on sentences (2); this issue is addressed in Section 4.2.

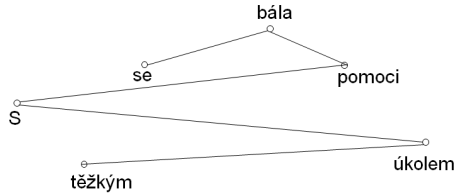
We will present the analysis of these interesting cases step by step, by looking at typical examples. For the sake of simplicity, we will present only 'optimal' branches of AR, i.e., those branches that require a minimal number of shifts (see principle 3 in Section 2.1). This is a purely technical simplification, we are looking for *minimal* necessary number, therefore investigating all possible branches does not make sense, it would give identical results as our 'optimal' approach.

4.1 Number of Necessary Shifts within AR

The crucial question is how many shifts are necessary. The first attempt to get some estimation of the maximal number of necessary shifts in Czech sentences described in (Kuboň et al., 2012) suggested that this number might equal one. This observation had been performed on a small sample of PDT. However, our further research, which included some additional interesting examples created introspectively, indicated that this number might be higher even if the principle of projectivity is not applied, i.e., if we allow for non-projective reductions.

First of all, let us present a counter example to the claim published in (Kuboň et al., 2012) concerning the number of necessary shifts (≤ 1) in Czech sentences. The following sentence requires at least two shifts in the course of the AR (note that the sentence is non-projective):

Example 3.
S těžkým se bála pomoci úkolem.
'with - difficult - *refl* - (she) was afraid - to help - task'
'With a *difficult* task, she wanted to help.'



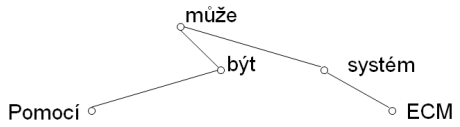
Due to the dependency relations present in the sentence there is only one possibility how to reduce it, the reduction of the adjective *těžkým* 'difficult'. Unfortunately, it results in syntactically incorrect word order: $\rightarrow_{delete}$ **S se bála pomoci úkolem.*
This situation can be corrected in two possible ways, we will sketch only one of them:
$\rightarrow_{shift}$ *S úkolem se bála pomoci* . (A shift of the noun *úkolem* 'task' next to the preposition.)
$\rightarrow_{delete}$ **Se bála pomoci* . (Unfortunately, the next reduction must remove the prepositional group *s úkolem* 'with task' making the sentence again ungrammatical.)
$\rightarrow_{shift}$ *Bála se pomoci.* (Now we can repair the sentence by shifting the verb *bála* to the left.)
The same result will be gained in other branches of AR.

Regardless of the possible reduction sequences, it is necessary to apply at least two shifts. However, although the sample sentence is rather strange, the splitting of the prepositional group is a grammatical construction in Czech. It allows to put a strong stress on an adjective modifying the noun and not on the whole prepositional group, see (Hajičová and Sgall, 2004).

4.2 Projectivization within AR

The principle of projectivity (Section 2.1) constitutes a relatively strong constraint on AR. Its role may be illustrated by the following example of a (simplified) sentence from PDT.

Example 4.
Pomocí může být systém ECM.
'help - can - to be - system - ECM'
'The ECM system may be a help.'



The first two steps are easy, we will get rid of the subject *systém ECM* 'the ECM system' by a stepwise deletion: $\rightarrow$ *Pomocí může být.*
The remaining three words constitute a non-projective 'core' of the original sentence with the non-projective edge *být – pomocí* 'to be – help'.

The AR with non-projective reductions (without the principle of projectivity applied) would not require any shift operation, the word *pomocí* 'help' would be reduce first, followed by the verb *být* 'to be'; these steps would result in the correct simplified sentence *Může.* '(It) can'. However, with the principle of projectivity we have to make the sentence projective first, otherwise no reductions would be possible. For this, we have the following options:
(a) We can make the sentence projective by shifting the *dependent* word pomocí 'help': → *Může pomocí být.* (or → *Může být pomocí.*)
(b) We can also make it projective by shifting the *governing* word *být*: → *Pomocí být může.* (or → *Být pomocí může.*).
The application of both options (a) and (b) in example (4) requires one shift, so the score with the principle of projectivity applied (i.e., only projective reductions are allowed) increases.

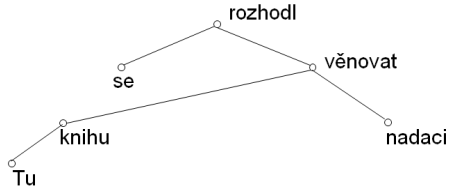
In general, it is also possible to use (c) a shift of the *main verb* of the sentence. If a non-projective core of the sentence has a simple structure with only a single non-projective edge involved, the shift of the main verb has the same results as either (a) or (b). However, in general (with more non-projective edges present in the core of the sentence), the shift of the main verb may result in a word order different from those achieved by the options (a) and (b), see esp. example (6) below.

4.2.1 Clitics and Non-Projectivity in Projective AR

The results on the test sample without the principle of projectivity applied showed that the number of non-projective constructions in a sentence and the number of clitics are not directly reflected in the necessary number of shifts (presented in (Kuboň et al., 2012)).

With the principle of projectivity (i.e., only projective reductions are allowed), the sentences requiring more than one shift not necessarily contain any special constructions, just the combination of clitics and non-projectivities is enough to raise the number of shifts over one:

Example 5.
Tu knihu se rozhodl věnovat nadaci.
'this - book - *refl* - decided - donate - to a foundation'
'He decided to donate this book to a foundation.'

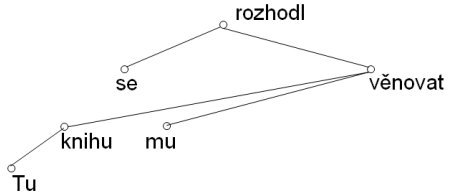


The first two deletions are obvious, the words *tu* 'this' and *nadaci* 'foundation' can be reduced in an arbitrary order: → *Knihu se rozhodl věnovat.*
There are two possibilities how to projectivize the sentence, (a) shifting the dependent word, or (b) shifting the governing word, as mentioned in example (4). Let us sketch here only the former variant (the latter results in the same number of shifts): $\rightarrow_{shift}$ **Se rozhodl knihu věnovat.* (Reduction of the dependent word *knihu* 'book'.)
This shift results in the ungrammatical sentence, therefore it is necessary to perform a shift operation again, this time by shifting the predicate of the sentence to the sentence first position (thus eliminating the ungrammaticality caused by the clitic in the first position).
$\rightarrow_{shift}$ *Rozhodl se knihu věnovat.*
The remaining reductions are then obvious:
$\rightarrow_{delete}$ *Rozhodl se věnovat.* $\rightarrow_{delete}$ *Rozhodl se.*
Regardless of the variant used, we arrive at a score of 2 shifts. This actually indicates that the constraints applied to the AR help to capture the interplay of clitics and non-projectivities in a more subtle way than the original measure presented in (Kuboň et al., 2012).

4.2.2 Number of Shifts in Projective AR

Let us now show that the resulting number of shifts cannot be simply calculated as a sum of the number of non-projective constructions and the number of clitics in a sentence. The following sentence contains two instances of each phenomenon – the clitics *se* and *mu* and the non-projective dependency edges *knihu – věnovat* 'book – to donate', and *mu – věnovat* 'him – to donate', respectively:

Example 6.
Tu knihu se mu rozhodl věnovat.
'this - book - *refl* - him - (he) decided - donate'
'(He) decided to donate this book to him.'

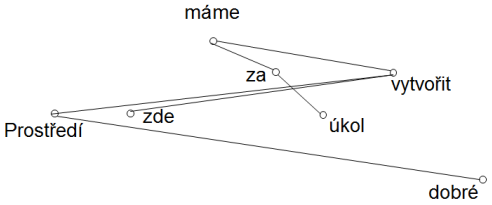


Let us now quickly sketch the AR.
$\rightarrow_{delete}$ *Knihu se mu rozhodl věnovat.* (Reduction of the pronoun *tu* 'this'.)
If we now apply the shift operation on the pronoun *mu* 'him' with the aim at reducing the number of non-projectivities, we will get $\rightarrow_{shift}$ *Knihu se rozhodl mu věnovat*.[8] After the reduction of the dependent pronoun *mu* 'him' we will get one of the intermediate results from the previous example (5), $\rightarrow_{delete}$ *Knihu se rozhodl věnovat*. We already know that in order to reduce this sentence completely we need two more shift operations, therefore the total number of shifts reaches 3.
However, in this case the application of the option (c) mentioned in example (4), shifting the main verb, Section 4.2, brings a better result. If (after the first projective reduction of *tu* 'this') we now shift the word *knihu* 'book' to a projective position $\rightarrow_{shift}$ **Se mu rozhodl knihu věnovat.*, a complementary second shift of the main verb *rozhodl* 'decided' will make the sentence projective (and grammatically correct) $\rightarrow_{shift}$ *Rozhodl se mu knihu věnovat.* and by subsequent application of the reduction of dependent words *mu* 'him' and *knihu* 'book' in an arbitrary order we will get $\rightarrow$ *Rozhodl se věnovat.* This sentence can be further reduced $\rightarrow$ *Rozhodl se.* Overall, only 2 shift operations are necessary in this case (regardless of the number of the studied phenomena involved).

Searching for the minimal necessary number of shifts may be relatively complicated even for sentences whose complexity is lower than in the previous examples. The naive estimation of a number of necessary shift operations for projective reduction in the course of AR can rely on a number of clitics and a number of non-projective edges. However, the next example shows that a single shift operation may 'fix' several non-projectivities. It also shows an example of a sentence where the complex word order is not caused by a clitic.

Example 7.
Prostředí zde máme za úkol vytvořit dobré.
'environment - here - (we) have - as a task - to create - good'
'We have a task to create a good environment here.'

[8]The group *Knihu se rozhodl* may be understood as a single unit, see (Hana, 2007), and thus the clitic *mu* 'him' still occupies the correct 'sentence second' position.



This sentence contains a topicalized noun *prostředí* 'environment'. The dependency tree includes three non-projective edges which are caused by the topicalization. Again, the first reduction is simple and straightforward, the prepositional group *za úkol* can be reduced immediately: $\rightarrow$ *Prostředí zde máme vytvořit dobré.* Then we have several possibilities in which order and by what type of shift to proceed. Again, we will focus on (one of) the 'optimal' sequences of reductions:
$\rightarrow_{shift}$ *Prostředí máme zde vytvořit dobré.* (The reduction of *zde* 'here' is possible only after a shift moves it next to the governing infinite verb *vytvořit* 'create'.)
$\rightarrow_{delete}$ *Prostředí máme vytvořit dobré.*
$\rightarrow_{shift}$ *Máme vytvořit dobré prostředí.* (Here we are shifting the governing noun *prostředí*.)
$\rightarrow_{delete}$ *Máme vytvořit prostředí.* (The reduction of the dependent adjective *dobré* 'good'.)
$\rightarrow_{delete}$ *Máme vytvořit.* (Reduction of the dependent noun *prostředí* 'environment'.)
$\rightarrow_{delete}$ *Máme.* (Final reduction.)
In this 'optimal' branch we have achieved all reductions with the help of only 2 shifts.

This example shows that even without clitics we need at least 2 shifts in the process of projective AR.

Conclusion and Perspectives

In this paper we have tried to achieve a deeper insight into the phenomenon of word order freedom. We have concentrated upon the relationship between (formal) dependencies and word order in Czech. The investigation of this relationship has been performed by means of a semi-automatic analysis of a subset of a large corpus. This analysis proved the applicability of AR on a vast majority of sentences and, at the same time, it helped us to identify problematic phenomena.

Further, manual analysis of complicated sentences proved that the relationship between the number of necessary shifts in the process of AR is orthogonal to the number of clitics or non-projective constructions in a sentence.

Our research helped to analyze concrete phenomena in Czech which influence the word order, namely strict position of clitic(s) and non-projective constructions, and their mutual interplay. The number of necessary shifts with a constraint on projectivity of reductions allows for a more subtle expression of differences between certain configurations of a word order than the measures introduced in previous papers. The range of values of the original measure of word order freedom has been increased.

In the future we would like to continue the research by examining more linguistic phenomena, by testing the measure on other languages with various degrees of word order freedom and by experimenting with a different or modified set of constraints applied on the shift operation. We would also like to expand the research scope to other important phenomena, especially coordination. It would also be interesting to develop a (semi-)automatic method for an optimal application of the shift operation.

Acknowledgments

The research reported in this paper has been supported by the Czech Science Foundation GA ČR, grant No. GA P202/10/1333. This work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of MŠMT (project LM2010013).

References

Tania Avgustinova and Karel Oliva. 1997. On the Nature of the Wackernagel Position in Czech. In *Formale Slavistik*, pages 25–47. Vervuert Verlag, Frankfurt am Main.

František Daneš, Miroslav Grepl, and Zdeněk Hlavsa, editors. 1987. *Mluvnice češtiny 3*. Academia, Praha.

Kim Gerdes and Sylvain Kahane. 2011. Defining dependencies (and constituents). In *Proceedings of DepLing 2011*, pages 17–27, Barcelona.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, M. Mikulová, Z. Žabokrtský, and M. Ševčíková. 2006. *Prague Dependency Treebank 2.0*. LDC, Philadelphia.

Eva Hajičová and Petr Sgall. 2004. Degrees of Contrast and the Topic-Focus Articulation. In *Information Structure – Theoretical and Empirical Aspects*, volume 1, pages 1–13. Walter de Gruyter, Berlin; New York.

Eva Hajičová, Jiří Havelka, Petr Sgall, Kateřina Veselá, and Daniel Zeman. 2004. Issues of Projectivity in the Prague Dependency Treebank. *The Prague Bulletin of Mathematical Linguistics*, 81:5–22.

Jiri Hana. 2007. *Czech Clitics in Higher Order Grammar*. Ph.D. thesis, The Ohio State University.

Tomáš Holan, Vladislav Kuboň, Karel Oliva, and Martin Plátek. 2000. On Complexity of Word Order. *Les grammaires de dépendance – Traitement automatique des langues (TAL)*, 41(1):273–300.

Petr Jančar, František Mráz, Martin Plátek, and Jörg Vogel. 1999. On monotonic automata with a restart operation. *Journal of Automata, Languages and Combinatorics*, 4(4):287–311.

Vladislav Kuboň, Markéta Lopatková, and Martin Plátek. 2012. Studying formal properties of a free word order language. In G. Youngblood and Philip McCarthy, editors, *Proceedings of FLAIRS 25*, pages 300–3005, Palo Alto. AAAI Press.

Marco Kuhlmann and Joakim Nivre. 2006. Mildly non-projective dependency structures. In *Proceedings of the COLING 2006 and ACL 2006 (Poster Sessions)*, pages 507–514, Sydney.

Markéta Lopatková, Martin Plátek, and V. Kuboň. 2005. Modeling Syntax of Free Word-Order Languages: Dependency Analysis by Reduction. In *Proceedings of TSD 2005*, volume 3658 of *LNAI*, pages 140–147, Berlin Heidelberg. Springer-Verlag.

Solomon Marcus. 1965. Sur la notion de projectivité. *Zeitschrift fur Mathematische Logik und Grundlagen der Mathematik*, 11(1):181–192.

Igor A. Mel'čuk. 2011. Dependency in language. In *Proceedings of DepLing 2011*, pages 1–16, Barcelona.

Friedrich Otto. 2006. Restarting Automata. In *Recent Advances in Formal Languages and Applications, Studies in Computational Intelligence*, volume 25, pages 269–303, Berlin. Springer-Verlag.

Petr Pajas and Jan Štěpánek. 2008. Recent Advances in a Feature-Rich Framework for Treebank Annotation. In *Proceedings of CoLING 2008*, volume 2, pages 673–680, Manchester, UK. The Coling 2008 Organizing Committee.

Petr Pajas and Jan Štěpánek. 2009. System for Querying Syntactically Annotated Corpora. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 33–36, Singapore. ACL.

Martin Plátek, František Mráz, and Markéta Lopatková. 2010. (In)Dependencies in Functional Generative Description by Restarting Automata. In *Proceedings of NCMA 2010*, volume 263 of *books@ocg.at*, pages 155–170, Wien. Österreichische Computer Gesellschaft.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel, Dordrecht.

Lucien Tesnière. 1959. *Eléments de syntaxe structurale*. Librairie C. Klincksieck, Paris.