FACULTY OF MATHEMATICS AND PHYSICS
CHARLES UNIVERSITY IN PRAGUE

22nd THEORIETAG

AUTOMATA AND FORMAL LANGUAGES

OCTOBER 3–5, 2012, PRAGUE
PROCEEDINGS

František Mráz (Ed.)

matfyzpress
PRAGUE 2012

ISBN 978-80-7378-221-4

F. Mráz (Ed.): Theorietag 2012, Prague, 3. – 5.10.2012
Faculty of Mathematics and Physics, Charles University in Prague pp. 71 – 76

How to Measure Word Order Freedom for Natural Languages?

Vladislav Kuboň Markéta Lopatková$^{(A)}$ Martin Plátek$^{(A)}$

Charles University in Prague, Faculty of Mathematics and Physics
{vk,lopatkova}@ufal.mff.cuni.cz, martin.platek@.mff.cuni.cz

1. Introduction

In this paper we would like to clarify some basic features and notions which may play a key role in the investigations of the word order freedom. For this purpose we are going to exploit the elementary method of *analysis by reduction* (AR) and the formal data type derived from this method, so-called D-trees. A complete description of both the method and the data type can be found for example in ([7]). Let us remind that the analysis by reduction has served as a motivation for a family of so called restarting automata, see ([6]). The first step in the direction of more formal treatment of the word order freedom has been done in ([2]), where the authors discussed it without the exploitation of the analysis by reduction and without setting the constraints on unchanged morphological and syntactic properties of individual words. We will focus on some examples cited there and modify them according to the methods mentioned in ([7]). We also exploit sample sentences from the Prague Dependency Treebank (PDT),[1] a large-scale treebank of Czech ([1]).

2. The Background of our Experiments

We would like to introduce the notion of a shift operation in the course of the AR, a key notion for the investigation of a measure of word order freedom. In order to be able to define the shift operation, it is useful to introduce the data structure we are exploiting.

The D-tree (Delete or Dependency trees), see e.g. [7], is a rooted ordered tree with edges oriented from its leaves to its root. Nodes of each tree correspond to individual occurrences of word forms in a sentence. Moreover, we suppose a total ordering on the nodes that reflects word order in a sentence.

Let us remind that the concept of D-tree reflects the analysis by reduction (without rewriting) – its structure corresponds to a way how individual words of a sentence are deleted in the course of the corresponding steps of the analysis by reduction. (Informally, each edge of a D-tree connects a word form to some other word form if the latter cannot be deleted earlier then the first word form in (any branch of) analysis by reduction of the same sentence.)

$^{(A)}$The paper reports on the research within the NoSCoM project, supported by the grant of the Grant Agency of the Czech Republic No. P202/10/1333.

[1] http://ufal.mff.cuni.cz/pdt.html

Measures of Non-projectivity and Shift Operation

Non-projectivity. When considering word order freedom, we have to take into account one phenomenon which is common in languages with higher degree of word order freedom, namely non-projective constructions (for previous usage of this term see esp. [5, 4]). In order to classify this phenomenon, it is useful to define certain notions allowing for an easy definition of projectivity/non-projectivity and also for the introduction of measures of non-projectivity (these notions are formally defined in [2]).

The *coverage of a node u of a* D*-tree* identifies nodes from which there is a path to u in the D-tree (including empty path). It is expressed as a set of horizontal positions/indices (expressing total ordering on nodes in a D-tree, see above) of nodes directly or indirectly dependent upon a particular node.

The notion of a coverage leads directly to a notion of *a hole in a subtree*. Such a hole exists if the set of indices in the coverage is not a continuous sequence.

We say that D-tree T is *projective* if none of its subtrees contains a hole; otherwise, T is *non-projective*.

Shift operation. In order to be able to describe necessary word order shifts in the course of AR, we need to define a notion of equivalence for D-trees. Such equivalence (denoted as DP-equivalence) is defined as follows: DP*-equivalent trees* are those D-trees which have (i) the 'same' sets of nodes, i.e., the nodes describing the same set of lexical bundles, and (ii) their edges always connect 'identical' pairs of nodes (nodes with identical lexical bundles). It actually means that a particular set of DP-equivalent trees contains the D-trees representing sentences created by a permutation of the words of the original sentence but having the same dependency relations.

Let T be a D-tree; the set of D-trees which are DP-equivalent to T will be denoted DPE(T). In other words, DPE(T) is a set of D-trees which differ only in the word order of their characteristic sentence.

The previous concepts allow us to introduce a new feature, a *number of reduction steps enforcing a shift* in a single branch of AR. Shifts make it possible to change word order and thus 'recover' from incorrect word order that may be incurred by an AR deleting step. The *shift operation* is such a change in a D-tree when (i) the ordering of all nodes except for one is preserved, and (ii) the edges are preserved (connecting 'identical' pairs of nodes with respect to described lexical bundles). It means that both the original D-tree T and the modified one belong to the same set DPE(T).

Let T be a D-tree, $T \notin$ CT. Our goal is to find – if possible – a modified D-tree T' such that T' is a correct surface tree (i.e., $T' \in$ CT) and T' is DP-equivalent to T (i.e., $T' \in$ DPE(T)) by applying as small number of shift operations as possible.

3. Towards a Measure of Word Order Freedom

3.1. Data

The investigation focuses upon an interplay of two phenomena related to word order: the *non-projectivity* of a sentence and the *number of word order shifts* within the analysis by reduction. This interplay is exemplified on a set of 'suspicious' types of sentences identified in previous work on Czech word order freedom [2]. The sample set was enriched with sentences from the Prague Dependency Treebank (PDT), a large-scale treebank of Czech [1], namely the sentences with a non-projectivity given by a modal verb (typically in combination with clitics [3]). These sentences were manually annotated using the method of analysis by reduction.

3.2. Principles of Data Analysis

The following principles are applied during the analysis of sample data.

Principle 1: *'Preprocessing' – we simplify the input sentences using* AR *in such a way that only words related to these phenomena are preserved.*
In other words, we focus on those branches of AR where the words which do not contribute to the examined structures are already processed and thus deleted (if it is possible without shifting). Let us exemplify this on sentence (3) (shortened sentence from PDT) and its initial simplification:

Example:

(2) *Naše firma by se možná mohla tvářit, že se jí premiérova slova netýkají (. . .).*
'Perhaps our firm might pretend that the prime minister's words do not apply to it (. . .).'
→ *Firma by se mohla tvářit.*
'The firm might pretend.'

Principle 2: *Minimality – we focus especially on those branches of* AR *that allows us to reduce a sentence with minimal number of shifts.*
Typically, there are several possibilities how to analyze a simplified sentence. In our example (2), we can start with reducing the noun *firma* 'firm'. This results in the string starting with clitics *by* and *se* – thus a shift in word order positions must by applied to ensure the correctness of the simplified sentence. We have two possibilities of shifting:
(a) We can SHIFT the verb *tvářit* 'to pretend' to the first position, which results in the correct sentence *Tvářit by se mohla.* However, the only possible subsequent reduction step means deleting the pair *tvářit se* 'to pretend + REFL', which requires another SHIFT *By mohla.* $\rightarrow_{SHIFT}$ *Mohla by.*
Or, (b) we can SHIFT the verb *mohla* 'may' to the first position *Mohla by se tvářit.* The subsequent reduction of the pair *se tvářit* 'REFL + to pretend' does not require another shifting.

This example shows that if we aim at the minimal necessary number of shifts then we must apply the second type of shifting.

Principle 3. *Restriction on the shift operation – the application of the shift operation is limited to cases where it is enforced by the correctness preserving principle of* AR (i.e., to cases where a simple deletion would violate the principle of correctness imposed on AR).

Principle 4: *Non-projectivity – we allow for non-projective reductions.*
Long distance dependencies are allowed, i.e., depending word in a distant (non-projective) position may be deleted.

Example:

(3) *Marii se Petr tu knihu rozhodl nekoupit.*
'to-Mary – REFL – Peter – that/the book – decided – not-to-buy'
'To Mary, Peter decided not to buy the book.'

The word *Marii* (indirect object of the verb *nekoupit* 'not-to-buy') is reduced even though it is 'separated' from its governing verb by the main predicate *rozhodl* 'decided' (i.e., by the root of the dependency tree); the relation *Marii – nekoupit* 'to-Mary – not-to buy' creates a non-projective edge, [2].

Principle 5: *Locality – we limit our observations to simple sentences/clauses containing interesting phenomena.*
This principle allows us to focus on an interplay of several phenomena affecting a single surface syntactic construction (based on principle 1, all 'uninteresting' words are already processed, coordination is simplified etc.); in case of more than one interesting construction in a sentence (prototypically a complex sentence consisting of several clauses), they are processed separately. The reason is simple – if we want to achieve results reflecting the properties of the investigated phenomenon, we have to eliminate chances to construct a complex sentence with an arbitrary number of shifts simply by coordinating a desired number of clauses requiring one shift each (or by inserting a relative clause with a shift).

3.3. How to Measure Word Order Freedom?

The previous work led to the proposal of a measure based on (minimal) number of shifts within an analysis by reduction of a given sentence, see esp. [3]. We can characterize this approach by principles 1-5 mentioned above, i.e., as an analysis by reduction enhanced with the possibility of word order changes. The results proved that the number of shifts is an important factor providing different information than already existing measures reflecting the complexity of word order of individual sentences.

However, the granularity of the proposed measure seemed to be too low as all the inspected sentences from PDT were analyzed with at most one shift operation, regardless the number of holes and number of clitics in a sentence. This result was improved when we subsequently inspected 'suspicious' sentences analyzed in [2]. We have found a construction where at least two shifts are necessary (even when the principle of non-projectivity is applied, i.e. we allow for non-projective reductions).

Example:
(4) *S těžkým se mu bála pomoci úkolem.*
'with – difficult – REFL – him – (she) was afraid – to help – task'
'With a *difficult* task, she was afraid to help him.'
This sentence is rather special Czech surface construction when – due to the stress on the adjectival attribute *těžkým* 'difficult' – the prepositional group *s těžkým úkolem* 'with difficult task' is split and the preposition and adjective are moved to the beginning of the sentence.

The only possible correctness preserving reduction lies in deleting the pronoun *mu* 'him'. With respect to the dependency relations in the sentence, the subsequent reduction step must delete the adjective, but this step results in an ill-formed sentence:
$\rightarrow_{DEL}$ * *S se bála pomoci úkolem.*
Thus a word order correction is enforced:
(a) We can SHIFT the noun *úkolem* 'task' to obtain the (correct) continuous noun group *s úkolem* 'with (the) task'. The reduction of this noun group is then the only reduction possibility, again resulting in the sentence with an incorrect word order. Now, the 'optimal' SHIFT of the main predicate is enforced; the final deletion results in a correct simplified sentence:
$\rightarrow_{SHIFT}$ *S úkolem se bála pomoci.* $\rightarrow_{DEL}$ * *Se bála pomoci.* $\rightarrow_{SHIFT}$ *Bála se pomoci.* $\rightarrow_{DEL}$ *Bála se.*

(b) Alternatively, the preposition *s* ''with' is shifted to create continuous noun group *s úkolem* 'with (the) task', followed by the 'optimal' shift of the main predicate; then the sentence can be reduced by applying simple delete operations:
$\rightarrow_{SHIFT}$ * *Se bála pomoci s úkolem.* $\rightarrow_{SHIFT}$ *Bála se pomoci s úkolem.* $\rightarrow_{DEL} \ldots \rightarrow$ *Bála se.*

In both branches of AR, (at least) two shift operations are necessary to analyze sentence (4) (contrary to the hypothesis made on the basis of corpus data [3]).

This observation, however, does not refine the measure itself, it only increases the range of its values for Czech.

3.4. A proposed refinement of the original measure

It is quite obvious that applying stricter constraints on the delete or shift operations would bring a more refined measure. There are actually at least two possible ways – (i) we can distinguish a type of necessary shifts (e.g., a shift of a verb / a shift across a verb), (ii) the deletion can be limited to adjacent word forms, or (iii) the deletion can be limited to projective reductions (i.e., dependent and governing words may be 'separated' only by word forms (transitively) dependent on the latter one, contrary to Principle 4). So far, we have focused on the third restriction.

Example:
(5) *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 a subject (the ECM system) by a stepwise deletion:
→ *Pomocí může být.*

The remaining three words constitute a non-projective 'core' of the original sentence providing the following options typical for such a case:
(a) We can make the sentence PROJECTIVE by shifting the **dependent** word → *Může pomocí být.*
(b) We can also make it PROJECTIVE by shifting the **governing** word → *Pomocí být může.*
(c) The projectivization mentioned in (a) and (b) can also be achieved by means of a shift of the main verb, in (a) it would represent a shift of the main verb *Může* to the first position in the sentence, in b) to the last one. This option actually only increases the number of possibilities without bringing anything really new. Even worse, shifting the main verb of the sentence may bring additional complications in case that the non-projective core of the input sentence is bigger and more complex than in our example. Therefore it is better to avoid this type of a shift entirely and to concentrate on the shifts under options (a) and (b).

Let us now look at a more complicated example with a clitic combined with a non-projectivity.

Example:
(6) *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.*
Let us now perform the two variants of further reduction according to the options mentioned above:
(a) Let us make the sentence PROJECTIVE by means of shifting the **dependent** word *knihu*
$\rightarrow_{SHIFT}$ * *Se rozhodl knihu věnovat.*
This shift results in an 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:
$\rightarrow_{DEL}$ *Rozhodl se věnovat.* $\rightarrow_{DEL}$ *Rozhodl se.*

(b) In a similar way we can make the sentence PROJECTIVE by means of shifting the **governing** word *věnovat* $\rightarrow_{SHIFT}$ *Knihu věnovat se rozhodl.* This shift results in a sentence which looks like syntactically incorrect one due to the clitic becoming a third word in a sentence, not a second one. However, in

this case, the group *věnovat knihu* may be understood as a single unit and thus the clitic still occupies the sentence second position and we may proceed with a simple reduction:
$\rightarrow_{DEL}$ *Věnovat se rozhodl.* $\rightarrow_{DEL}$ ** Se rozhodl.* This reduction is the only possible and the ungrammaticality of the resulting sentence has to be corrected by a second shift:
$\rightarrow_{SHIFT}$ *Rozhodl se.*

So, again, regardless of the option used, we are arriving at a score of 2 shifts. This actually indicates that the refined measure captures the interplay of clitics and non-projectivities in a more subtle way than the original measure.

4. Conclusion and Future Work

In this paper we have presented the results of a detailed investigation of the phenomenon of word order freedom for a particular natural language, Czech. We have shown, on the basis of several examples, that if we leave the safe grounds of data contained in a syntactically annotated corpus of Czech, we may found sentences exemplifying the complexity of the problem. We have shown that the range of values of the original measure of word order freedom presented in previous papers may be bigger in certain cases, we have also discussed the method how to obtain an exact value for this measure, and, last but not least, we have suggested a refinement of the original measure wchich more adequately captures the interplay of various phenomena.

In the future we would like to continue the research in this direction by examining more linguistic phenomena, by testing the measure on other languages with various degree of word order freedom and by experimenting with a different or modified set of constraints applied on the shift operation.

References

[1] J. Hajič, J. Panevová, E. Hajičová, P. Sgall, P. Pajas, J. Štěpánek, J. Havelka, M. Mikulová, Z. Žabokrtský, M. Ševčíková-Razímová, *Prague Dependency Treebank 2.0*. Linguistic Data Consortium, Philadelphia, PA, USA, 2006.

[2] T. Holan, V. Kuboň, K. Oliva, M. Plátek, On Complexity of Word Order. *Les grammaires de dépendance – Traitement automatique des langues (TAL)* **41** (2000) 1, 273–300.

[3] V. Kuboň, M. Lopatková, M. Plátek, Studying Formal Properties of a Free Word Order Language. In: G. Youngblood, P. McCarthy (eds.), *Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference*. AAAI Press, Palo Alto, California, 2012, 300–3005.

[4] J. Kunze, *Die Auslassbarkeit von Satzteilen bei koordinativen Verbindungen im Deutschen*. Number 14 in Schriften zur Phonetik, Sprachwissenschaft und Kommunikationsforschung, Akademie-Verlag, Berlin, 1972.

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

[6] F. Otto, Restarting Automata and Their Relation to the Chomsky Hierarchy. In: Z. Ésik, Z. Fülöp (eds.), *Proceedings of DLT 2003*. LNCS 2710, Springer-Verlag, Berlin, 2003, 55–74.

[7] M. Plátek, F. Mráz, M. Lopatková, (In)Dependencies in Functional Generative Description by Restarting Automata. In: H. Bordihn, R. Freund, T. Hinze, M. Holzer, M. Kutrib, F. Otto (eds.), *Proceedings of NCMA 2010*. books@ocg.at 263, Österreichische Computer Gesellschaft, Wien, Austria, 2010, 155–170.