The Prague Bulletin of Mathematical Linguistics
NUMBER 90 DECEMBER 2008

EDITORIAL BOARD

Published twice a year by Charles University in Prague

Editorial office and subscription inquiries:
ÚFAL MFF UK, Malostranské náměstí 25, 118 00, Prague 1, Czech Republic
E-mail: pbml@ufal.mff.cuni.cz

ISSN 0032-6585

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 90 DECEMBER 2008

CONTENTS

Articles

Reviews

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 90 DECEMBER 2008 5–32

Netgraph Query Language for the Prague Dependency Treebank 2.0

Jiří Mírovský

Abstract

We study the annotation of the Prague Dependency Treebank 2.0 (PDT 2.0) and assemble a list of requirements on a query language that would allow searching for and studying all linguistic phenomena annotated in the treebank. We propose an extension to the query language of an existing search tool Netgraph 1.0 and show that the extended query language satisfies the list of requirements. We demonstrate how all principal linguistic phenomena annotated in the treebank can be searched for with the proposed query language and compare the query language to some other treebank search systems. The proposed query language has been implemented in the search tool Netgraph – we talk about features of a search tool that can simplify the searching and make it more powerful. We also present a table that shows the extent of usage of various features of the implemented query language by the users of Netgraph and mention several usages of Netgraph for other treebanks than PDT 2.0.

1. Introduction

Linguistically annotated treebanks play an essential role in modern computational linguistics. The more complex the treebanks become, the more sophisticated tools are required for using them, namely for searching in the data. A search tool helps extract useful information from the treebank, in order to study the language, the annotation system or even to search for errors in the annotation. The Prague Dependency Treebank 2.0 (Hajič et al. 2006), which is a sequel to the Prague Dependency Treebank 1.0 (Hajič et al. 2001), is one of the most advanced manually annotated treebanks in the linguistic world.

Three sides existed whose connection needed to be solved. First, it was the Prague Dependency Treebank 2.0 with its extensive annotation. Second, there existed a very limited but extremely intuitive search tool – Netgraph 1.0 (Ondruška 1998). Third, there were users longing for such a simple and intuitive tool that would be powerful enough to search in the Prague Dependency Treebank.

Please cite this article as: Jiří Mírovský, Netgraph Query Language for the Prague Dependency Treebank 2.0. The Prague Bulletin of Mathematical Linguistics No. 90, 2008, 5–32.

We study the annotation of the Prague Dependency Treebank 2.0 (PDT 2.0), especially on the tectogrammatical layer, which is by far the most complex layer of the treebank, and assemble a list of requirements on a query language that would allow searching for and studying all linguistic phenomena annotated in the treebank. We propose an extension to the query language of the existing search tool Netgraph 1.0 and show that the extended query language satisfies the list of requirements. We also demonstrate how all principal linguistic phenomena annotated in the treebank can be searched for with the proposed query language and compare the query language to some other treebank search systems. The proposed query language has also been implemented in the search tool Netgraph – we talk about features of a search tool that can simplify the searching and make it more powerful. We also present a table that shows the extent of usage of various features of the implemented query language by the users of Netgraph and mention several usages of Netgraph for other treebanks than PDT 2.0.

More details about the topics of this contribution can be found in Mírovský (2008e).

2. The Analysis of the Problem

We study linguistic phenomena annotated in PDT 2.0, in order to decide what features a query language of a search tool needs to have to allow searching for these phenomena and studying them (Mírovský 2008d). Afterwards, we summarize the features and formulate a concise list of linguistic requirements on a query language for PDT 2.0.

2.1. Linguistic Phenomena in PDT 2.0

PDT 2.0 has three layers of annotation: the morphological layer (Hana et al. 2005), the analytical layer (Hajič et al. 1997), and the tectogrammatical layer (Hajičová 1998). To be exact, there is one more layer – the word layer – which only keeps the tokenized original data and (apart from the tokenization) does not contain any annotation.

Our work is focused on the two structured layers – the analytical layer and the tectogrammatical layer. We intend to access the morphological information only from the higher layers, not directly. Since there is a 1:1 relation among nodes on the analytical layer (but for the technical root) and tokens on the morphological layer, the morphological information can be easily merged into the analytical layer – the nodes only get additional attributes. We study two ways of accessing the data of PDT 2.0:

- the analytical layer directly, the morphological and word layer information merged into the analytical layer; the tectogrammatical layer inaccessible,
- the tectogrammatical layer directly, the analytical layer "through" this layer, the morphological and word layer annotation merged into the analytical layer.

The difference between these two approaches is not only in the presence of the tectogrammatical layer, but also in the way of accessing the information from the lower layers, which is inevitably caused by the non-1:1 relation between the analytical and the tectogrammatical layer.

Since the tectogrammatical layer is by far the most complex layer in the treebank, we start

our analysis with a study of the annotation manual for the tectogrammatical layer (t-manual, Mikulová et al. 2006) and focus also on the requirements on accessing the lower layers with non-1:1 relations. Afterwards, we add some requirements on the query language concerning the annotation of the lower layers – the analytical layer and the morphological layer.

During the studies, we have to keep in mind that we do not only want to search for a phenomenon, but also need to study it, which can be a much more complex task. Therefore, it is not sufficient e.g. to find a predicative complement, which is a trivial task, since the attribute `functor` of the complement is set to the value `COMPL`. In this particular example, we also need to be able in the query to specify properties of the node the second relation of the complement goes to, e.g. that it is an Actor.

2.1.1. The Tectogrammatical Layer

Basic Principles

The basic unit of annotation on the tectogrammatical layer of PDT 2.0 is a sentence as a basic means of conveying meaning (t-manual, page 8). The representation of the tectogrammatical annotation of a sentence is a rooted dependency tree. (More exactly, a tectogrammatical tree structure, TGTS, see Sgall (2001) for the differences between a TGTS and a theoretically substantiated tectogrammatical representation.) It consists of a set of nodes and a set of edges. One of the nodes is marked as the root. Each node is a complex unit accompanied by a set of attribute-value pairs. The edges express dependency relations between the nodes. The edges do not have their own attributes; attributes that logically belong to the edges (e.g. a type of the dependency) are represented as node-attributes (t-manual, page 9).

This implies the first and most basic requirement on the query language: one result of the search is one sentence along with the tree belonging to it. Also, the query language should be able to express the node evaluation and the tree dependency among nodes in the most direct way.

Valency

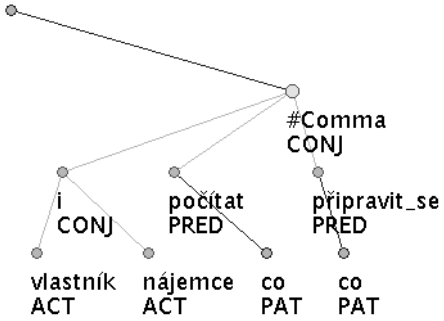
Valency (Hajičová, Panevová 1984) of verbs, valency of verbal nouns, valency of nouns that represent the nominal part of a complex predicate and valency of some adverbs are annotated fully in the trees (t-manual, pages 162-3). Since the valency of verbs is the most complete in the annotation and since the requirements on searching for valency frames of nouns are the same as those concerning verbs, we will (for the sake of simplicity in expressions) focus on the verbs only. Verbs usually have more than one meaning; each is assigned a separate valency frame. Every verb has as many valency frames as it has meanings (t-manual, page 105).

Therefore, the query language has to be able to distinguish valency frames and search for each one of them, at least as long as the valency frames differ in their members and not only in their index. (Two or more identical valency frames may represent different verb meanings (t-manual, page 105).) The required features include a presence of a son, its absence, and a possibility to control the number of sons of a node.

Coordination and Apposition

A tree dependency is not always a linguistic dependency (t-manual, page 9). Coordination and apposition are examples of the phenomenon (t-manual, page 282). If a Predicate governs two coordinated Actors, these Actors depend on a coordinating node and this coordinating node depends on the Predicate. The query language should be able to skip such a coordinating node. In general, there should be a possibility to skip any type of node.

Skipping a given type of node helps but is not sufficient. The coordinated structure can be more complex, for example the Predicate itself can be coordinated too. Then, the Actors do not even belong to the subtree of any of the Predicates. In the following example, the two Predicates ("PRED") are coordinated with conjunction ("CONJ"), as well as the two Actors ("ACT"). The linguistic dependencies go from each of the Actors to each of the Predicates but the tree dependencies are quite different:



In Czech: *S čím mohou vlastníci i nájemci počítat, na co by se měli připravit?*
In English: *What can owners and tenants expect, what should they get ready for?*

The query language should therefore be able to express the linguistic dependency directly. The information about the linguistic dependency, as well as many other phenomena, is annotated in the treebank by means of references (see Coreferences below).

Other Phenomena

Similarly, other phenomena annotated in the treebank are studied in Mírovský (2008e). For the lack of space in this paper, let us only briefly list the phenomena and their requirements.

Idioms (Phrasemes) etc.

The query language has to offer at least a basic searching procedure in the linear form of the sentences, to allow searching for any idiom or phraseme, regardless of the way it is or is not captured in the tectogrammatical tree. It can even help in a situation when the user does not know how a certain linguistic phenomenon is annotated on the tectogrammatical layer.

Complex Predicates

There are problematic cases of annotation of complex predicates where the expressed valency modification occurs in the same form in the valency frames of both components of the complex predicate (t-manual, page 362).

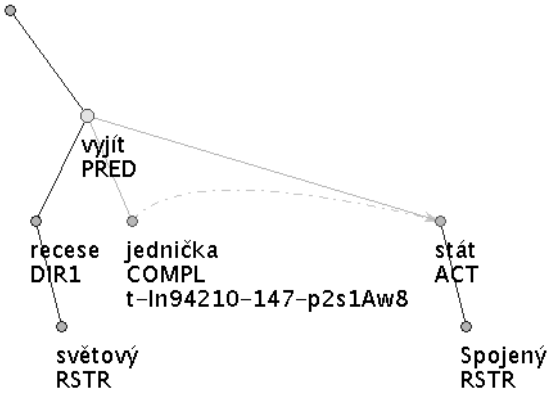
To study these special cases of valency, the query language has to offer a possibility to define that a valency member of the verbal part of a complex predicate is at the same time a valency member of the nominal part of the complex predicate, possibly with a different function. The identity of valency members is annotated by means of references, which is explained later (see Coreferences below).

Predicative Complement (Dual Relations)

The predicative complement is a non-obligatory free modification (adjunct) which has a dual semantic relation. It simultaneously modifies a noun and a verb (which can be nominalized). These two relations are represented by different means (t-manual, page 376):

- the relation to a verb is represented by means of an edge (which means it is represented in the same way as other modifications),
- the relation to a noun is represented by means of attribute `compl.rf`, the value of which is the identifier of the modified noun.

In the following example, the predicative complement ("`COMPL`") has one relation to the verb ("`PRED`") and another (dual) relation to the noun ("`ACT`"):



In Czech: *Ze světové recese vyšly jako jednička Spojené státy.*
In English: *The United States emerged from the world recession as number one.*

The second form of relation, represented once again with references (still see Coreferences just below), has to be expressible in the query language.

Coreferences

Grammatical coreference and textual coreference (Kučová et al. 2003) are annotated on the tectogrammatical layer. The current way of representing coreference uses references (t-manual, page 996).

Let us finally explain what references are. References are pointers, a technical way of expressing various relations between nodes[1]. They make use of the fact that every node of every tree has an identifier (the value of the attribute `id`), which is unique within PDT 2.0. If coreference, dual relation, or valency member identity is merely a link between two nodes (one node referring to another), it is enough to specify the identifier of the referred node in an appropriate attribute of the referring node. Reference types are distinguished by different referring attributes. Individual reference subtypes can be further distinguished by the value of another attribute.

The essential point in references (for the query language) is that at the time of forming a query, the value of the reference is unknown. For example, in the case of dual relation of the predicative complement, we know that the value of the attribute `compl.rf` of the complement must be the same as the value of the attribute `id` of the governing noun, but the value itself differs tree from tree and therefore is unknown at the time of creating the query. The query language has to offer a possibility to bind these unknown values.

Communicative Dynamism

Communicative dynamism (underlying order of nodes, cf. Hajičová et al. 1998) requires that the relative order of nodes in the tree from left to right can be expressed. The order of nodes is controlled by the attribute `deepord`, which contains a non-negative real (usually natural) number that sets the order of the nodes in the tree from left to right. Therefore, we will again need to refer to a value of an attribute of another node but this time with a relation other than "equal to".

Focus Proper and Quasi-Focus

Focus proper is the most dynamic and communicatively significant contextually non-bound part of the sentence. Focus proper is placed on the rightmost path leading from the effective root of the tectogrammatical tree, even though it occupies a different position in the surface structure. The node representing this expression will be placed rightmost in the tectogrammatical tree (t-manual, page 1129).

Quasi-focus is constituted by a contextually bound expression, on which the focus proper is dependent. The focus proper can immediately depend on the quasi-focus, or it can be a more deeply embedded expression. In the underlying word order, nodes representing the quasi-focus, although they are contextually bound, are placed to the right from their governing node.

[1] References are a technical term. They should not be confused with linguistic terms like coreferences.

Nodes representing the quasi-focus are therefore contextually bound nodes on the rightmost path in the tectogrammatical tree (t-manual, page 1130).

The ability of the query language to distinguish the rightmost node in the tree and the rightmost path leading from a node is therefore necessary.

Focalizers

Focalizers are expressions whose function is to signal the topic-focus articulation categories in the sentence, namely the communicatively most important categories – the focus and the contrastive topic.

Focalizers bring a further requirement on the query language – an ability to control the distance between nodes (in the terms of deep word order); at the very least, the query language has to distinguish an immediate brother and the relative horizontal position of nodes.

(Non-)Projectivity

Projectivity (Havelka 2007) is defined as follows: between a father and its son there can only be direct or indirect sons of the father (t-manual, page 1135).

The relative position of a node (node A) and an edge (nodes B, C) that together cause a non-projectivity forms four different configurations: ("B is to the left from C" or "B is to the right from C") x ("A is on the path from B to the root" or "it is not").

To be able to search for all configurations in one query, the query language should be able to combine several queries into one multi-query. We do not require that a general logical expression can be set above the single queries. We only require a general OR combination of the single queries.

2.1.2. Accessing Lower Layers

Studies of many linguistic phenomena require a multilayer access. For example, the query "find an example of a Patient that is more dynamic than its governing Predicate (with greater `deepord`) but on the surface layer is on the left side from the Predicate" requires information both from the tectogrammatical layer and the analytical layer (for the study of these phenomena, see Hajičová 2007).

As we have already said, information from the lower layers can be easily compressed into the analytical layer, since there is a 1:1 relation among tokens/nodes of the layers (with some rare exceptions like misprints on the w-layer). The interrelationship between the tectogrammatical layer and the analytical layer is much more complex. Several nodes from the analytical layer may be (and often are) represented by one node on the tectogrammatical layer and new nodes without an analytical counterpart may appear on the tectogrammatical layer. It is necessary that the query language addresses this issue and allows access to the information from the lower layers.

2.1.3. The Analytical Layer (and Lower Layers)

Requirements (on a query language) of most linguistic phenomena annotated on the analytical layer have already been covered in the previous section, discussing the tectogrammatical layer. The lower layers only supplement a few additional requirements.

Morphological Tags

In PDT 2.0, morphological tags are positional. They consist of 15 characters, each representing a certain morphological category.

The query language has to offer a possibility to specify a part of the tag and leave the rest unspecified. It has to be able to set such conditions on the tag as "this is a noun", or "this is a plural in the accusative". Some conditions might include negation or enumeration, like "this is an adjective that is not in the accusative", or "this is a noun either in the dative or the accusative". This is best done with some sort of wild cards. The latter two examples suggest that such a powerful tool as regular expressions may be needed.

Agreement

There are several cases of agreement in the Czech language, like agreement in case, number and gender in attributive adjective phrases, agreement in gender, person and number between predicate and subject (though it may be complex), or agreement in case in apposition.

To study agreement, the query language has to allow to make a reference to only a part of a value of an attribute of another node, e.g. to the fifth position of the morphological tag for case.

Word Order

Word order is a linguistic phenomenon widely studied on the analytical layer, because this layer offers a perfect combination of word order (the same as in the sentence) and syntactic relations between the words. The same technique as with the deep word order on the tectogrammatical layer can be used here. The order of words (tokens) and also nodes in the analytical tree is controlled by the attribute `ord`.

The only new requirement on a query language is an ability to measure the horizontal distance between words, to satisfy linguistic queries like "find trees where a preposition and the head of the noun phrase are at least five words apart".

2.2. Linguistic Requirements

Let us summarize what features a query language has to have to suit PDT 2.0. We list the features from the previous section and also add some obvious requirements that have not been mentioned so far but are very useful generally, regardless of a corpus.

2.2.1. Complex Evaluation of a Node

- multiple attributes evaluation (an ability to set values of several attributes at one node)
- alternative values (e.g. to define that `functor` of a node is either a disjunction or a conjunction)
- alternative nodes (alternative evaluation of the whole set of attributes of a node)
- wild cards (regular expressions) in values of attributes (e.g. `m/tag=''N...4.*''` defines that the morphological tag of a node is a noun in the accusative, regardless of other morphological categories)
- negation (e.g. to express "this node is not an Actor")
- relations lower than ("<") , higher than (">") (for numerical attributes)

2.2.2. Dependencies Between Nodes (Vertical Relations)

- immediate, transitive dependency (existence, non-existence)
- vertical distance (from root, from one another)
- number of sons (zero for leaves)

2.2.3. Horizontal Relations

- precedence, immediate precedence (positive, negative)
- horizontal distance
- secondary edges, secondary dependencies, coreferences, long-range relations

2.2.4. Other Features

- multi-tree queries (combined with general `OR` relation)
- skipping a node of a given type (for skipping simple types of coordination, apposition etc.)
- skipping multiple nodes of a given type (e.g. for recognizing the rightmost path)
- references (for matching values of attributes unknown at the time of creating the query)
- accessing several layers of annotation at the same time with a non-1:1 relation (for studying relations between layers)
- searching in the surface form of the sentence

3. The Query Language

In this contribution, we only shortly and selectively introduce a query language that satisfies linguistic requirements stated in the previous section. We present the language informally on a series of examples. A full description of the query language, as well as a formal definition of the textual form of the query language, can be found in Mírovský (2008e). The query language is an extension to the existing query language of Netgraph 1.0 (Ondruška 1998).

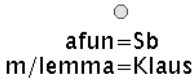
A query in Netgraph is always a tree (or a multi-tree, see below) that forms a subtree in the result trees. The treebank is searched tree by tree and whenever the query is found as a subtree of a tree, the tree becomes a part of the result.

3.1. The Basics

The simplest possible query is a simple node without any evaluation:

This query matches all nodes of all trees in the treebank, each tree as many times as how many nodes there are in the tree.

Values of attributes of the node can be specified in the form of `attribute=value` pairs:



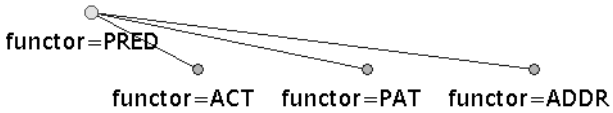
The query searches for all trees containing a node evaluated as Subject ("`Sb`") with lemma Klaus.

3.2. Regular Expressions

A Perl-like regular expression (Hazel 2007) can be used as a whole value of an attribute. If the value of an attribute is enclosed in quotation marks, the value is considered an anchored regular expression.

3.3. Dependencies Between Nodes

Dependencies between nodes are expressed directly in the syntax of the query language. Since the result is always a tree, the query also is a tree (or a multi-tree, see Section below) and the syntax does not allow non-tree constructions. The following query searches for Predicates ("`PRED`") that directly govern an Actor ("`ACT`"), a Patient ("`PAT`") and an Addressee ("`ADDR`"):



It is important to note that the query does not prevent other nodes in the result being sons of the Predicate and that the order of the sons as they appear in the query can differ from their order in the result trees.

3.4. Meta-Attributes

Meta-attributes are attributes that are not present in the corpus, yet they pretend to be ordinary attributes and users can treat them the same way as normal attributes. There are eleven

meta-attributes, each adding some power to the query language, enhancing its semantics, while keeping the syntax of the language on the same simple level. To be easily recognized, names of the meta-attributes start with an underscore ("_").

3.4.1. _transitive

This meta-attribute defines a transitive edge. It has two possible values: the value `true` means that a node may appear anywhere in the subtree of a node matching its query-father, the value `exclusive` means, in addition, that the transitive edge cannot share nodes in the result tree with other exclusively transitive edges.

3.4.2. _optional

The meta-attribute `_optional` defines an optional node. It may but does not have to be in the result tree at a given position. Its father and its son (in the query) can be the direct father and son in the result. Only the specified node can appear (once or more times as a chain) between them in the result tree. Possible values are:

- `true` – There may be a chain of unlimited length (even zero) of nodes matching the optional node in the result tree between nodes matching the query-father and the query-son of the optional node.
- *a positive integer* – There may be a chain of length from zero up to the given number of nodes matching the optional node in the result tree between nodes matching the query-parent and the query-son of the optional node.

3.4.3. _#sons

The meta-attribute `_#sons` ("number of sons") controls the exact number of sons of a node in the result tree.

3.4.4. _#hsons

The meta-attribute `_#hsons` ("number of hidden sons") is similar to the meta-attribute `_#sons`. It controls the exact number of hidden sons of a node in the result tree. Hidden sons are used to access lower layers of annotation from the tectogrammatical layer, see Section .

3.4.5. _depth

The meta attribute `_depth` controls the distance of a node in the result tree from the root of the result tree.

3.4.6. _#descendants

The meta-attribute `_#descendants` ("number of descendants") controls the exact number of all descendants of a node (number of nodes in its subtree), excluding the node itself.

3.4.7. _#lbrothers

The meta-attribute `_#lbrothers` ("number of left brothers") controls the exact number of left brothers of a node in the result tree.

3.4.8. _#rbrothers

Similarly, the meta-attribute `_#rbrothers` ("number of right brothers") controls the exact number of right brothers of a node in the result tree.

3.4.9. _#occurrences

The meta-attribute `_#occurrences` ("number of occurrences") specifies the exact number of occurrences of a particular node at a particular place in the result tree. It controls how many nodes of the kind can occur in the result tree as sons of the father of the node (including the node itself). Zero value can be used to express that a particular type of node is not among sons of a node.

3.4.10. _name

The meta-attribute `_name` is used to name a node for a later reference, see Section "???-??¿' below.

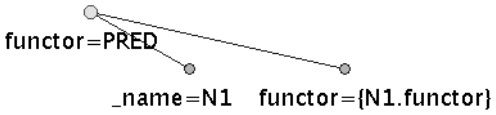
3.4.11. _sentence

The meta-attribute `_sentence` can be used to search in the linear surface form of the trees – in the sentences.

3.5. References

References are used in the queries to refer to values of attributes in the result trees, to values unknown at the time of creating the query. We use the word "references" as a technical term that simply means "a pointer". It should not be confused with linguistically motivated terms like "coreference".

First, a node in the query has to be named using the meta-attribute `_name`. Then, references to values of attributes of this node can be used at other nodes of the query. The following query searches for a Predicate with two sons with the same functor in the result tree, whatever the functor may be:



References can refer to the whole value (as shown above) or only to one character of the value. The required position is separated from the name of the attribute with another dot ("`.`"). It is also possible that references only form a substring of a defined value and appear several times in a definition of an attribute (but they cannot be a part of a regular expression).

3.6. Multi-Tree Queries

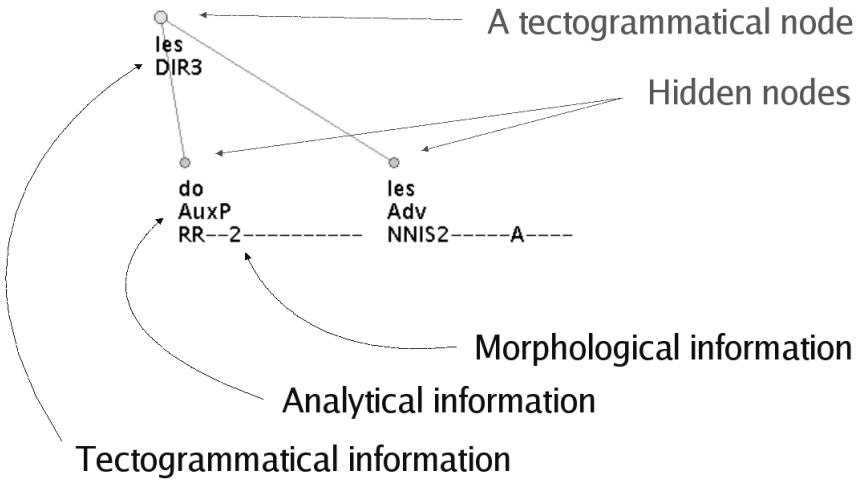
A multi-tree query consists of several trees combined either with a general `AND` or a general `OR`. In the case of `AND`, all the query trees are required to be found in the result tree at the same time (different nodes in the query cannot be matched with one node in the result), while in the case of `OR`, at least one of the query trees is required to be found in the result tree.

3.7. Hidden Nodes

Hidden nodes are nodes that are marked as hidden by setting the attribute `hide` to `true`. Their visibility in result trees can be switched on and off. Hidden nodes can serve as a connection to the lower layers of annotation with non-1:1 relations or generally to any external source of information.

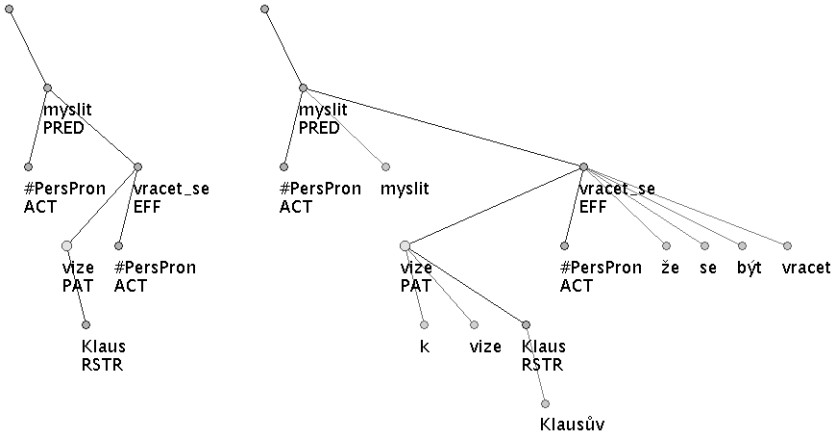
Netgraph uses the hidden nodes as a connection to the lower layers of annotation with non-1:1 relations. It presents all the available information in one tree (Mírovský 2006). The tectogrammatical nodes contain only the tectogrammatical information, while all the information from the lower layers is kept at the hidden nodes. Each tectogrammatical node has as many hidden sons as there are analytical nodes corresponding to the tectogrammatical node. (Hidden nodes were first introduced with the Prague Dependency Treebank 1.0; they were used in a slightly different way there.)

The principle of using hidden nodes for representing information from several layers of annotation in one tree is demonstrated in the following picture, which shows how the phrase "do lesa" ("to the forest") is annotated on several layers of annotation and how it is represented using the hidden nodes:



One node on the tectogrammatical layer with `t_lemma=les` ("the forest") and `functor=DIR3` (representing the direction "to") has two hidden sons, representing a preposition `do` ("to") and an adverbial `les` ("the forest"). The information from the morphological layer is merged into the analytical layer.

The hidden nodes are usually not displayed – they are "hidden". The following picture demonstrates two possible ways of displaying a tectogrammatical tree in Netgraph. On the left side, there is a tectogrammatical tree with the hidden nodes hidden. In the same tree on the right side, the hidden nodes are displayed:



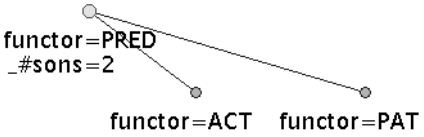
In Czech: *Myslím, že ke Klausově vizi se budeme vracet.*
In English: *I think that to Klaus's vision we will get back.*

4. Using the Query Language

Let us show a few representative examples of searching for some of the linguistic phenomena annotated in PDT 2.0. It was shown in much more detail in Mírovský 2008e that Netgraph Query Language fulfils the requirements stated in Section : The query language meets the general requirements on a query language for PDT 2.0, listed in Section ; it can be used for searching for all linguistic phenomena from PDT 2.0 listed in Section (Mírovský 2008c).

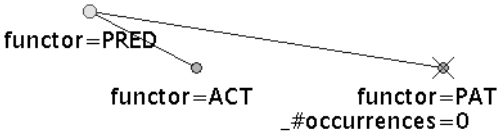
Valency

We present two queries for studying valency. The first query searches for Predicates governing an Actor, a Patient and nothing else (the Actor and the Patient are members of the valency frame, no other member is present):



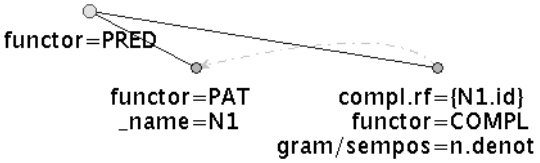
The meta-attribute `_#sons` makes sure that there are no other sons of the Predicate in the result trees.

The second query searches for Predicates governing an Actor and not governing a Patient. Since Patient has to be the second inner participant of any valency frame that has at least two inner participants (t-manual, page 102), the query searches for occurrences of Predicates with only one inner participant in its valency frame – the Actor:



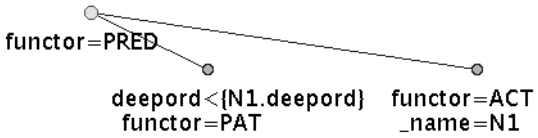
Predicative Complement (Dual Relation)

The example query uses references, the referential information is stored in the attribute `compl.rf`. The query searches for those cases of the predicative complement where the second relation goes to a Patient:



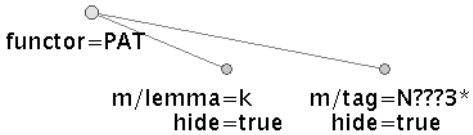
Topic-Focus Articulation

The communicative dynamism requires that the relative order of nodes in the tree from left to right can be expressed. The order of nodes is controlled by the attribute `deepord`, which contains a non-negative real (usually natural) number that sets the order of nodes from left to right. The following query demonstrates searching for a Predicate governing an Actor and a Patient, the Patient less dynamic (on the left side in the tree) than the Actor:



Accessing Lower Layers

Let us present an example query that accesses the lower layers of annotation from the tectogrammatical layer. It searches for Patients (on the tectogrammatical layer) that are expressed with a preposition "k" and a noun in the dative on the morphological layer:



The Patient has (at least) two hidden sons, the former with lemma "k", the latter with a morphological tag that states that the node is a noun in the dative.

5. Comparison to Other Treebank Query Systems

To show the power of FS Query Language, we use an indirect approach of comparing the language to four other query languages, languages of TGrep (Pito 1994), TGrep2 (Rohde 2005), TigerSearch (Brants et al. 2002), and fsq (Kepser 2003). We present a table showing to what extent the five tools (Netgraph and the other four tools) fulfil the requirements stated in Section 2.2. Please note that the table is biased in favour of Netgraph, because Netgraph has been designed to fulfil the requirements. The table does not contain query language features that do not belong to the requirements. The other tools have been designed for different corpora and may implement features that Netgraph does not. A detailed unbiased comparison of the expressive power of Netgraph Query Language and the query languages of TGrep, TGrep2 and TigerSearch is presented in Mírovský (2008a) and Mírovský (2008e).

In the table, the following marks are used:

+ ... the feature is supported

- ... the feature is not supported

* ... the feature is partially supported

N/A ... the feature is not applicable to the query language

Complex Evaluation of a Node	**TGrep**	**TGrep2**	**TigerSearch**	**fsq**	**Netgraph**
multiple attributes evaluation (the ability to set values of several attributes at one node)	−	−	+	+	+
alternative values (e.g. to define that functor of a node is either a disjunction or a conjunction)	+	+	+	+	+[I]
alternative nodes (alternative evaluation of the whole set of attributes of a node)	N/A	+	+	+	+
wild cards (regular expressions) in values of attributes	+	+	+	+	+
negation (e.g. to express "this node is not an Actor")	+	+	+	+	+
relations lower than (<) , higher than (>)	−	−	−	−	+
Dependencies Between Nodes (Vertical Relations)	**TGrep**	**TGrep2**	**TigerSearch**	**fsq**	**Netgraph**
immediate, transitive dependency (existence, non-existence)	+	+	*[II]	+	+
vertical distance (from root, from one another)	−	−	*[III]	*[III]	+
number of sons (zero for leaves)	+	+	+	+	+
Horizontal Relations	**TGrep**	**TGrep2**	**TigerSearch**	**fsq**	**Netgraph**
precedence, immediate precedence (positive, negative)	+	+	*[IV]	+	+
horizontal distance	−	−	*[V]	*[V]	+
secondary edges, secondary dependencies, coreferences, long-range relations	*[VI]	*[VI]	+	+	+
Other Features	**TGrep**	**TGrep2**	**TigerSearch**	**fsq**	**Netgraph**
multi-tree queries (combined with the general OR relation)	−	+[VII]	+[VIII]	+[IX]	+[X]
skipping a node of a given type (for skipping simple types of coordination, apposition etc.)	−	+[XI]	+[XII]	+	+
skipping multiple nodes of a given type (e.g. for recognizing the rightmost path)	_[XIII]	_[XIII]	_[XIV]	+	+
references (for matching values of attributes unknown at the time of creating the query)	−	+	+	−	+
accessing several layers of annotation for studying relations between layers with non-1:1 relations	N/A	N/A	N/A	N/A	+
searching in the surface form of the sentence	+[XV]	+[XV]	+[XV]	+	+

Notes referred to from the table:

I: Only OR relation is supported.

II: Variables (nodes in the query) are existentially quantified. If the query specifies that A does not dominate B, then B must appear somewhere else in the tree.

III: Vertical distance can only be measured for nodes that are in the transitive dependency relation.

IV: Variables (nodes in the query) are existentially quantified. If the query specifies that A does not precede B, then B must appear somewhere else in the tree.

V: Horizontal distance can be measured for leaf nodes.

VI: Only one type of dependency can be set (although multiple times at a node, expressing relations to several nodes).

VII: Full Boolean expressions on patterns are supported.

VIII: Boolean expressions without negation on patterns are supported.
IX: At least first-order logic formula can be used.
X: Only the general OR or general AND are supported.
XI: Thanks to general Boolean expressions on patterns.
XII: Thanks to Boolean expressions on patterns.
XIII: But there are special predicates for the rightmost/leftmost descendant of a node.
XIV: But there are special predicates for the rightmost/leftmost leaf descendant of a node.
XV: Using predicates for precedence and immediate precedence on terminals.

6. The Tool

We have implemented Netgraph Query Language in a search tool called Netgraph. As a basis, we used Netgraph 1.0. We present a list of features that we consider important for a search tool for a treebank, especially for the Prague Dependency Treebank 2.0. We do not include general features that can be expected from any graphically oriented tool, like saving or printing capability. We rather focus on features that are connected with searching in treebanks. All these features have been implemented in Netgraph, so we present them this way. Some of the features have been implemented on a request from users:

- **client-server architecture**

With the client-server architecture, data can reside at one place in the Internet. Multiple users (clients) can access the server simultaneously (Mírovský, Ondruška 2002a, Mírovský et al. 2002b). The version control has been implemented in the tool, in order to keep the server and the client compatible.

- **authentication of users**

In order to protect the data, the authentication of users is available. Each user gets a login name and a password to access the server. Different users can have different permissions (maximum number of found trees, a permission to change the password, a permission to save the result trees to the local disc).

- **graphical creation of the query**

Especially for non-programmers, a graphical creation of the query, in our case a full implementation of Netgraph Query Language, is important.

- **browsing the result trees**

Obviously, users have to be able to browse the result of a query. A graphical representation of the trees is again an important feature. It includes displaying coreferential arrows and other references, as well as hidden nodes on request.

- **access to context trees**

Since the annotation on the tectogrammatical layer captures the linguistic meaning of the sentence in its context, the context of the sentence has to be accessible as well. The tool allows displaying context trees in both directions (forward and backward).

- **chained queries**

To refine a result of a query, another query can be set on top of the previous query. The second query searches only in the result of the previous query. This way, queries can be chained unlimitedly.

- **inverted search**

Some queries can be much simpler if the inversion of matching is available. We can simply define a query that represents a phenomenon that we do not want in the result trees and invert the search. Only trees that do not match the query become a part of the result.

- **search only for the first occurrence in each tree**

If we are only interested in the result trees and not in multiple occurrences of a query in the result trees, a possibility to search only for the first occurrence in the result trees can be useful. Although the tool allows the user to browse the result trees in such a way that multiple occurrences of a query in one tree are skipped, they are still searched for (thus the search slows down); searching only for the first occurrence makes the search faster. It is also very useful for chained queries if the subsequent query does not search in several same trees representing multiple occurrences of the previous query in one tree.

- **removing trees from the result**

Sometimes, it is difficult to refine a query further to obtain the exact set of result trees a user wishes. Therefore, a possibility to remove an unwanted result tree from the result is available (e.g. before the result is saved to the local disc).

- **right-left trees**

Some languages, like Arabic, require right-left ordering of nodes in the trees, as well as of the tokens in the sentence. The tool has to offer this feature.

- **multi-language support**

UTF-8 has become a standard in coding characters of natural languages. Thanks to this universal coding, all major languages are supported in Netgraph, even at the same time (in one corpus).

- **basic statistics**

The tool has to provide at least the most basic statistics about the result. It provides the following numbers: number of searched trees, number of found (result) trees, number of found occurrences in the found trees, and also number of the actually displayed tree/occurrence.

- **external command**

For further processing of the found tree, an external command can be run from the tool. Several variables for identifying the file, the tree and the position in the tree are substituted before the external command is launched.

- **speed/portability**

For the server, speed is the most important factor. Therefore, C programming language (Herout 2002) has been chosen for the implementation.

On the other hand, the most important factor for choosing the programming language for the client is portability. Java 2 (Eckel 2006) belongs to the best portable programming languages

and it has also a very good support for various natural languages; it uses its own fonts and supports UTF-8 very naturally. Therefore, Java 2 has been chosen as a programming language for the client.

6.1. Changes since Version 1.0

The actual version of Netgraph is 1.93. We call the original version of Netgraph, programmed by Roman Ondruška, Netgraph 1.0. Here, we describe the main changes that have been done to the tool since this 1.0 version.

Let us start with several numbers representing code lines. The Netgraph client 1.0 had 1 526 lines of code. The Netgraph client 1.93 consists of more than 21 thousand lines. The Netgraph server 1.0 had 3 973 lines of code. The Netgraph server 1.93 has more than 11 thousand lines.

The following lists contain the most important changes that have been done since the version 1.0. The first list describes extensions to the query language, the second list describes changes in the tool.

6.1.1. Main Extensions to the Query Language

- Meta-attributes have been introduced to the system.
- References to values of attributes of (other) nodes can be set in the query.
- Regular expressions in values of attributes can be used.
- Other relations than equation can be used for setting values of attributes.
- Arithmetic operations in numerical values of attributes can be used.
- Multi-tree queries are supported.
- Support for hidden nodes has been added.

6.1.2. Main Extensions to the Tool

- The tool now supports the tectogrammatical trees (with hidden nodes and coreferences), both in searching and displaying; a configuration file defining how to display individual references is available.
- Authentication of users has been implemented.
- Queries can be chained.
- The matching of a query can be inverted.
- History of queries is created; queries or the whole history can be saved to the local disc; a list of selected files for searching can also be saved.
- Result trees can be printed or saved to the local disc.
- The tool now supports the UTF-8 encoding.
- Right-left trees are supported.
- Version control has been implemented.
- A query is created in a fully graphical way.
- Basic statistics about the search are provided.

- Context trees can be displayed.
- Individual trees can be removed from the result.
- An external command with variables substitution can be launched from the tool.

7. Real World

7.1. Netgraph Query Language and PDT 2.0

After we have presented Netgraph Query Language and shown what can be searched for with the language, it might be interesting to know to what extent the features have been put to use by the users and what the users really do search for. There are about 40 registered users and an anonymous access to the server for PDT 2.0 is also available.

Since October 2002, the Netgraph server stores all queries to a log file. By then, only the analytical trees were searched through in Netgraph. Since February 2005, also the tectogrammatical trees (though not publicly released yet) have been made available in Netgraph for the internal usage of our institute, and later (after PDT 2.0 publication) the tectogrammatical trees were made available for the registered public users, too.

From these two servers (the analytical and the tectogrammatical trees), all queries entered by users have been stored in log files. However, we have not had access to queries that had been processed on local installations of the Netgraph server, e.g. on notebooks, which are quite numerous. All the following numbers come only from the two public servers mentioned above (from the dates stated above up to March 24, 2008). For obvious reasons, before any statistics were counted, we excluded all queries that we had entered.

Number of:	Total	Analytical Trees	Tectogrammatical Trees
all queries	16 870	10 299	6 571
one-node queries	10 146	7 180	2 966
structured queries (more than one node)	6 724	3 119	3 605
queries without a meta-attribute	15 575	9 989	5 586
queries with a meta-attribute	1 295	310	985
`_transitive`	174	81	93
`_optional`	172	18	154
`_#sons`	91	22	69
`_#hsons`	36	–	36
`_depth`	51	11	40
`_#descendants`	103	24	79
`_#lbrothers`	35	25	10
`_#rbrothers`	11	0	11
`_#occurrences`	197	12	185
`_name`	397	116	281
`_sentence`	28	1	27
queries with a reference	363	110	253
queries with a hidden node	1 194	–	1 194
queries with an alternative value	884	314	570
queries with an alternative node	94	19	75

The table shows numbers of queries using various features of the query language, both on the analytical layer and on the tectogrammatical layer. The total usage is also counted.

Some values in the table should be equal but they are not. The number of queries that use the meta-attribute `_name` should be equal to the number of queries that use a reference. The discrepancy is caused by errors in some queries (e.g. queries that contain a named node but the name is never used).

A representative selection of queries put in by the users can be found in Mírovský (2008e).

7.2. Other Usages of Netgraph

The query tool Netgraph and its query language are general enough to be used with other treebanks than PDT 2.0. Netgraph can be used both for dependency trees and for constituent structure trees, provided the treebank is transformed to FS File Format (Mírovský 2008e), and also other kinds of usage are possible.

7.2.1. Czech Academic Corpus 1.0 and 2.0

During the work on the re-annotation of the Czech academic corpus (Králík and Hladká 2006), Netgraph was used for searching for errors in the process of re-annotation of the data from the original annotation scheme to a PDT-like annotation scheme. The first version of the

"new" Czech academic corpus contained only the morphological annotation (Vidová-Hladká et al. 2007). During its preparation, the data was searched for errors on the morphological layer. Since there is no structure in the morphological annotation (but Netgraph only works with trees), flat morphological "trees" were used, in which a technical root had all the words of the sentence as its sons.

During the preparation of the second version of the Czech Academic Corpus (version 2.0), which is going to be released in LDC in the Fall of 2008, the morphologically annotated files were first automatically parsed on the analytical layer (Ribarov et al. 2006). Netgraph was then used for searching for errors on the analytical layer. The annotation was almost identical to the analytical layer of PDT 2.0, therefore a similar set of checks as for PDT 2.0 (Štěpánek 2006) was used.

7.2.2. Latin IT Treebank

Index Thomisticus (IT) Treebank is an ongoing project, which is a part of the Lessico Tomistico Biculturale (LTB) project by Father Roberto Busa.[2] IT-Treebank wants to make IT a Treebank.

The annotation on the analytical layer is performed on the basis of the annotation guidelines for the Prague Dependency Treebank and according to guidelines specifically written for Latin, shared and developed with the Latin Dependency Treebank of the Perseus Project in Boston. Presently, IT-Treebank is composed of 32 880 tokens, for a total of 1 479 syntactically parsed sentences from the Scriptum super Sententiis Magistri Petri Lombardi.

During the development of the Latin treebank, Netgraph is used for browsing the data and searching in the data.

7.2.3. Arabic Trees

In the year 2003, Netgraph was installed in LDC (Linguistic Data Consortium) in Philadelphia, University of Pennsylvania[3], to be used with their Arabic treebank. In cooperation with LDC, the Prague Arabic Dependency Treebank (Smrž et al. 2005) was developed at ÚFAL (Institute of Formal and Applied Linguistics) at Charles University in Prague[4]. Netgraph was used during the annotation work for studying the treebanks. Right-left ordering of nodes in trees was implemented for purposes of the Arabic treebanks.

7.2.4. Chinese Treebank

Netgraph has also been used for a work on a Chinese treebank at ÚFAL. Since Java supports Chinese language and Netgraph works with files encoded in UTF-8, no adaptation of the tool

[2] `http://gircse.marginalia.it/ passarotti/`. IT is considered as the pathfinder of Computer Sciences applications in the Humanities; it retains the opera omnia by Thomas Aquinas (118 texts), plus works by other 61 authors related to Thomas (61 texts). It is a corpus of around 11 millions of tokens (150.000 types; 20.000 lemmas).

[3] LDC – `http://www.ldc.upenn.edu/`

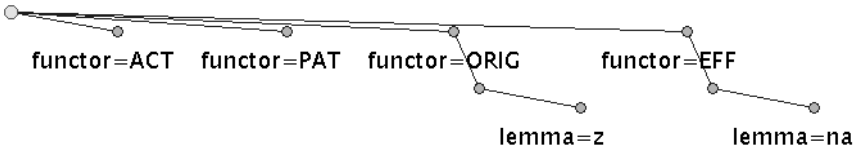
[4] ÚFAL – `http://ufal.mff.cuni.cz`

was necessary. This is an example of the use of Netgraph with constituent-structure trees.

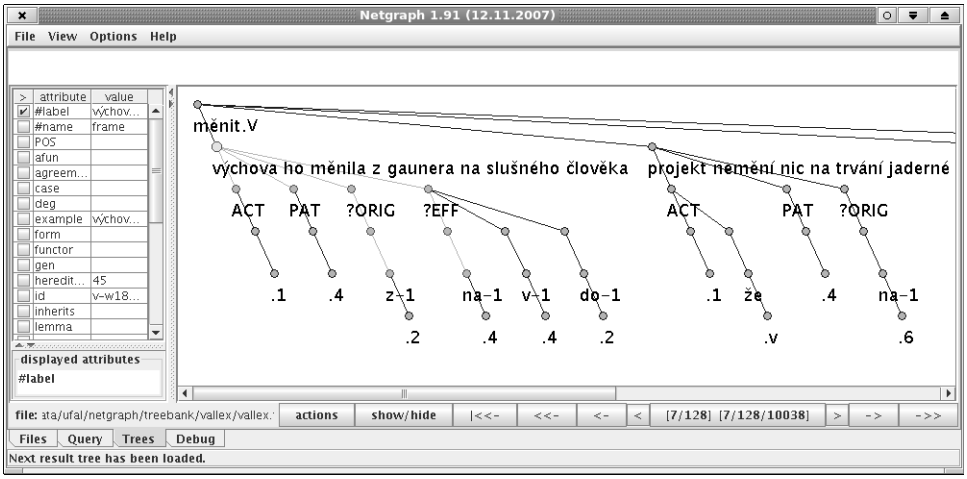
7.2.5. Vallex

Vallex is a valency lexicon of Czech (Žabokrtský and Lopatková 2007). A recent usage of Netgraph for sophisticated searching in this "treebank" belongs to interesting applications of the tool. Thanks to Petr Pajas and his tool TrEd (Pajas 2007), Vallex has been transformed to FS File Format and can be searched through with Netgraph.

The following query searches for valency frames of the type "přešila panenku z kašpárka na čerta" ("she altered the puppet from the Punch to the devil"), i.e. valency frames consisting of an Actor, a Patient, an Origin and an Effect. The query also requires that on the surface, the Origin is expressed with the preposition "z" and the Effect is expressed with the preposition "na":



The following picture shows one of the results in Netgraph:



In Czech: *výchova ho měnila z gaunera na slušného člověka*
In English: *education was changing him from a scrounger to a decent man*

8. Conclusion

In the paper, we have studied the Prague Dependency Treebank 2.0 and created a list of linguistic phenomena annotated in the treebank that bring a requirement on a query language for searching in the treebank. We have assembled a list of requirements that any query language should satisfy in order to fit the Prague Dependency Treebank 2.0.

We have proposed Netgraph Query Language – a simple to use and graphically oriented language that meets the requirements.

The proposed query language is an extension to an existing query language – a query language of Netgraph 1.0. The following three features are the most important additions to the query language:

- *meta-attributes* – for setting complex types of relation between nodes and complex properties of the nodes
- *hidden nodes* – for accessing lower layers of annotation with non-1:1 relation among nodes
- *references* – for setting relations between values of attributes of nodes that are unknown at the time of creating the query

The proposed query language meets the requirements on a query language for the Prague Dependency Treebank 2.0 (Mírovský 2008e).

We have compared the proposed query language to some other query languages.

We have also studied to what extent the features of the query language have been put to use by real users

The proposed query language has been implemented in Netgraph, which is also an extension to the existing search tool – Netgraph 1.0. Thus, a comfortable, simple to use and fully graphically oriented client-server system for searching in the Prague Dependency Treebank 2.0 has been created.

Acknowledgement

The research and work presented in the paper were supported by the Grant Agency of the Academy of Sciences of the Czech Republic, project IS-REST (No. 1ET101120413).

9. References

Brants S. et al. (2002): The TIGER Treebank. *In: Proceedings of TLT 2002, Sozopol, Bulgaria, 2002.*

Eckel B. (2006): Thinking in Java (4th edition). *Prentice Hall PTR, 2006.*

Hana J., Zeman D., Hajič J., Hanová H., Hladká B., Jeřábek E. (2005): Manual for Morphological Annotation, Revision for PDT 2.0. *ÚFAL Technical Report TR-2005-27, Charles University in Prague, 2005.*

Hajič J., Vidová-Hladká B., Panevová J., Hajičová E., Sgall P., Pajas P. (2001): Prague Dependency Treebank 1.0 (Final Production Label). *CD-ROM LDC2001T10, LDC, Philadelphia, 2001.*

Hajič J. et al. (1997): A Manual for Analytic Layer Tagging of the Prague Dependency Treebank. *ÚFAL Technical Report TR-1997-03, Charles University in Prague, 1997.*

Hajič J. et al. (2006): Prague Dependency Treebank 2.0. *CD-ROM LDC2006T01, LDC, Philadelphia, 2006.*

Hajičová E. (2007): Information Structure from the Point of View of the Relation of Function and Form. *In: The Prague Bulletin of Mathematical Linguistics 88, 2007, pp. 53-71.*

Hajičová E. (1998): Prague Dependency Treebank: From analytic to tectogrammatical annotations. *In: Proceedings of 2nd TST, Brno, Springer-Verlag Berlin Heidelberg New York, 1998, pp. 45-50.*

Hajičová E, Panevová J. (1984): Valency (case) frames. *In P. Sgall (ed.): Contributions to Functional Syntax, Semantics and Language Comprehension, Prague, Academia, 1984, pp. 147-188.*

Hajičová E., Partee B., Sgall P. (1998): Topic-Focus Articulation, Tripartite Structures and Semantic Content. *Dordrecht, Amsterdam, Kluwer Academic Publishers, 1998.*

Havelka J. (2007): Beyond Projectivity: Multilingual Evaluation of Constraints and Measures on Non-Projective Structures. *In: Proceedings of ACL 2007, Prague, pp. 608-615.*

Hazel P. (2007): PCRE (Perl Compatible Regular Expressions) Manual Page. *Available from http://www.pcre.org/*

Herout P. (2002): Učebnice jazyka C. *Kopp 2002.*

Kepser S. (2003): Finite Structure Query – A Tool for Querying Syntactically Annotated Corpora. *In Proceedings of EACL 2003, pp. 179-186.*

Králík J., Hladká B. (2006): Proměna Českého akademického korpusu (The transformation of the Czech Academic Corpus). *In: Slovo a slovesnost 3/2006, pp. 179-194.*

Kučová L., Kolářová-Řezníčková V., Žabokrtský Z., Pajas P., Čulo O. (2003): Anotování koreference v Pražském závislostním korpusu. *ÚFAL Technical Report TR-2003-19, Charles University in Prague, 2003.*

Mikulová M., Bémová A., Hajič J., Hajičová E., Havelka J., Kolářová V., Kučová L., Lopatková M., Pajas P., Panevová J., Razímová M., Sgall P., Štěpánek J., Urešová Z., Veselá K., Žabokrtský Z. (2006): Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. *Tech. Report 30, ÚFAL MFF UK, 2006.*

Mírovský J. (2008e): Netgraph – a Tool for Searching in the Prague Dependency Treebank 2.0. *PhD Thesis, Charles University in Prague, 2008.*

Mírovský J. (2008d): PDT 2.0 Requirements on a Query Language. *In: Proceedings of ACL 2008, Columbus, Ohio, USA, 16th - 18th June 2008, pp. 37-45.*

Mírovský J. (2008c): Does Netgraph Fit Prague Dependency Treebank? *In: Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008), Marrakech, Marocco, 28th - 30th May 2008.*

Mírovský J. (2008a): Towards a Simple and Full-Featured Treebank Query Language. *In: Proceedings of ICGL 2008, Hong Kong, 9th - 11th January 2008, pp. 171-178.*

Mírovský J. (2006): Netgraph: a Tool for Searching in Prague Dependency Treebank 2.0. *In Proceedings of TLT 2006, Prague, pp. 211-222.*

Mírovský J., Ondruška R., Průša D. (2002b): Searching through Prague Dependency Treebank

- Conception and Architecture. *In Proceedings of The First Workshop on Treebanks and Linguistic Theories, Sozopol, 2002, pp. 114—122.*

Mírovský J., Ondruška R. (2002a): NetGraph System: Searching through the Prague Dependency Treebank. *In: The Prague Bulletin of Mathematical Linguistics 77, 2002, pp. 101-104.*

Ondruška R. (1998): Tools for Searching in Syntactically Annotated Corpora. *Master Thesis, Charles University in Prague, 1998.*

Pajas P. (2007): TrEd User's Manual. Available from http://ufal.mff.cuni.cz/ pajas/tred/

Pito R. (1994): TGrep Manual Page. *Available from http://www.ldc.upenn.edu/ldc/online/treebank/*

Ribarov et al. (2006): When a Statistically Oriented Parser Was More Efficient Than a Linguist: A Case of Treebank Conversion. *In: The Prague Bulletin of Mathematical Linguistics 86, 2006, pp. 21-38.*

Rohde D. (2005): TGrep2 User Manual. *Available from http://www-cgi.cs.cmu.edu/ dr/TGrep2/ tgrep2.pdf*

Sgall P. (2001): Underlying Structures in Annotating Czech National Corpus. *In: Current Issues in Formal Slavic Linguistics, edited by G. Zybatow, U. Junghanns, G. Mehlhorn and L. Szucsich, Peter Lang, Frankfurt a/Main, 2001, pp. 499-505.*

Smrž O., Pajas P., Žabokrtský Z., Hajič J., Mírovský J., Němec P. (2005): Learning to Use the Prague Arabic Dependency Treebank. *In: Elabbas Benmamoun. Proceedings of Annual Symposium on Arabic Linguistics (ALS-19). Urbana, IL, USA, Apr. 1-3: John Benjamins, 2005.*

Štěpánek J. (2006): Post-Annotation Checking of Prague Dependency Treebank 2.0 Data. *In: The Prague Bulletin of Mathematical Linguistics 85, 2006, pp. 23-33.*

Vidová-Hladká B., Hajič J., Hana J., Hlaváčová J., Mírovský J., Votrubec J. (2007): Czech Academic Corpus 1.0 Guide. *Karolinum - Charles University Press, 2007, ISBN: 978-80-246-1315-4*

Žabokrtský Z., Lopatková M. (2007): Valency Information in VALLEX 2.0. *In: The Prague Bulletin of Mathematical Linguistics 88, 2007, pp. 41-59.*

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 90 DECEMBER 2008 33–56

Acquisition du tchèque par les francophones: Analyse automatique des erreurs de déclinaison

Ivan Šmilauer

Abstract

This paper is a summary of our PhD thesis (Šmilauer, 2008) that presents the concept and the implementation of a platform of computer-assisted language learning, featuring on-line fill-in-the-blank exercises with feedback on errors in Czech declension (www.cetlef.fr). Morphological annotation of required forms enables a didactic presentation of the morphological system on the learning platform, as well as the implementation of a procedure of automatic error diagnosis that is carried out by the comparison of an erroneous production with hypothetical forms generated from the stem of the required form. The device can be used as a source of data for a research into second langauge acquisition.

1. Introduction

La langue tchèque, avec sa flexion très riche, offre un matériau intéressant du point de vue de l'acquisition de la morphologie par des étudiants étrangers.

Dans notre thèse, nous nous sommes concentrés sur les erreurs commises par les apprenants francophones dans un cadre expérimental restreint : celui d'exercices de déclinaison dans lesquels il faut décliner le lemme d'une forme (substantif, adjectif, pronom, numéral) au sein d'une phrase.

Une erreur de déclinaison se manifeste par un écart entre une ou plusieurs formes requises, dans un certain contexte syntaxique, et la forme erronée produite par l'apprenant. Nous avons établi l'hypothèse qu'une telle forme peut être générée automatiquement à partir du lemme de la forme requise, à l'aide des moyens formels de la déclinaison (choix d'une désinence, réalisation des alternances), employés d'une manière incorrecte. En nous basant sur cette hypothèse, nous avons proposé un module de diagnostic automatique des erreurs dont l'objectif est de générer un message de retour spécifiant le type de l'erreur au niveau morphologique.

Ce diagnostic, possible grâce à l'annotation morphologique des formes requises, a été implémenté sur une plateforme d'enseignement de langue assisté par ordinateur qui représente la

Please cite this article as: Ivan Šmilauer, Acquisition du tchèque par les francophones: Analyse automatique des erreurs de déclinaison. The Prague Bulletin of Mathematical Linguistics No. 90, 2008, 33–56.

partie appliquée de notre thèse. Cette plateforme, nommée CETLEF[1], est une application Web dynamique (disponible librement sur www.cetlef.fr) contenant une base de données relationnelle gérée par MySQL et une interface XHTML avec des éléments dynamiques en Javascript. Les procédures automatiques sont implémentées en langage PHP. CETLEF contient une plateforme auteur qui sert pour la création des exercices, et une plateforme apprenant qui est destinée aux étudiants.Pendant l'inscription sur cette plateforme, les apprenants fournissent des informations qui peuvent aider pendant l'interprétation de leur productions (âge, durée de l'apprentissage du tchèque, autres langues maîtrisées, etc.).

2. Motivation de CETLEF

Bien que l'enseignement pratique du tchèque langue étrangère (TLE) soit d'une tradition relativement riche, voir (Hrdlička, 2002), ce n'est qu'à partir des années 1980 que commencent à apparaître des travaux préliminaires, incitant à la constitution d'un champ de recherche autonome dont l'objet serait une méthodologie spécifique pour l'enseignement du TLE. La présentation didactique de la déclinaison est un des problèmes principaux dans l'enseignement, voir par exemple (Poldauf and Špruňk, 1968) ou (Nekula, 2007).

2.1. CETLEF comme source de données pour l'analyse des erreurs

En adoptant l'analyse des erreurs comme moyen privilégié pour étudier l'acquisition d'une langue étrangère, voir par exemple (Porquier, 1977), (Besse and Porquier, 1991), (Gaonac'h, 1991), le premier de nos objectifs a été de concevoir un outil qui puisse servir comme source de données pour l'étude des erreurs dans la déclinaison.

2.1.1. Productions libres

L'avantage des *productions libres* est l'authenticité des données qui reflètent l'emploi effectif de la langue. L'inconvénient principal est une collecte de données coûteuse en temps et effort. Les productions libres sont également affectées par la volonté d'utiliser des structures et un vocabulaire que l'apprenant estime maîtriser suffisamment bien pour pouvoir s'en servir dans la communication. On parle dans ce cas des «stratégies d'évitement», voir par exemple (Porquier, 1977), (Bautier-Castaing, 1977).

Afin de disposer de plus de données pour l'analyse, les collectes de corpus électroniques de productions d'apprenants commencent à émerger depuis une quinzaine d'années, voir (Granger, Hung, and Petch-Tyson, 2002), (Pravec, 2002), (Tono, 2003).

2.1.2. Productions issues des exercices

Les *données sollicitées* dans un cadre expérimental, permettant de mieux contrôler les facteurs situationnels, doivent nécessairement contenir les phénomènes spécifiques qui ont été

[1] Acronyme de *Connaître / Comprendre / Corriger les Erreurs en Tchèque Langue Étrangère pour les Francophones.*

établis comme objet de l'investigation. Néanmoins, l'authenticité de ces données peut être contestée, ainsi que leur ambition de refléter l'état réel de la compétence de l'apprenant. Par rapport à un corpus de productions libres, le recueil de données produites dans des exercices ciblés sur une compétence spécifique peut apporter plus rapidement des données pertinentes. Les informations sur l'emploi d'une certaine structure, obtenues à l'aide des exercices, seraient beaucoup plus éparses dans un corpus de productions libres et le nombre de leurs occurrences serait proportionnel à sa taille.

2.1.3. Arguments pour les exercices

CETLEF permet de collecter les données au sein d'exercices grammaticaux contenant des tâches de déclinaison : les formes requises dans de telles tâches peuvent être facilement accompagnées par une annotation morphologique (la catégorie lexicale, les catégories morphologiques, le type paradigmatique et l'indication d'une éventuelle alternance), ajoutée manuellement ou avec des méthodes semi-automatiques lors de la création des exercices. Le stockage des productions dans une base de données relationnelle permet leur exploitation efficace. De plus, la plateforme proposant des exercices peut intégrer des fonctionnalités supplémentaires à visée didactique.

2.2. CETLEF comme un outil d'enseignement de langue assisté par ordinateur

L'enseignement ou l'apprentissage des langues assisté par ordinateur (ELAO ou ALAO, CALL pour Computer Assisted Language Learning) est un domaine pluridisciplinaire dont l'objet est l'intégration d'outils informatiques dans l'enseignement des langues, pour des revues synthétiques sur la discipline, voir par exemple (Levy, 1997), (Nerbonne, Jager, and van Essen, 1998), (Cameron, 1999), (Hanson-Smith, 2003). Ces outils sont considérés plutôt comme un complément de l'enseignement traditionnel qu'une alternative à celui-ci, voir (Bertin, 2001).

D'après (Karttunen, 1986), (Zock, 1996), (Nerbonne, 2003) et d'autres, l'enseignement des langues assisté par ordinateur est un domaine idéal pour la vérification des fonctionnalités des techniques de TAL, car la tâche d'assister un apprenant dans son apprentissage implique virtuellement tous les objectifs visés par cette discipline. L'intégration de messages de diagnostic des erreurs, dans les outils d'enseignement assistés par ordinateur, est considéré comme un point positif au niveau didactique, voir par exemple (Heift and Schulze, 2003), (L'haire and Vandeventer-Faltin, 2003), (Heift and Schulze, 2007).

Des méthodes de correction et de diagnostic des erreurs peuvent être appliquées soit sur des productions libres, soit sur des productions provenant de tâches fermées comme dans les exercices grammaticaux. Pour le traitement des productions libres, différentes techniques sont expérimentées pour adapter les correcteurs orthographiques et grammaticaux, destinés à l'usage universel, afin qu'ils puissent prendre en compte les spécificités des textes produits par des apprenants étrangers.

Par rapport à l'imperfection actuelle des outils disponibles pour la correction des productions libres, (Holland and Kaplan, 1995), (Kraif et al., 2004), (Tschichold, 2006) estiment néces-

saire l'adoption d'une approche «pédagogiquement responsable», favorisant l'emploi de techniques de base qui sont suffisamment bien maîtrisées pour réduire le bruit ou le silence à la sortie du traitement. Ces imperfections, qui peuvent être acceptables pour certaines applications dans leur usage «non pédagogique», se révèlent particulièrement perturbantes pour un apprenant au sein d'un didacticiel.

Dans la perspective de l'emploi des techniques de TAL pour la correction des erreurs dans un outil ELAO, nous estimons qu'un diagnostic des erreurs, issues des exercices grammaticaux à trous, peut être effectué par des procédés relativement simples et fiables, basés sur la génération morphologique.

Fig. *1. Page d'entrée de la plateforme apprenant de CETLEF*

3. Annotation morphologique

Dans le cadre de CETLEF, nous avons élaboré un modèle de la déclinaison du tchèque répondant aux objectifs suivants :

1. Annoter les productions des apprenants par des métadonnées linguistiques qui seraient utiles pour leur analyse. Comme dans le cas des corpus annotés, il s'agit de **faciliter la**

recherche de l'information à l'aide des étiquettes linguistiques.

2. Servir pour le **diagnostic automatique des erreurs**. L'indication des valeurs des catégories grammaticales, du type paradigmatique et de l'alternance est une information cruciale pour l'interprétation automatique d'une erreur de déclinaison.
3. Être affiché, dans un format adapté pour une **présentation didactique**, sur la plateforme apprenant en tant qu'assistance dans l'apprentissage.

3.1. Définition des types paradigmatiques

Le classement des types paradigmatiques a été fondé sur la tradition grammaticale tchèque (14 types de déclinaison de bases), nous définissons cependant une classification détaillée des différents sous-types et des exceptions. Contrairement aux autres annotations morphologiques utilisées pour le traitement automatique du tchèque, voir (Hajič, 2004) et (Osolsobě, 1996), cette classification est établie uniquement par les différences dans les ensembles de désinences qui s'attachent au radical du lexème pour créer une forme déclinée. La réalisation des alternances vocaliques ou consonantiques n'est pas prise en compte pour la définition des types paradigmatiques ce qui permet de diminuer leur nombre.

3.1.1. Représentation d'un paradigme

Un exemple de la représentation complète du paradigme de désinences casuelles d'un certain sous-type est présenté sur la table 1. Cette table représente le paradigme de désinences casuelles du sous-type *hoch*, défini par rapport au sous-type modèle *pán* du type *pán*. Il se distingue par le remplacement (valeur de l'attribut ET - écart du type) de la désinence *-e* par *-u* dans le vocatif singulier (*pan-e* × *hoch-u*), par le remplacement de la désinence *-ech* par *-ích* dans le locatif pluriel et par l'ajout d'une variante de registre *-ách* pour le même cas. Les désinences qui sont des variantes fonctionnelles ou des variantes de registre sont marquées par la valeur correspondante de l'attribut var (variante).

3.2. Alternances vocaliques et consonantiques

Afin de pouvoir annoter la réalisation des alternances vocaliques ou consonantiques, et de déterminer les règles de leur réalisation, nous avons entrepris une étude basée sur l'ensemble des 50 000 lexèmes du tchèque les plus fréquents dans (Čermák and Křen, 2004). Les différentes configurations dans lesquelles une alternance peut être effectuée ont été examinées afin de définir les règles de leur réalisation et des listes d'exceptions.

3.3. Annotation des tâches dans les exercices

Avec ce répertoire d'étiquettes morphologiques, formatées dans des fichiers XML *pdgm.xml* pour les types paradigmatiques et *alt.xml* pour les alterances, la forme requise dans une tâche est annotée manuellement par l'auteur des exercices. Vu le nombre de tâches qui sont proposées

cas	num	gen	var	ET	term
nom	sg	m			#
gen	sg	m			a
dat	sg	m	fnct		u
dat	sg	m	fnct		ovi
acc	sg	m			a
voc	sg	m		R	u
loc	sg	m	fnct		u
loc	sg	m	fnct		ovi
inst	sg	m			em

cas	num	gen	var	ET	term
nom	pl	m	fnct		i
nom	pl	m	fnct		ové
gen	pl	m			ů
dat	pl	m			ům
dat	pl	m	reg		um
acc	pl	m			y
voc	pl	m	fnct		i
voc	pl	m	fnct		ové
loc	pl	m		R	ích
loc	pl	m		A	ách
inst	pl	m			y
inst	pl	m	reg		ama

Tab. *1. Représentation du paradigme de désinences casuelles d'un sous-type paradigmatique (Valeurs de l'attribut* `var` *:* `fnct` *- variante fonctionnelle,* `reg` *- variante de registre)*

dans le cadre d'une enquête contenant des exercices, une annotation manuelle ne représente pas un travail inabordable.

L'annotation d'une forme requise, telles qu'elle est stockée dans la base de données, contient les informations suivantes :

requis	lemme	tagLex	tagMorph	pdgm	cas	num	gen	alt
hub	houba	subst	N	zn	gen	pl	f	ou > u

Tab. *2. Exemple de l'annotation d'une forme requise*

Explication des noms des attributs : `requis` : *la forme requise*, `lemme` : *lemme*, `tagLex` : *catégorie lexicale*, `tagMorph` : *type morphologique*, `pdgm` : *type paradigmatique*, `cas` : *cas*, `num` : *nombre*, `gen` : *genre*, `alt` : *alternance*.

3.4. Exploitation didactique de l'annotation

Une des fonctions principales du modèle de la déclinaison est l'affichage des informations d'ordre didactique à l'apprenant. L'annotation des formes requises sont affichées pendant la correction de l'exercice. Pour l'exemple, voir la figure 2 avec l'affichage de l'annotation de la forme requise, dont l'annotation a été présentée dans la table ci-dessus (tab. 2). Il s'agit de la correction d'un exercice effectué.

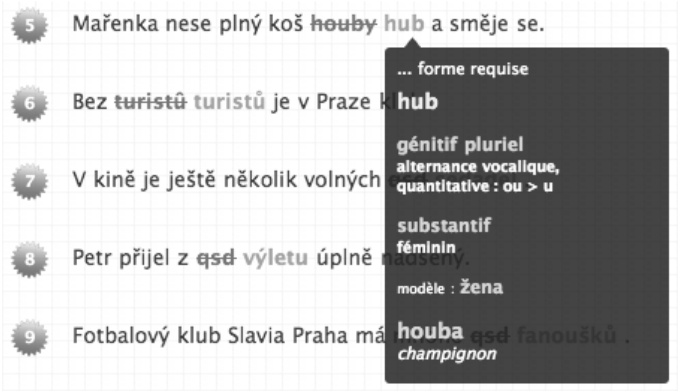


Fig. *2. Annotation d'une forme requise sur la plateforme apprenant*

Les éléments linguistiques que l'apprenant rencontre dans les exercices, remplissent sa base de données personnelle qui peut être consultée à tout moment et qui peut servir comme outil d'apprentissage. L'augmentation du volume de cette base peut être un facteur motivant pour l'apprenant. Les éléments qui nourrissent cette «base de connaissances» sont classées en quatre sections : le lexique contenu dans les exercices terminés ; les paradigmes des formes à décliner ; les alternances à effectuer dans les formes à décliner ; le cas et le nombre des formes à décliner.

3.4.1. Lexique

La section *Lexique* contient une liste alphabétique des mots tchèques et de leur traduction française qui apparaissent dans les exercices soit en tant que formes requises, soit comme appartenant au contexte gauche ou droit d'une phrase donnée. Le lien hypertexte derrière chaque entrée permet d'afficher la tâche dans laquelle cette entrée est apparue.

3.4.2. Paradigmes

Dans la section *Paradigmes*, l'apprenant peut choisir l'un des types paradigmatiques pour afficher la liste des terminaisons (voir fig. 3) et les mots rencontrés dans les exercices qui appartiennent à ce paradigme. Seules les terminaisons du modèle de type paradigmatique sont présentées, les variantes du registre ne sont pas affichées. Les différences dans la déclinaison des mots dans le lexique par rapport à leur type modèle sont mises en évidence (voir fig. 4).

3.4.3. Alternances

La section *Alternances* permet de visualiser les différentes occurrences des alternances rencontrées dans les tâches des exercices. Pour afficher les exemples de la réalisation d'une al-

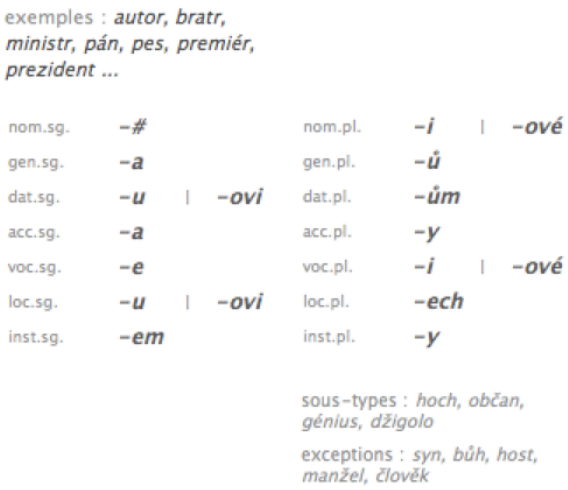


Fig. *3. Présentation d'un type paradigmatique*

Fig. *4. Affichage des différences dans la déclinaison d'un sous-type*

ternance dans les exercices, l'apprenant est invité d'abord à choisir un type d'alternance (par exemple vocalique quantitative, palatalisation A, mouillure etc.), puis une alternance particulière (par exemple *k* > *c*). Au passage de la souris sur le couple *lemme – forme alternée*, il est possible de visualiser une vignette portant des informations sur le mot alterné et l'alternance elle même (voir fig. 5).

Fig. *5. Vignette accompagnant une alternance*

3.4.4. Cas

La section *Cas* permet de consulter les formes requises dans les exercices en fonction du nombre grammatical et du cas (voir fig. 6).

Fig. *6. Consultation des formes requises par nombre et par cas*

4. Diagnostic des erreurs

Dans le diagnostic automatique des erreurs, une production erronée est considérée comme une combinaison inappropriée du radical de la forme requise et d'une désinence casuelle. Grâce

à l'annotation d'une forme requise, il est possible de générer automatiquement différentes formes hypothétiques qui pourraient être produites par un apprenant et ces formes hypothétiques sont comparées à la production erronée. S'il y a une correspondance, la production erronée est interprétée par les différentes propriétés de la forme hypothétique correspondante. Cette approche est basée sur les hypothèses présentées ci-dessous.

4.1. Hypothèses sur les erreurs possibles

Les hypothèses sur les erreurs possibles, commises dans les exercices de déclinaison par des apprenants francophones, peuvent être établies sur la base de la comparaison du système nominal tchèque et français. La déclinaison du tchèque présente pour un apprenant français, ou pour tout autre apprenant dont la langue maternelle ne dispose pas de la déclinaison, une sorte d'**idiosyncrasie** par rapport au système de sa langue maternelle. La variation des formes fléchies rajoute de la complexité dans la production langagière et peut être naturellement une source d'erreurs.

Du point de vue de l'activité de production langagière de l'apprenant, il est possible de distinguer plusieurs étapes, qui sont nécessaires pour produire une forme casuelle correcte au sein d'une tâche de déclinaison et qui peuvent mener à l'erreur : (1) le choix des valeurs de la catégorie du cas, du nombre et du genre ; (2) le classement du lexème dans le paradigme approprié et le choix de la désinence correspondante aux valeurs des catégories grammaticales ; (3) la réalisation, si cela est nécessaire, des alternances vocaliques ou consonantiques.

4.1.1. Choix des valeurs des catégories grammaticales

L'étape (1) est effectuée en fonction des critères purement syntaxiques pour l'attribution du cas. Les valeurs du genre et du nombre sont attribuées en fonction de l'accord entre la forme requise et son régisseur où en fonction des critères d'ordre sémantiques, relatifs au contenu cognitif exprimé par la phrase. Un dysfonctionnement dans cette opération serait reflété dans la production par le choix de désinences exprimant les valeurs inappropriées des catégories respectives.

4.1.2. Attribution du paradigme casuel au lexème

L'attribution d'un certain type paradigmatique au lemme de la forme requise pendant l'étape (2) délimite le répertoire des désinences permettant d'exprimer les valeurs des catégories grammaticales choisies à l'étape précédente. Nous pouvons supposer l'existence de formes produites par les apprenants qui peuvent être correctes, en ce qui concerne les valeurs des catégories grammaticales liées à la désinence choisie pour la génération d'une certaine forme, mais qui ne sont pas appropriées pour exprimer les significations grammaticales au sein du paradigme propre à la forme requise. Par exemple, le génitif pluriel *turist* du lexème *turista* n'est pas généré d'après le paradigme approprié (type *předseda*), demandant la désinence *-ů* dans le génitif pluriel, mais d'après le type *žena* qui emploie effectivement la désinence zéro -# pour exprimer le cas et le nombre correspondants.

4.1.3. Réalisation des alternances

L'étape (3) représente une opération sur la forme composée du radical et la désinence. La réalisation des alternances est conditionnée par des facteurs phonologiques, morphologiques et lexicaux et la maîtrise des règles de leur réalisation est nécessaire pour la création adéquate des formes casuelles. Nous pouvons supposer, qu'un apprenant fera des erreurs dans les alternances, comme par exemple *houb* au lieu de *hub* dans le génitif pluriel du substantif *houba* (*champignon*), etc.

4.2. Définition des types d'erreurs

Les différents types d'erreurs sont définis sur la base des propriétés morphologiques des formes hypothétiques, générées à partir du radical de la production requise et des désinences employées pour la génération des formes casuelles du tchèque.

4.2.1. Définition préliminaires

1. Soit un alphabet L, ensemble fini de caractères ; soit un langage $L^\star$, ensemble infini de toutes les chaînes possibles sur l'alphabet L ;
2. Soit un alphabet $L_{CZ} \subset L$, ensemble fini contenant tous les caractères du tchèque à part l'espace, les chiffres et les signes de ponctuation, et qui est une union des ensembles de caractères minuscules, majuscules, minuscules diacritées et majuscules diacrités ; soit un langage $L_{CZ}{}^\star \subset L^\star$, ensemble infini de toutes les chaînes possibles sur l'alphabet L_{CZ} ;
3. Soit une fonction Min, opération de minusculisation qui attribue à chaque mot $m \in L_{CZ}{}^\star$ sa forme correspondante uniquement en minuscules ; soit une fonction Dia, opération d'enlèvement du diacritique qui attribue à chaque mot $m \in L_{CZ}{}^\star$ sa forme correspondante en caractères sans diacritique ; soit une fonction St, opération de standardisation telle que $St(m) = Dia(Min(m))$;
4. Soit un langage $N \subset L_{CZ}{}^\star$, ensemble fini de toutes les formes lexicales des mots tchèques appartenant aux types morphologiques nominal, adjectival, adjectival mixte, pronominal ou numéral : N = {*abatyše, ..., mládě, mláděte, ..., žížalami*}.
5. Soit un ensemble $R \subset L_{CZ}{}^\star$, ensemble fini de tous les radicaux extraits par l'enlèvement de la désinence casuelle de la forme du lemme des mots tchèques appartenant au type morphologique nominal, adjectival, adjectival mixte, pronominal ou numéral ; soit un ensemble $D \subset L_{CZ}{}^\star$, ensemble fini de toutes les désinences des types paradigmatiques des mots tchèques appartenant au type morphologique nominal, adjectival, adjectival mixte, pronominal.
6. Soit un langage $H \subset L_{CZ}{}^\star$, ensemble fini de toutes les formes lexicales hypothétiques des mots tchèques h appartenant au type morphologique nominal, adjectival, adjectival mixte, pronominal ou numéral ; ainsi que leurs formes minusculisées $Min(h)$, sans diacritique $Dia(h)$ et standardisées $St(h)$. Ces formes sont le résultat de la concaténation

des couples contenus dans le produit des ensembles $R \times D$, avec ou sans la réalisation de l'alternance sur la chaîne résultante

4.2.2. Définition de la forme requise et de la production erronée

1. La **forme requise** r dans une tâche x est un mot tel que $r \in N$. Chaque r est caractérisée par son annotation morphologique.
2. La **production erronée** p dans une tâche x est un mot tel que $p \in L^{\star}$ et $p \neq r$.
3. Une production erronée p **peut être interprétée morphologiquement** si p correspond à une des formes lexicales hypothétiques $h \in H$, générées à partir du radical de la forme requise r.
4. Une production erronée p **ne peut pas être interprétée morphologiquement** si p ne correspond à aucune des formes hypothétiques $h \in H$, générées à partir du radical de la forme requise r.

4.3. Interprétation morphologique

Chaque forme requise r dans le cadre d'une tâche x est caractérisée par son annotation morphologique. Pour les besoins de la description formelle des erreurs, cette annotation peut être représentée à l'aide d'une **structure de traits**. Pour une forme requise r, la structure de traits est la suivante :

$$\begin{bmatrix} cas : cas \\ num : nombre \\ gen : genre \\ alt : identifiant\,de\,l'alternace \\ pdgm : soustype\,paradigmatique \\ tagMorph : type\,morphologique \end{bmatrix}$$

Les différentes valeurs des attributs dans cette structure sont instanciées en fonction des propriétés morphologiques de r inscrites dans l'annotation. Les noms des attributs sont identiques à ceux utilisés dans l'annotation morphologique sur CETLEF.

Par exemple, pour une forme requise *r = matce* qui est le datif singulier du substantif *matka*, l'instanciation de la structure de traits est la suivante :

$$\begin{bmatrix} cas : dat \\ num : sg \\ gen : f \\ alt : k > c \\ pdgm : zn\_Re \\ tagMorph : N \end{bmatrix}$$

L'ensemble des formes hypothétiques h, générées à partir du radical *matka*, est créé par toutes les combinaisons possibles du radical *matk* avec toutes les désinences dans D, avec ou sans la réalisation des alternances, avec ou sans diacritique et avec toutes les possibilités dans la casse des caractères : *matka, matky, ..., matek, matk, ..., matkám, matkam, ..., matkách, matkach, ..., matkovi, matkem, ..., matkému, ...*

À chacune de ces formes h peut être assignée au moins une structure de traits. Les valeurs des attributs dans la structure de h sont déterminées uniquement sur la base de propriétés formelles de ses composants en fonction (1) des différentes valeurs des catégories morphologiques qui peuvent être exprimées par la désinence employée, (2) de la réalisation d'une alternance sur le radical ou (3) de sa forme graphique. Une structure de traits assignée à une forme h est appelée son **interprétation**.

Le nombre des interprétations des formes hypothétiques n'est pas déterminé uniquement par l'homonymie des formes casuelles existantes pour un certain lemme, mais également par toutes les combinaisons possibles du radical et des désinences appartenant aux autres paradigmes, ainsi que par les variation dans la diacritique.

4.4. Erreur d'après l'attribut atteint

En fonction des attributs qui diffèrent dans les structures de r et de p (qui sont atteints par l'erreur), les différents types d'erreurs sont représentés dans la table 3.

attribut	type d'erreur
cas	erreur de cas
num	erreur de nombre
gen	erreur de genre
alt	erreur d'alternance
pdgm	erreur de type paradigmatique
pdgm	erreur de sous-type paradigmatique
tagMorph	erreur de type morphologique
dia	erreur de diacritique
casse	erreur de casse

Tab. *3. Erreurs d'après l'attribut atteint*

Ces erreurs peuvent se combiner librement entre elles en fonction des attributs atteints par l'erreur dans une interprétation donnée. Il peut exister par exemple une erreur de cas et de nombre, une erreur de cas et d'alternance, une erreur de nombre et de graphie, etc. Dans ces appellations, chaque attribut qui contient une valeur différente par rapport à la forme requise doit être spécifié.

4.5. Erreur par rapport au paradigme de la forme requise

Sur la base des observations des erreurs authentiques produites par les apprenants dans les exercices de déclinaison, nous avons établi quatre groupes dans lesquels l'ensemble des interprétations des formes hypothétiques h, employées pour la recherche d'une correspondance avec une production erronée p pour une forme requise r, établi par rapport au paradigme de la forme requise : erreur locale, erreur verticale, erreur horizontale interne, erreur horizontale externe.

4.5.1. Erreur locale

La désinence appartient au sous-type paradigmatique de la forme requise r avec la même valeur de cas, de genre et de nombre. Dans l'exemple (1), la production erronée *Olge* est une erreur locale d'alternance.

(1) *Petr vzal* ***Olge*** *všechny peníze.*
Olze dat.sg.f.
Pierre a pris à Olga tout argent
'Pierre a pris à Olga tout l'argent'

$$\begin{array}{c} Olze \\ \left[\begin{array}{l} cas : dat \\ num : sg \\ gen : f \\ \textbf{alt : g > z} \\ pdgm : zn\_Re \\ tagMorph : N \\ dia : 1 \\ casse : 1 \end{array}\right] \end{array} \neq \begin{array}{c} Olge \\ \left[\begin{array}{l} cas : dat \\ num : sg \\ gen : f \\ \textbf{alt : sans} \\ pdgm : zn\_Re \\ tagMorph : N \\ dia : 1 \\ casse : 1 \end{array}\right] \end{array}$$

4.5.2. Erreur verticale

La désinence appartient au sous-type paradigmatique de la forme requise r avec la valeur de cas autre que celle de la forme requise. Dans l'exemple (2), la production erronée *průvodci* est une erreur verticale de cas.

(2) *Výklad našeho* ***průvodci*** *byl velice zajímavý.*
průvodce gen.sg.m.
Exposé notre guide était très intéressant
'L'exposé de notre guide a été très intéressant'

průvodce		*průvodci*
cas : gen	≠	**cas : dat | loc**
num : *sg*		*num* : *sg*
gen : *m*		*gen* : *m*
alt : *sans*		*alt* : *sans*
pdgm : *sc*		*pdgm* : *sc*
tagMorph : *N*		*tagMorph* : *N*
dia : 1		*dia* : 1
casse : 1		*casse* : 1

4.5.3. Erreur horizontale interne

La désinence appartient aux autres sous-types dans le type paradigmatique de la forme requise r avec les mêmes valeurs de cas, de nombre et de genre. Dans l'exemple (3), la production erronée *výleta* est une erreur horizontale interne de sous-type paradigmatique.

(3) *Petr přijel z* ***výleta** *v Bretani.*
výletu gen.sg.i.
Petr est rentré de voyage en Bretagne

'Pierre est rentré du voyage en Bretagne'

výletu		*výleta*
cas : *gen*	≠	*cas* : *gen*
num : *sg*		*num* : *sg*
gen : *i*		*gen* : *i*
alt : *sans*		*alt* : *sans*
pdgm : hd		**pdgm : hd_l**
tagMorph : *N*		*tagMorph* : *N*
dia : 1		*dia* : 1
casse : 1		*casse* : 1

4.5.4. Erreur horizontale externe

La désinence appartient aux autres types paradigmatiques dans le cadre du même type morphologique avec les mêmes valeurs de cas et de nombre et qui ont la même désinence dans le nominatif singulier comme le lemme de r. Dans l'exemple (4) la production erronée *sole* est une erreur horizontale externe de type paradigmatique (le choix de la désinence *-e* du type *píseň* au lieu de la désinence *-i* du type *kost*).

(4) *Maso bez* ***sole** *není většinou příliš chutné*
soli gen.sg.f
viande sans sel n'est pas d'habitude très appétissant

'La viande sans sel n'est pas d'habitude très appétissante'

$$\begin{array}{ccc} soli & \neq & sole \\ \left[\begin{array}{l} cas : gen \\ num : sg \\ gen : f \\ alt : \mathring{u} > o \\ \textbf{pdgm : kt\_n} \\ tagMorph : N \\ dia : 1 \\ casse : 1 \end{array}\right] & & \left[\begin{array}{l} cas : gen \\ num : sg \\ gen : f \\ alt : \mathring{u} > o \\ \textbf{pdgm : ps\_Re} \\ tagMorph : N \\ dia : 1 \\ casse : 1 \end{array}\right] \end{array}$$

4.6. Diagnostic morphologique d'une production erronée

Le diagnostic morphologique d'une production erronée p dans une tâche x est l'ensemble de ses interprétations qui sont le plus plausibles du point de vue de l'activité langagière de l'apprenant. La plausibilité d'une interprétation peut être établie sur la base des critères morphologiques pour les erreurs locales et horizontales. Par contre, pour les erreurs verticales, où la valeur de la catégorie du cas est une variable, des facteurs syntaxiques entrent nécessairement en jeu.

Le diagnostic automatique sur CETLEF ne peut prendre en compte que les informations morphologiques et son ambition n'est que de proposer la meilleure solution dans le cadre donné. Cette solution peut être par la suite confirmée ou rejetée à l'aide d'une étude «manuelle», effectuée par un humain qui prend en compte des critères divers qui lui permettent de choisir l'interprétation la plus probable.

Dans le cadre morphologique, nous définissons donc que *la plausibilité d'une interprétation est déterminée par le nombre d'attributs atteints par l'erreur*. Moins il y a d'attributs qui diffèrent dans la structure de r et dans une certaine interprétation de p, plus cette interprétation est plausible.

4.7. Description de la procédure de diagnostic des erreurs

La procédure *Diagnostic* est employée pour diagnostiquer les productions qui ne correspondent à aucune des formes requises dans le cadre d'une tâche. Ce diagnostic est effectué par la recherche de différentes interprétations de la production erronée et par le choix de celles qui sont les plus plausibles.

Les données à l'entrée de la procédure sont : la *production erronée*, la *forme requise*, le *lemme* de la forme requise et l'*annotation* de la forme requise. À la sortie de la procédure, un message qui spécifie l'erreur est généré. Ce message d'erreur sert comme critère pour les recherches des productions dans la base de données en fonction des différents types d'erreurs et pour la génération du message de diagnostic, spécifiant la nature de l'erreur sur la plateforme apprenant.

4.7.1. Traitement non morphologique

D'abord, l'interprétation de l'erreur est effectuée à l'aide des techniques simples qui n'impliquent pas l'utilisation des données morphologiques. Le but est d'identifier, à l'aide des calculs sur les

caractères et les chaînes de caractères qui représentent la forme requise et la production erronée, une différence trop importante entre ces deux éléments (distance de Levenshtein (Levenshtein, 1966) trop grande, différence importante de longueur des deux chaînes, etc.), pour qu'il puisse y avoir une interprétation morphologique . Si cette étape n'a pas été suffisante pour déterminer le bon diagnostic, une série de tests morphologiques est commencée pour interpréter l'erreur comme une forme hypothétique générée à partir du radical de la forme requise.

4.7.2. Traitement morphologique

Les tests morphologiques utilisent les informations linguistiques dans l'annotation de la forme requise, le modèle de la déclinaison, structuré dans les fichiers *pdgm.xml* et *alt.xml*, et la procédure *AlterneRadical*, qui assure les changements du radical au contact d'une désinence susceptible de provoquer une alternance vocalique ou consonantique.

Pendant ce traitement, des formes hypothétiques sont générées à partir du radical de la forme requise. Ces formes doivent nécessairement observer les restrictions posées sur les erreurs locales, verticales, horizontales internes et horizontales externes. Chaque forme hypothétique est ensuite systématiquement comparée à la production erronée. S'il y a une correspondance, l'erreur est interprétée sur la base des propriétés morphologiques de cette forme et cette interprétation est inscrite parmi les autres possibles.

4.7.3. Filtrage des interprétations

Pendant cette étape, le message d'erreur est filtré pour réduire au minimum le nombre des interprétations possibles afin d'en retenir uniquement celles qui sont les plus plausibles. Cette réduction est effectuée en fonction du nombre d'attributs atteints par l'erreur qui détermine leur classement dans l'échelle de plausibilité pour un diagnostic.

Prenons les différentes interprétations de l'erreur dans la tâche suivante :

(5) *David nemůže jíst* ****rajče.***
rajčata acc.pl.n.
David ne peut pas manger les tomates
'David ne peut pas manger des tomates'

La comparaison de la production erronée avec les formes hypothétiques détermine qu'il peut s'agir des erreurs suivantes : (a) une erreur verticale de nombre d'après l'accusatif singulier ; (b) une erreur verticale de cas et de nombre d'après le nominatif singulier ; (c) une erreur verticale de cas et de nombre d'après le vocatif singulier ; (d) une erreur horizontale externe de type paradigmatique d'après l'accusatif pluriel du type *moře* ; (e) une erreur horizontale externe de type paradigmatique et de genre d'après l'accusatif pluriel du type *růže* ; (f) une erreur horizontale externe de type paradigmatique et de genre d'après l'accusatif pluriel du type *soudce*.

Pendant l'étape de filtrage des interprétations possibles, les interprétations retenues comme les plus probables sont les interprétations (a) et (d) : l'apprenant se trompe soit dans le nombre et met la forme du singulier au lieu de la forme du pluriel ; soit il confond le type paradigmatique

kuře avec le type *moře*, qui a le même genre. Cette décision est prise sur la base du nombre de traits morphologiques atteintes par l'erreur : il s'agit d'un seul trait pour les interprétations (a) et (d) ; et de deux traits pour les interprétation (b), (c), (e) et (f). Sur la base de ce critère, les interprétations (b), (c), (e), (f) peuvent être rejetées, car deux interprétations, classées plus haut sur l'échelle de la plausibilité, ont été trouvées.

4.7.4. Formatage du diagnostic

La dernière étape de la procédure *Diagnostic* consiste en une «traduction» des interprétations retenues dans le message filtré en langue naturelle pour qu'elles puissent être publiées sur la plateforme apprenant afin de servir comme une explications des production erronées. Cette procédure est basée sur un principe simple de transfert des valeurs contenues dans le message d'erreur filtré dans des phrases préformatées avec des variables à instancier. L'exemple d'un message de diagnostic d'une production erronée est affiché sur la figure 7.

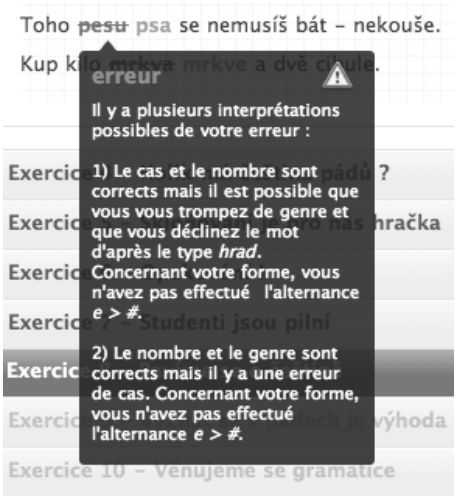


Fig. *7. Exemple d'un message de diagnostic*

5. Évaluation

Nous avons menée une étude pilote visant à tester le dispositif avec des données authentiques recueillies par deux enquêtes différentes. Il s'agit d'une illustration des possibilités de CETLEF dont l'objectif principal est de montrer des exemples d'enquêtes qui sont menées grâce à cet outil et qui peuvent être utilisées pour une recherche sur l'acquisition de la déclinaison du

tchèque par les francophones.

Dans l'enquête publique, qui a eu lieu sur www.cetlef.fr, 159 exercices ont été envoyés à la correction au cours d'une période de deux mois. Au sein de ces exercices, 1551 tâches ont été effectuées. Le nombre de productions erronées dans ces tâches est égal à 442 (28,5 % de toutes les productions). Pour 61 productions erronées (13,8 % de toutes les productions erronées), le diagnostic automatique n'a pas réussi à identifier la nature de l'erreur. Une représentation schématique de cette situation est proposée sur la figure 8.

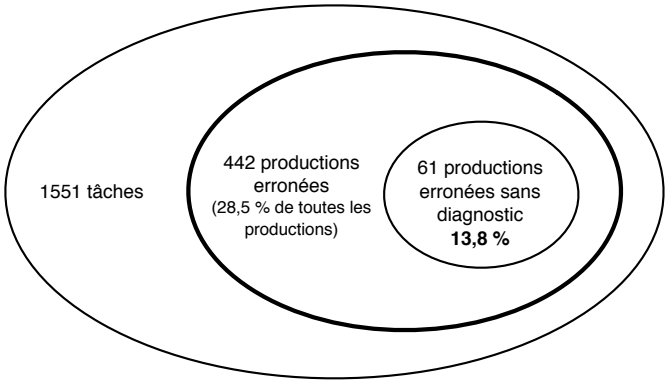


Fig. *8. Pourcentage de productions erronées et sans diagnostic*

Parmi toutes les productions erronées, il y en a 373 (84,4 %) qui ont été diagnostiquées comme une erreur avec une ou plusieurs interprétations morphologiques. Le nombre de productions erronées, diagnostiquées comme un des types d'erreurs établies d'après les restrictions sur les formes hypothétiques (ou une combinaison possible de deux types différents), est présentée dans la table 4.

5.0.5. Erreurs locales

Dans l'enquête publique, les erreurs les plus fréquentes parmi les erreurs locales sont les **erreurs de diacritique** (93 productions erronées, 75,6 %). L'explication de ce fait devrait être cherché dans les raisons techniques de saisie des caractères diacrités.

Les **erreurs d'alternance** sont le second type d'erreurs locales le plus fréquent (20 productions erronées, 16,3 %). Il est intéressant de constater que ce sont uniquement les formes requises avec les alternances vocaliques quantitatives qui étaient à l'origine de ces erreurs (*kráv* au lieu de *krav*, *jmén* au lieu de *jmen*, *díl* au lieu de *děl*, *houb* au lieu de *hub*, *můžem* au lieu de *nožem*, *penízmi* au lieu de *penězmi*, *smlouv* au lieu de *smluv*, etc.).

type de l'erreur	*productions*	%
erreur locale	123	33,0
erreur verticale	145	38,9
erreur horizontale interne	2	0,5
erreur horizontale externe	19	5,1
erreur locale ou verticale	39	18,5
erreur verticale ou horizontale interne	1	0,2
erreur verticale ou horizontale externe	40	10,7
erreur horizontale interne ou externe	4	1,1
sans diagnostic	61	

Tab. *4. Pourcentage des types d'erreurs*

5.0.6. Erreurs verticales

Parmi les erreurs verticales, les productions les plus fréquentes sont les **erreurs de cas** (71 productions, 49,0 %) qui reflètent probablement des dysfonctionnements au niveau syntaxique. Les erreurs les plus courantes sont celles où la production erronée a été laissée au nominatif singulier.

Le second groupe d'erreurs verticales les plus fréquentes sont les **erreurs de nombre** (19 productions avec un remplacement d'une forme de pluriel par une forme de singulier ; 8 productions avec un remplacement d'une forme de singulier par une forme de pluriel).

5.0.7. Erreurs horizontales externes

Contrairement aux erreurs horizontales internes (seulement deux occurrences), le diagnostic des erreurs horizontales externes se montre plutôt satisfaisant. Les plus nombreuses sont les **erreurs de type paradigmatique et de genre**, causées par la confusion du type paradigmatique en fonction des ambiguïtés qui peuvent exister dans l'attribution de ce type à un mot en fonction de sa forme. Par exemple, dans les tâches (5) et (6), il s'agit d'une confusion entre les types *předseda* et *žena*.

(5) *Bez* ****turist*** *je v Praze klid*
turistů gen.sg.m
sans touristes est à Prague calme

'Sans touristes, Prague est calme'

Diagnostic : *Le cas et le nombre sont corrects mais peut être que vous vous trompez de genre et que vous déclinez le mot d'après le type* žena.

(6) *Hanička se líbí* ****Jirce****.*
Jirkovi gen.sg.f
Hanička *refl* plaît à Jirka

'Hanička plaît à Jirka'

Diagnostic : *Le cas et le nombre sont corrects mais peut être que vous vous trompez de genre et que vous déclinez le mot d'après le type* žena.

5.0.8. Erreurs locales ou verticales

La majorité des productions erronées pour lesquelles le diagnostic propose soit une interprétation locale, soit une interprétation verticale, sont des **erreurs de diacritique** en ce qui concerne l'interprétation locale. Le diagnostic peut être jugé comme adéquat si la seconde interprétation – l'interprétation verticale – est **une erreur de cas ou de nombre** sans contenir une erreur de diacritique.

5.0.9. Erreurs non morphologiques

Les productions qui ont été diagnostiquées comme n'ayant aucune interprétation morphologique mais qui remplissent cependant une des conditions posées dans les tests non morphologiques sont les suivantes : la production *pesmrkev* au lieu de *pes* a été diagnostiquée comme trop longue. Il s'agit évidemment d'une inattention, les productions de deux tâches distinctes ont été saisies dans un seul champ de formulaire sur la plateforme apprenant.

La production *rad* au lieu de *radost* a été diagnostiquée comme trop courte. Il s'agit ici probablement d'une confusion avec l'adjectif nominal *rád*, figurant dans la locution *Jsem rád, že ...* (*Je suis content que ...*), et qui remplace dans cette tâche la forme casuelle appropriée du substantif *radost*.

Deux productions ont été éliminées du traitement morphologique grâce au test sur la distance de Levehnstein entre la forme requise et la production erronée. Il s'agit des productions *houbovych* au lieu de *hub* et *vejcatach* au lieu de *vajec*.

5.0.10. Erreurs sans diagnostic

Les productions qui n'ont pas été diagnostiquées parmi les types d'erreurs présentés ci-dessus représentent 13,8 % de toutes les productions erronées. La plus grande partie de ces productions contient des fautes de frappe, manifestées le plus souvent par un ajout, un remplacement ou un effacement d'un graphème dans le radical de la production erronée. Cette modification rend le radical distinct par rapport au radical de la forme requise mais cette différence n'est pas assez prononcée pour que la production erronée puisse être diagnostiquée dans les traitements non morphologiques. Il s'agit par exemple de la production *spolužci* au lieu de *spolužáci*, *mínosti* au lieu de *místnosti*, *pkoje* au lieu de *pokoje*, *studenky* au lieu de *studentky*, *písnchí* au lieu de *písní*, etc.

6. Conclusion et perspectives

Nous estimons que l'apport principal de notre travail est l'intégration d'une riche représentation morphologique dans un outil d'enseignement de langue assisté par ordinateur.

Notre hypothèse que les erreurs de déclinaison sont calculables a été éprouvée dans le diagnostic automatique. L'application du diagnostic sur un échantillon de productions authentiques, collectées sur CETLEF, a permis de vérifier que cette hypothèse est vraie pour une grande partie de productions. La majorité des erreurs peut être interprétée automatiquement comme une combinaison du radical de la forme requise et d'une désinence.

Pour l'évaluation globale du diagnostic automatique, il est nécessaire de considérer la situation spécifique dans laquelle les erreurs analysées ont été produites. En déclinant une forme au sein d'un exercice grammatical, l'apprenant et la machine procèdent effectivement d'une manière assez semblable au niveau de l'analyse et de la génération. De ce point de vue, il serait intéressant d'étudier les occurrences des erreurs décrites dans ce travail dans les productions libres où l'apprenant n'est pas limité à un cadre défini aussi strictement que dans la tâche d'un exercice.

L'utilité du diagnostic pour la recherche des différents types d'erreurs dans la base de données est indéniable. Grâce à l'annotation morphologique des formes requises et grâce au message d'erreur caractérisant les productions erronées, des recherches basées sur cette annotation peuvent être effectuées facilement et servir des analyses variées, comme nous l'avons illustré avec un échantillon de données recueilli sur CETLEF. Avec le nombre croissant de productions dans la base de données au cours du temps, il sera possible d'entreprendre des études d'une envergure plus grande.

L'adéquation du diagnostic du point de vue didactique est une question qui reste ouverte pour le moment. Des modifications du diagnostic se révéleront nécessaires au fur et à mesure du service de CETLEF, avec le volume croissant de différentes productions erronées. Comme une des perspectives, il serait souhaitable de l'améliorer par l'intégration d'informations syntaxiques qui permettraient d'enrichir les critères pour le choix de l'interprétation la plus probable, pour identifier automatiquement des erreurs d'accord, des erreurs dans l'attribution d'une rection à un mot, etc.

L'outil CETLEF permet des analyses de volumes de données plus importantes que celles qui ont été exploitées dans notre travail. Ceci devrait permettre d'effectuer une analyse des erreurs dans l'acquisition de la déclinaison, qui serait menée non pas uniquement sur la base de critères morphologiques, comme c'est le cas dans notre travail, mais qui pourrait prendre en compte des critères plus complexes, comme les interférences entre les deux langues.

Remarque Cet article est le résumé de la thèse (Šmilauer, 2008), élaborée dans le cadre d'un doctorat en cotutelle entre le laboratoire LALIC-CERTAL (Langues, Logiques, Informatique, Cognition – Centre de Recherche en Grammaire et Traitement Automatique des Langues) de l'INALCO (Institut National des Langues et Civilisations Orientales) à Paris, et le laboratoire ÚTKL (Institute of Theoretical and Computational Linguistics) de la Faculté des Lettres de l'Université Charles de Prague.

Bibliographie

Bautier-Castaing, Elisabeth. 1977. Acquisition comparée de la syntaxe du français par des enfants fran-

cophones et non francophones. Étude expérimentale de quelques stratégies d'apprentissage. *Étude de linguistique appliquée*, 27 :19–41.

Bertin, Jean-Claude. 2001. *Des outils pour des langues. Multimédia et Apprentissage.* Ellipses Éditions, Paris.

Besse, Henri and Rémy Porquier. 1991. *Grammaires et didactiques des langues.* Hatier / Didier, Paris.

Cameron, Keitt, editor. 1999. *CALL : Media, Design and Applications*. Swets & Zeitlinger, Lisse.

Čermák, František and Michal Křen. 2004. *Frekvenční slovník češtiny.* Nakladatelství Lidové Noviny, Praha.

Gaonac'h, Daniel. 1991. *Théories d'apprentissage et acquisition d'une langue étrangère.* Hatier / Didier, Paris.

Granger, Sylviane, Joseph Hung, and Stephanie Petch-Tyson. 2002. *Computer learner corpora, second language acquisition and foreign language teaching.* Benjamins, Amsterdam.

Hajič, Jan. 2004. *Disambiguation of Rich Inflection. Computational Morphology of Czech.* Karolinum, Praha.

Hanson-Smith, Elizabeth. 2003. A brief history of CALL theory. *CATESOL Journal*, 15(1) :21–30.

Heift, Trude and Mathias Schulze. 2003. Error diagnosis and error correction in CAL : Introduction. *CALICO*, 20(3) :433–436.

Heift, Trude and Mathias Schulze. 2007. *Errors and Intelligence in Computer-Assisted Language Learning : Parsers and Pedagogues*. Routledge, UK.

Holland, V. Melissa and Jonathan D. Kaplan. 1995. NLP techniques in CALL : Status and instructional issues. *Instructional Science*, 23 :352–380.

Hrdlička, Milan. 2002. *Cizí jazyk čeština.* ISV, Praha.

Karttunen, F. 1986. A linguist looks at computer-assisted instruction. In Reinhold Freudenstein and James C. Vaughan, editors, *Confidence Through Competence in Modern Language Learning. CILT Reports & Papers 25*.

Kraif, Olivier, Georges Antoniadis, Sandra Echinard, Mathieu Loiseau, T. Lebarbé, and Claude Ponton. 2004. NLP tools for CALL : the simpler, the better. In *Proceedings of InSTIL / ICALL2004 – NLP and Speech Technologies in Advanced Language Learning Systems*, Venice, 17-19 June.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8) :707–710.

Levy, Michael. 1997. *Computer-Assisted Language Learning : Context and Conceptualization.* Clarendon Press, Oxford.

L'haire, Sébastien and Anne Vandeventer-Faltin. 2003. Diagnostic d'erreurs dans le projet FreeText. *ALSIC : Apprentissage des Langues et Systèmes d'Information et de Communication*, 6(2) :21–37.

Nekula, Marek. 2007. Systém a úzus. K výuce české deklinace se zřetelem k substantivům. In Jana Čemusová and Lída Holá, editors, *Sborník Asociace učitelů češtiny jako cizího jazyka 2006-2007*. Akropolis, Praha, pages 23–47.

Nerbonne, John. 2003. Natural language processing in computer-assisted language learning. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*. Oxford University Press, pages 670–698.

Nerbonne, John, Sake Jager, and Arthur van Essen, editors. 1998. *Language Teaching and Language Technology*. Swets & Zeitlinger, Lisse.

Osolsobě, Klára. 1996. *Algoritmický popis formální morfologie a strojový slovník češtiny*. Ph.D. thesis, Filozofická fakulta Masarykovy univerzity, Brno.

Poldauf, Ivan and Karel Špruňk. 1968. *Čeština jazyk cizí*. Státní pedagogické nakladatelství, Praha.

Porquier, Rémy. 1977. L'analyse des erreurs. Problèmes et perspectives. *Étude de linguistique appliquée*, 25 :23–43.

Pravec, Norma A. 2002. Survey of learner corpora. *ICAME Journal*, 26 :81–114.

Šmilauer, Ivan. 2008. *Acquisition du tchèque par les francophones : analyse automatique des erreurs de déclinaison*. Ph.D. thesis, FF UK, INALCO, Prague, Paris.

Tono, Yukio. 2003. Learner corpora : design, development and applications. In Dawn Archer, Paul Rayson, Andrew Wilson, and Tony McEnery, editors, *Proceedings of the Corpus Linguistics 2003 conference*, pages 800–809.

Tschichold, Cornelia. 2006. Intelligent CALL : The magnitude of the task. In P. Mertens, C. Fairon, A. Dister, and P. Watrin, editors, *Verbum ex machina. Actes de la 13e conférence sur le Traitement automatique des langues naturelles*, pages 806–814, Louvain-la-Neuve. Presses universitaires de Louvain.

Zock, Michael. 1996. Computational linguistics and its use in real world : The case of computer assisted-language learning. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, pages 1002–1004.

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 90 DECEMBER 2008 57–68

Towards English-to-Czech MT via Tectogrammatical Layer

Ondřej Bojar, Silvie Cinková, Jan Ptáček

Abstract

We present an overview of an English-to-Czech machine translation system. The system relies on transfer at the tectogrammatical (deep syntactic) layer of the language description. We report on the progress of linguistic annotation of English tectogrammatical layer and also on the first end-to-end evaluation of our syntax-based MT system.

1. Introduction

Current state-of-the-art machine translation (MT) systems are mostly statistical and phrase-based[1]. In recent years the performance of (surface) syntax-based systems has improved and as a result they are approaching state-of-the-art performance levels (Zollmann and Venugopal, 2006, Quirk and Menezes, 2006, Chiang, 2005).

Our long-term goal is to improve English-Czech MT quality by introducing a transfer step at a deep syntactic layer, making explicit use of linguistic theories and annotated data. For the time being, parts of the annotated data as well as the whole pipeline of automatic deep syntactic analysis, syntactic transfer and a generation component still constitute just work in progress. Nevertheless, we are able to deliver a first end-to-end evaluation that will serve as a baseline for the future improvements of the system.

In Section 2, we give a brief overview of the tectogrammatical representation. Section 3 summarizes our ongoing efforts in annotating English texts at the tectogrammatical layer. In Section 4, we describe both formal and implementational aspects of our MT system and Section 5 compares and discusses automatically assessed translation quality of several configurations of our system.

[1] See NIST evaluation: `http://www.nist.gov/speech/tests/mt/`

Please cite this article as: Ondřej Bojar, Silvie Cinková, Jan Ptáček, Towards English-to-Czech MT via Tectogrammatical Layer. The Prague Bulletin of Mathematical Linguistics No. 90, 2008, 57–68.

2. Overview of the Tectogrammatical Representation

2.1. Functional Generative Description and Treebank Annotation

The tectogrammatical language representation is an implementation of the Functional Generative Description (FGD, Sgall, Hajičová, and Panevová, 1986). FGD has been implemented in treebank annotations. The Prague Dependency Treebank (PDT 2.0, Hajič et al., 2006) consists of three interlinked annotation layers, corresponding to the three FGD-original levels: the morphological layer (m-layer; 2 million words), the analytical layer (a-layer, an auxiliary step reflecting surface syntax; 1.5 million words) and the tectogrammatical layer (t-layer; 0.8 million words).

The FGD as well as the treebank annotation focus on the tectogrammatical language (t-) level. Being a transition between syntax and semantics (sometimes also referred to as *underlying syntax/deep syntax*), the tectogrammatical language level captures the linguistic meaning of each sentence, describing mutual syntactic and semantic relations between the respective words in a sentence, including those of coreference and topic-focus articulation in a broader context scope. FGD has a strong valency theory (Panevová, 1980, Panevová, 1974, Panevová, 1975). The valency theory of FGD assigns valency frames to verbs, nouns, adjectives and certain types of adverbs, assigning semantic roles to their complementations.

2.2. Trees, Nodes and Edges

In the treebank annotation, every sentence is represented as a rooted dependency tree with labeled nodes and edges. The tree reflects the underlying (deep) structure of the sentence. Several types of edges specify whether the relation between two nodes is a dependency relation or not (e.g. the relation between the sentence predicate and an interjection or a disjunct is not that of dependency, although the predicate and the other node are connected by an edge).

Unlike the surface-syntax representation (a-layer), only autosemantic words[2] have their own nodes in the tectogrammatical tree structures. Function words like auxiliaries, subordinating conjunctions and prepositions as well as several cognitive, syntactic and morphological categories are attached to the respective nodes as a set of attribute-value pairs. The presence or absence of an attribute in a given node is determined by its node type.

2.3. Valency

Each occurrence of a part of speech that is considered to have valency is assigned a valency frame from a valency lexicon, interlinked with the data[3]. Obligatory complementations that are not present in the surface representation of the sentence get their tectogrammatical repre-

[2] Several artificially generated complementary nodes for coordination, apposition, reciprocity, etc., and the technical root node also have their own t-nodes, although they do not necessarily have a corresponding node in the surface structure.

[3] In the current annotation, this is restricted to verbs and certain types of nouns.

sentations by means of artificially added nodes. These nodes specify whether the missing information can be retrieved from the context (anaphora/cataphora, textual ellipsis) or whether it is only implied by common knowledge.

2.4. Machine Translation via Tectogrammatical Layer

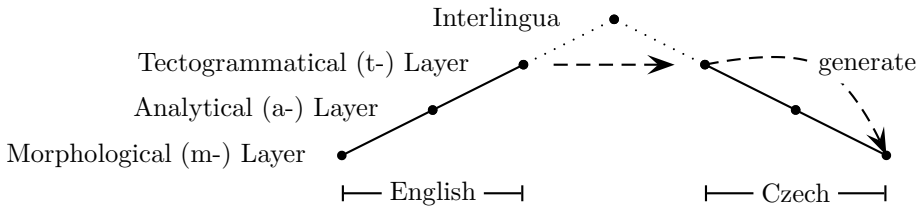


Figure 1. MT via tectogrammatical annotation.

Figure 1 illustrates an overall scheme of our MT system. The rationale to introduce additional layers of formal language description is to bring the source and target language closer to each other (see Figure 2). If the layers are designed appropriately, the transfer step will be easier to implement because (among others):

- t-structures exhibit less divergences, fewer structural changes will be needed in the transfer step.
- t-nodes correspond to autosemantic words only, all auxiliary words are identified in the source language and generated in the target language using language-dependent grammatical rules between t- and a- layers.
- t-nodes contain word lemmas, the whole morphological complexity of either of the languages is handled between m- and a- layers.
- t-layer abstracts away word-order issues, explicitly encoding topic-focus articulation (given/new) in node order.

3. English Tectogrammatical Layer: Ongoing Work

3.1. Prague Czech-English Dependency Treebank

The tectogrammatical representation remains language-specific in many concrete annotation decisions. Even so, its basic concepts are believed to be applicable to most languages. To prove this assumption, a parallel Czech-English treebank is being built. The Prague Czech-English Dependency Treebank (PCEDT 2.0) is based on PCEDT 1.0 (Cuřín et al., 2004), which comprises the Penn Treebank II - Wall Street Journal section (Marcus et al., 1994) converted into dependency trees on the a-layer, and a corpus of its Czech translations, parsed in the same

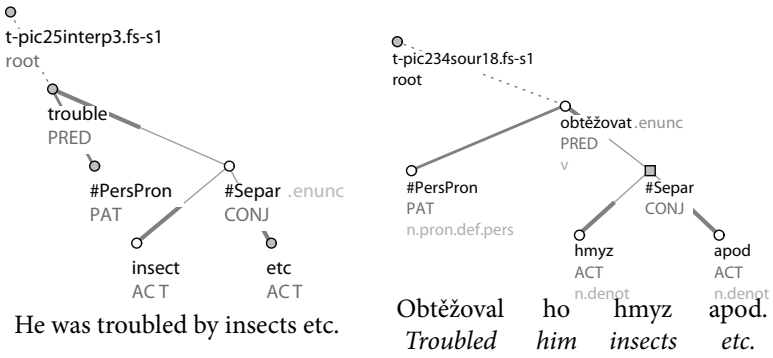


Figure 2. A pair of English and Czech t-trees of the same sentence.

way as PDT 1.0 (Hajič et al., 2001) was. As PDT 2.0 came into existence, the parallel texts were re-parsed to comply with the new format of PDT 2.0, and manual annotation of the automatically pre-processed t-layer trees was launched for both languages.

3.2. Prague English Dependency Treebank

The English counterpart (referred to as the Prague English Dependency Treebank, PEDT) comprises approx. 50 000 dependency trees, which have been obtained by an automatic conversion of the original Penn Treebank II constituency trees into FGD-compliant a-layer trees. These a-layer trees have been automatically converted into t-layer trees. EngVallex (Cinková, 2006), a valency lexicon of verbs contained in PTB-WSJ, was obtained by a semi-automatic conversion of the PropBank-Lexicon (Palmer, Gildea, and Kingsbury, 2005, Palmer et al., 2004) into an FGD-compliant valency lexicon (following the structure of the Czech PDT-Vallex (Hajič et al., 2003)) and its manual adjustment.

3.3. Annotation Manual

Three annotators and a coordinator have been working on the adaptation of the Czech annotation guidelines into English. An annotation manual for the English tectogrammatical representation was released (Cinková et al., 2006)[4]. So far, the annotation has concentrated on the following issues:

1. correct tree structure, including but not limited to:
   (a) rules for coordination, apposition, parenthesis
   (b) some specific constructions like comparison, restriction, consecutive clauses with quantifiers etc.

[4] http://ufal.mff.cuni.cz/~cinkova/TR_En.pdf

(c) determination of function words
2. assigning and completing valency frames in verbs
3. correct semantic labels (functors) in nodes
4. correct t-lemmas
5. correct links to a-layer

The following issues have been left aside for the moment:
1. coreference
2. topic-focus articulation
3. more fine-grained attributes in nodes (subfunctors, grammatemes)

3.4. Annotation Process

Three Czech annotators had first been trained in the Czech annotation and their proficiency in English had been checked before entering the English annotation. The annotation tool TrEd[5], used in the Czech annotation, was adopted to the specific features of the English annotation. Later on, the two configurations were re-unified to make it possible for the annotators to switch languages without having to learn two different ways of annotation with TrEd. This preparatory stage lasted from spring to fall 2006. The actual annotation was launched in September 2006.

The annotators are supposed to deliver 500 trees per month including the test files for agreement measurements, which should ensure about one half of PTB-WSJ to be manually annotated by 2008. Being slightly behind the schedule, we appointed and trained several new annotators. Simultaneously, special attention is being paid to tree pre-processing in order to decrease the extent of the manual annotation work. As the annotation manual has become quite stable now it is possible to formulate additional rules for the conversion of the original constituency trees into tectogrammatical trees, exploiting the rich original linguistic markup of PTB-WSJ in more depth than done so far, e.g. regarding cleft sentences and verb control.

4. Tree-to-tree Transfer

4.1. Synchronous Tree Substitution Grammars

Synchronous Tree Substitution Grammars (STSG) were introduced by Hajič et al., 2002 and later formalized by various authors. The exact definitions we use are summarized in Bojar and Čmejrek, 2007. STSG capture the basic assumption of syntax-based MT that a valid translation of an input sentence can be obtained by local structural changes of the input syntactic tree (and translation of node labels). Some training sentences may violate this assumption because human translators do not always produce literal translations but we are free to ignore such sentences.

As illustrated in Figure 3, STSG describe the tree transformation process using the basic unit of *treelet pair*. Both the source and the target tree are decomposed into treelets that fit

[5] http://ufal.mff.cuni.cz/pdt2.0/doc/tools/tred/

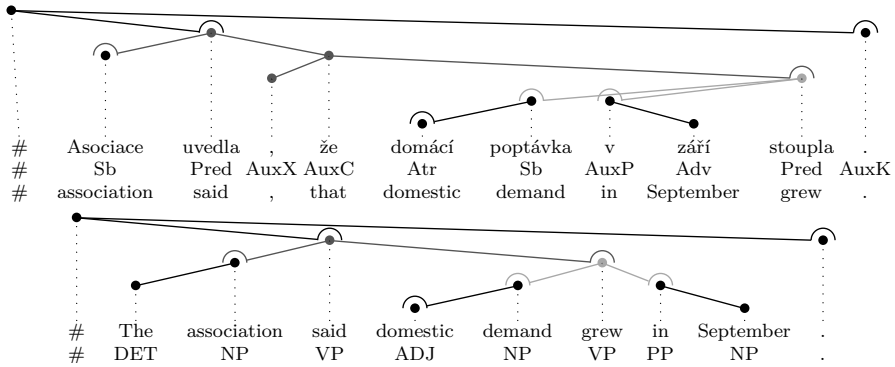


Figure 3. A sample pair of analytical trees synchronously decomposed into treelets.

together. Each treelet can be considered as representing a minimum translation unit. A treelet pair such as depicted in Figure 4 represents the structural and lexical changes necessary to

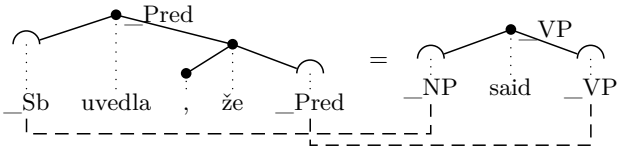


Figure 4. A sample analytical treelet pair.

transfer local context of a source tree into a target tree.

Each node in a treelet is either *internal* (●, constitutes treelet internal structure and carries a lexical item) or *frontier* (◠, represents an open slot for attaching another treelet). Frontier nodes are labelled with *state labels* (such as "_Sb" or "_NP"), as is the root of each treelet. A treelet can be attached at a frontier node only if its root state matches the state of the frontier.

A *treelet pair* (i.e. a *rule* in the synchronous grammar) describes also the *mapping* of the frontier nodes. A pair of treelets is always attached synchronously at a pair of matching frontier nodes.

Depending on our needs, we can encode ordering of nodes as part of each treelet. If only local ordering is used (i.e. we record the position of a parent node among its sons), the output tree will be always projective (see Sgall, Hajičová, and Panevová, 1986, p. 152). If we record global ordering of all nodes in a treelet, the final output tree may contain non-projectivities introduced by non-projective treelets (the attaching operation itself is assumed to be projective).

STSG is generic enough to be employed at or across various layers of annotation (e.g. En-

glish t-tree to Czech t-tree or English a-tree to Czech a-tree). Our primary goal is to perform transfer at the tectogrammatical layer.

4.2. STSG Decoder

The task of STSG "decoder" is to find the most likely target tree, given a source tree and a dictionary of treelet pairs.

Our current version of the decoder considers all possible decompositions of input tree. We traverse the input tree top-down, using the dictionary of treelet pairs to produce the output tree by attaching corresponding right hand treelets to open frontiers. Another option is to traverse the tree in bottom-up fashion in a parsing-like algorithm, as sketched in Čmejrek, 2006.

The research prototype of the transfer system can be obtained at the following URL: `http://ufal.mff.cuni.cz/euromatrix/`

4.3. Estimating STSG Model Parameters

Bojar and Čmejrek, 2007 provides formal details and expectation-maximization algorithms for training STSG using a parallel treebank. Considerations and experiments with this alignment system reported in Bojar, Janı̈ček, and Týnovský, 2008 unfortunately reveal serious problems with scalability of the system to moderately sized parallel corpora. If all possible decompositions of trees to treelets are taken into account, the lexicon of extracted treelet pairs is too big to fit in memory. If the extracted rules are heuristically pruned based on word-to-word alignment, memory requirements significantly decrease but so does the coverage of the rules: many tree pairs in the training data become unreachable using the lexicon of treelet pairs that survived the pruning. A plausible balance between the detail and coverage of the treelet pairs is still to be searched for. Ultimately, we may need to resort to a two-phase approach of preliminary alignment using very coarse-grained information from the trees (to avoid excessive number of distinct treelets in the beginning) followed by the selection of a single best choice from the set of preliminary alignments using more details from the data.

For the time being we restrict our training method to a heuristic based on GIZA++ (Och and Ney, 2000) word alignments. For each tree pair in the training data, we first read off the sequence of node labels and use GIZA++ tool to extract a possibly N-N node-to-node-alignment. Then we extract all treelet pairs from each aligned tree pair such that all the following conditions are satisfied:

- each treelet may contain at most 5 internal and at most 7 frontier nodes (the limits are fairly arbitrary),
- each internal node of each treelet, if aligned at all, must be aligned to a node in the other treelet,
- the mapping of frontier nodes has to be a subset of the node-alignment,
- each treelet must satisfy STSG property: if a node in the source tree is used as an internal node of the treelet, all immediate dependents of the node have to be included in the treelet as well (either as frontier or internal nodes). In other words, we assume no tree

adjuction operation was necessary to construct the training sentence.

All extracted treelet pairs and basic co-occurrence statistics constitute our "translation table".

4.4. Methods of Back-off

As expected, and also pointed out by Čmejrek, 2006, the additional structural information boosts data-sparseness problem. Many source treelets in the test corpus were never seen in our training data. To tackle the problem, our decoder utilizes a sequence of back-off models, i.e. a sequence of several translation tables where each subsequent table is based on less fine-grained description of the input tree.

Given a source treelet, we first search an "exact-match" translation table. If no translation candidate can be found, we disregard some of the detailed node attributes (such as verbal tense etc.) in the source treelet and search for a correspondingly reduced translation table. We also experiment with an alternative direction of source treelet simplification: we keep the full detail of internal nodes but remove all frontier nodes. When a target treelet is found (with no frontier nodes, because the source treelet we searched for had no frontier nodes either), we insert the original number of frontier nodes on the fly, guessing both their position in the treelet and their label using simple local statistics. As a last resort back-off, we keep the internal nodes in the source treelet untranslated and just guess target-side labels of all frontiers. The order and level of detail of the back-off methods is fixed but easily customizable in a configuration file.

4.5. Generating Surface from Czech Tectogrammatical Trees

The purpose of the generation component is to express the meaning given by the target t-tree in a sentence of the target language. In the terms of Figure 1, our objective is the transition given by the right side of the translation triangle.

We decompose the generation into a sequence of seven linguistically motivated steps: Formeme Selection, Agreement, Adding Functional Words (prepositions, subordinating conjunctions and other auxiliaries), Inflexion, Word Order, Punctuation and Vocalization. During each step the input t-tree is gradually changing - new node attributes and/or new nodes are added. After the last step, the nodes are ordered appropriately and each node bears a computed word form. The resulting sentence is then simply obtained by concatenation.

The Formeme Selection phase is where the syntactic shape of the final sentence is grounded. The input t-tree is traversed in depth-first fashion and a suitable morphosyntactic (surface) form is selected for each node. From the full repertoire of surface forms available in Czech language, a subset was selected and is implemented in the generator. Surface forms are identified in the system by a distinguishable label, which we call *formeme*. The formeme is stored as an attribute of a t-node once particular surface realization is picked out. Possible formeme values are for instance: simple case *gen* (genitive case), prepositional case *pod+7* (preposition *pod/under* and instrumental case), *adj* (syntactic adjective), *že+v-fin* (subordinating clause introduced with subordinating conjunction *že*), etc.

Surface forms suitable for a particular t-node are restricted both by syntax and semantics. The syntactic nature is given by the governor's and its own part of speech. As far as semantics is concerned, a particular choice of meaning-bearing preposition or subordinate conjunction is determined by an attribute of t-node called functor. Additional constraints can also be specified in a valency frame of t-node's governor; the frame is picked up from a valency dictionary. The six remaining steps of generation procedure materialize the syntactic and morphological aspects prescribed by the formeme.

Computation of word forms is accomplished using morphological tools characterized by Hajič, 2004. Vocalization rules specifying whether to append a vowel *-e/-u* to selected prepositions are based on Petkevič, 1995. A detailed description of the generation component is given in (Ptáček and Žabokrtský, 2006).

5. Experimental Results

Table 1 reports the BLEU (Papineni et al., 2002) scores of several configurations of our system. For the purposes of comparison with a phrase based system tuned for English-to-Czech, we train and test our system on the News Commentary corpus as available for the ACL 2007 workshop on machine translation (WMT)[6]. We use BLEU to compare the lowercased output of the system to a single lowercased reference translation.[7]

The values in column Generation indicate how strongly is the final production of string of words driven by an n-gram language model (LM). For phrase-based approaches, LM is a vital component. For our transfer to Czech a-layer, our decoder uses LM to score partial trees when enough consecutive internal nodes have been established. The generation component described in Section 4.5 employs no LM and has no access to the target side of the training corpus.

5.1. Discussion and Future Research

At the first sight, our preliminary results support common worries that with a more complex system it is increasingly difficult to obtain good results. However, we are well aware of many limitations of our current experiments:

1. BLEU is known to favour methods employing n-gram based language models (LMs). In future experiments we plan to attempt both, employing some LM-based rescoring when generating from the t-layer, as well as using other automatic metrics of MT quality.
2. All components in our setup deliver only the single best candidate. Any errors will therefore accumulate over the whole pipeline. In the future, we would like to pass and accept several candidates, allowing each step in the calculation to do any necessary rescoring.
3. The rule-based generation system has been designed to generate from full-featured manual Czech tectogrammatical trees from the (monolingual) PDT. There are so far no man-

[6] `http://www.statmt.org/wmt07/`

[7] For methods using the generation system as described in section 4.5, we tokenize the hypothesis and the reference using the rules from the official NIST `mteval-v11b.pl` script. For methods that directly produce sequence of output tokens, we stick to the original tokenization.

Transfer Mode	Generation	Dev	DevTest
English t → Czech t preserving structure	rule-based	5.38±0.43	5.12±0.49
English t → Czech t changing structure	rule-based	5.14±0.43	4.74±0.46
English t → Czech a	LM-guided	7.01±0.50	6.27±0.56
English a → Czech t	rule-based	3.21±0.37	3.18±0.35
English a → Czech a	LM-guided	9.88±0.58	8.61±0.57
Phrase-based as reported by Bojar, 2007			
Vanilla	LM-driven	-	12.9±0.6
Factored to improve target morphology	LM-driven	-	14.2±0.7

Table 1. Preliminary English-to-Czech BLEU scores for syntax-based MT evaluated on Dev and DevTest datasets of ACL 2007 WMT shared task.

ual Czech trees for a parallel corpus. Our target-side training trees are the result of an automatic analytical and tectogrammatical parsing procedure as implemented by McDonald et al., 2005 and Klimeš, 2006, resp. The errors in automatic target-side training trees, together with errors in the tree-to-tree transfer process, pose new challenges to the generation system. A more thorough analysis of which component causes most frequent errors still has to be done.

4. For the purposes of source-side English analysis, we still rely on simple rules similar to those used by Čmejrek, Cuřı⊠n, and Havelka, 2003 to convert Collins, 1996 parse trees to analytical and tectogrammatical dependency trees. We hope to improve the English-side pipeline soon, using recent parsers and improved tectogrammatical analysis, based on the PEDT manual t-trees described above.

Surprisingly, preserving the structure of English t-tree achieves (insignificantly) better BLEU score than allowing the decoder to use larger treelets to produce structurally different Czech t-trees. One possible explanation is that our current heuristic tree-alignment method performs poorly for t-trees. For all other modes of transfer (t→a, a→t, a→a), tree structure modifications gain significant improvements and we use them.

6. Conclusion

We have described the current status of our ongoing effort to translate from English to Czech via deep syntactic (tectogrammatical) structure. The process involves adaptation of the tectogrammatical layer definition for English, parallel treebank annotation and automatic procedures of source sentence analysis, tree-based transfer and target sentence generation.

Our first empirical results do not reach the phrase-based benchmark and we give several reasons why this is the case. However, the presented system is a finished pipeline that establishes a baseline and makes it possible to evaluate how modifications to individual components influence the end-to-end performance in syntax-based machine translation.

7. Acknowledgment

The work on this project was partially supported by the grants FP6-IST-5-034291-STP (EuroMatrix), GA405/06/0589, 1ET101120503, 1ET201120505, and GAUK 7643/2007. We would like to thank Zdeněk Žabokrtský for his rules performing automatic annotation of English t-layer.

Bibliography

Bojar, Ondřej and Martin Čmejrek. 2007. Mathematical Model of Tree Transformations. Project Euromatrix - Deliverable 3.2, ÚFAL, Charles University.

Bojar, Ondřej, Miroslav Janı̈̆ček, and Miroslav Týnovský. 2008. Implementation of Tree Transfer System. Project Euromatrix - Deliverable 3.3, ÚFAL, Charles University.

Bojar, Ondřej. 2007. English-to-Czech Factored Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 232–239, Prague, Czech Republic, June. Association for Computational Linguistics.

Chiang, David. 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Cinková, Silvie. 2006. From PropBank to EngValLex: Adapting the PropBank-Lexicon to the Valency Theory of the Functional Generative Description. In *Proc. of LREC*, pages 2170–2175.

Cinková, Silvie, Jan Hajič, Marie Mikulová, Lucie Mladová, Anja Nedolužko, Petr Pajas, Jarmila Panevová, Jiřı̈̆ Semecký, Jana Šindlerová, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2006. Annotation of English on the tectogrammatical level. Technical report, ÚFAL MFF UK.

Čmejrek, Martin. 2006. *Using Dependency Tree Structure for Czech-English Machine Translation.* Ph.D. thesis, ÚFAL, MFF UK, Prague, Czech Republic.

Čmejrek, Martin, Jan Cuřı̈̆n, and Jiřı̈̆ Havelka. 2003. Czech-English Dependency-based Machine Translation. In *EACL 2003 Proceedings of the Conference*, pages 83–90. Association for Computational Linguistics, April.

Collins, Michael. 1996. A New Statistical Parser Based on Bigram Lexical Dependencies. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 184–191.

Cuřín, Jan, Martin Čmejrek, Jiří Havelka, Jan Hajič, Vladislav Kuboň, and Zdeněk Žabokrtský. 2004. Prague Czech-English Dependency Treebank Version 1.0. LDC2004T25, ISBN: 1-58563-321-6.

Hajič, Jan. 2004. *Disambiguation of Rich Inflection (Computational Morphology of Czech).* Nakladatelství Karolinum, Prague.

Hajič, Jan, Eva Hajičová, Petr Pajas, Jarmila Panevová, Petr Sgall, and Barbora Vidová Hladká. 2001. Prague Dependency Treebank 1.0. LDC2001T10, ISBN: 1-58563-212-0.

Hajič, Jan, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová-Řezníčková, and Petr Pajas. 2003. PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In J. Nivre and E. Hinrichs, editors, *Proc. of The Second Workshop on Treebanks and Linguistic Theories*, pages 57–68, Vaxjo, Sweden. Vaxjo University Press.

Hajič, Jan, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, and Magda Ševčíková Razímová. 2006. Prague Dependency Treebank 2.0. LDC2006T01, ISBN: 1-58563-370-4.

Hajič, Jan, Martin Čmejrek, Bonnie Dorr, Yuan Ding, Jason Eisner, Daniel Gildea, Terry Koo, Kristen Parton, Gerald Penn, Dragomir Radev, and Owen Rambow. 2002. Natural Language Generation in the Context of Machine Translation. Technical report, Johns Hopkins University, Center for Speech and Language Processing. NLP WS'02 Final Report.

Klimeš, Václav. 2006. *Analytical and Tectogrammatical Analysis of a Natural Language*. Ph.D. thesis, ÚFAL, MFF UK, Prague, Czech Republic.

Marcus, M., G. Kim, M. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. 1994. The Penn treebank: Annotating predicate argument structure. In *Proc. of ARPA Human Language Technology Workshop*.

McDonald, Ryan, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-Projective Dependency Parsing using Spanning Tree Algorithms. In *Proceedings of HLT/EMNLP 2005*, October.

Och, Franz Josef and Hermann Ney. 2000. A Comparison of Alignment Models for Statistical Machine Translation. In *Proceedings of the 17th conference on Computational linguistics*, pages 1086–1090. Association for Computational Linguistics.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Palmer, Martha, Paul Kingsbury, Olga Babko-Malaya, Scott Cotton, and Benjamin Snyder. 2004. Proposition Bank I. LDC2004T14, ISBN: 1-58563-304-6, Sep 01.

Panevová, Jarmila. 1974. On verbal frames in Functional Generative Description I. *Prague Bulletin of Mathematical Linguistics*, 22:3–40.

Panevová, Jarmila. 1975. On verbal frames in Functional Generative Description II. *Prague Bulletin of Mathematical Linguistics*, 23:17–52.

Panevová, Jarmila. 1980. *Formy a funkce ve stavbě české věty [Forms and functions in the structure of the Czech sentence]*. Academia, Prague, Czech Republic.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL 2002, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania.

Petkevič, Vladimı⊠r. 1995. A New Formal Specification of Underlying Representations. *Theoretical Linguistics*, 21:7–61.

Ptáček, Jan and Zdeněk Žabokrtský. 2006. Synthesis of Czech Sentences from Tectogrammatical Trees. In *Proc. of TSD*, pages 221–228.

Quirk, Christopher and Arul Menezes. 2006. Dependency Treelet Translation: The Convergence of Statistical and Example-Based Machine-Translation? *Machine Translation*, 20(1):43–65.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands.

Zollmann, Andreas and Ashish Venugopal. 2006. Syntax Augmented Machine Translation via Chart Parsing. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 138–141, New York City, June. Association for Computational Linguistics.

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 90 DECEMBER 2008 69–82

Semantic Network Manual Annotation and its Evaluation

Václav Novák

Abstract

The present contribution is a brief extract of (Novák, 2008). The Prague Dependency Treebank (PDT) is a valuable resource of linguistic information annotated on several layers. These layers range from morphemic to deep and they should contain all the linguistic information about the text. The natural extension is to add a semantic layer suitable as a knowledge base for tasks like question answering, information extraction etc. In this paper I set up criteria for this representation, explore the possible formalisms for this task and discuss their properties. One of them, Multilayered Extended Semantic Networks (MultiNet), is chosen for further investigation. Its properties are described and an annotation process set up. I discuss some practical modifications of MultiNet for the purpose of manual annotation. MultiNet elements are compared to the elements of the deep linguistic layer of PDT. The tools and problems of the annotation process are presented and initial annotation data evaluated.

1. Motivation

The longterm goal of the research in the field of Artificial Intelligence has been to create a machine which would understand natural language input and be able to perform the reasoning necessary to perform the desired actions. It is obvious that such a machine must be capable of storing the acquired information in its memory in a form suitable for the necessary reasoning. We will call this form the *knowledge representation*. Let's discuss the criteria which should be imposed upon the form of the information representation, and the existing systems for knowledge representation and their properties with respect to the given criteria.

There are several reasons why Tectogrammatical Representation (TR) may not be sufficient in a question answering system or machine translation:

1. There is no information about sorts of concepts represented by TR nodes. Sorts (the upper conceptual ontology) are an important source of constraints for semantic relations. Every relation has its signature which in turn reduces ambiguity in the process of text analysis and inferencing.

Please cite this article as: Václav Novák, Semantic Network Manual Annotation and its Evaluation. The Prague Bulletin of Mathematical Linguistics No. 90, 2008, 69–82.

2. The syntactic functors Actor and Patient disallow creating inference rules for cognitive roles like *Affected object* or *State carrier*. For example, the axiom stating that an affected object is changed by the event cannot be feasibly expressed in the TR framework. However, if needed, this information can be stored in the lexicon for individual verb frames.
3. Lexemes of TR have no hierarchy; this limits especially the search for an answer in a question answering system. In TR there is no counterpart of SUB, SUBR, and SUBS MultiNet relations, which connect subordinate concepts to superordinate ones and individual object representatives to corresponding generic concepts.
4. In TR, each sentence is isolated from the rest of the text, except for coreference arrows connected to preceding sentences. This, in effect, complicates inferences combining knowledge from multiple sentences in one inference rule.
5. Nodes in TR always correspond to a word or a group of words in the surface form of a sentence or to a structure which is deleted on the surface (e.g., obligatory verb argument, coordination member). There are no means for representing knowledge generated during the inference process, if the knowledge does not have the form of a TR. For example, consider the axiom of temporal precedence transitivity (1):

$$(a \text{ ANTE } b) \wedge (b \text{ ANTE } c) \rightarrow (a \text{ ANTE } c) \tag{1}$$

   In TR, we cannot add an edge denoting $(a$ ANTE $c)$. We would have to include a proposition like "a precedes c" as a whole new clause.

For all these reasons we need to extend our text annotation to a form suitable to more advanced tasks. It is shown in (Helbig, 2006) that MultiNet is capable of solving all the above mentioned issues.

2. Criteria

In order to efficiently retrieve and process the knowledge acquired in the form of natural language input, these criteria should be fulfilled by the internal knowledge representation format:[1]

I. **Associativity**: The knowledge concerning a concept should be available without the necessity to iterate over the whole knowledge base. A representation lacking this property would not be scalable to real problems.

II. **Local interpretability**: The knowledge necessary for interpretation of an object should be limited to an easily identifiable local neighborhood of the concept (the knowledge may include a contextual embedding which is crucial for the concept interpretation).

III. **Inference friendliness**: The knowledge data format should allow for further inclusion of new facts, acquired both by new texts and by automatic inferencing. A practical system should be robust with respect to contradictions to avoid a situation where every proposition is true.

[1] Criteria II., A., B. and C. are modifications of some of the criteria imposed by (Helbig, 2006). I formulated criteria I. and III.

Apart from the overall necessary requirements, there are also further criteria necessary for a representation if it is to be annotated manually:

A. **Consistency**: Analogous facts should be treated analogously.
B. **Cognitive Adequacy**: The representation must be understandable to the annotators and easy to visualize and review.
C. **Communicability**: The instructions should contain applicable operational criteria (Hajičová and Sgall, 1980), definitions, and standards.

The next requirement for the representational formalism is to integrate smoothly into the layered nature of the PDT (Karcevskij, 1929, Callmeier et al., 2004).

Why are these requirements crucial?

Without associativity (I.), the query for information would always require a search through the whole knowledge base. Furthermore, for queries which cannot be answered using only one sentence, one would have to create a kind of associative structure on the fly to make use of disambiguation, coreferences etc.

Local interpretability (II.) is needed for concepts embedded in a way that changes their mode of existence. Consider the clause "*If I were you*". We do not want to extract the information that *I* refers to the same person as *you*. However, this is what we would infer if we ignored the contextual embedding associated with the word *if*. Therefore the knowledge representation must ensure this information is readily available for every piece of information without the need to iterate through the whole knowledge base.

Inference friendliness (III.) allows us to enrich the acquired knowledge by applying inference rules. If we know that "*Mrs. Hill is the current vice president finance*", we can infer for instance that "*The current vice president finance is Mrs. Hill*". An inference friendly representation will allow a compact representation of such an inference. Without this compactness (e.g., in the case where the inference must be included as a whole new sentence) the scale of practical inferences would be very limited.

Without consistency (A.) the annotation process is unimaginable, because annotators are able to use only a limited set of instructions and they always treat the new sentences by analogy. If this were not the correct way to annotate, they could not produce meaningful results.

Cognitive adequacy (B.) is practical when the annotators must deal with complicated sentences. There are few people who understand modal operators and first order logic axioms, but there are many people who understand the sentences in *The Wall Street Journal*. Ideally, the complexity of annotating a sentence should be 100% correlated with the complexity of understanding its meaning. Without cognitive adequacy of the representation, the annotation cannot leave the realm of toy sentences.

Communicability (C.) is another key to the success of annotation. A mere learning by example can prove to be useful, but it fails in the case of border cases. Unfortunately, however contradictory this may sound, border cases make up a significant percentage of decisions and can be found in every Wall Street Journal sentence.

3. Existing Meaning Representations

In this section we will discuss various formalisms of knowledge representation and their conformance to the criteria presented in Section 2.

3.1. Representations Based on First Order Logic

The first attempts to formalize natural language were made using the predicate calculus (Frege, 1892). Since then various approaches have been trying to fix the problems of using first order logic purely extensional interpretation of the meaning. First, intensional semantics was developed (Montague, 1972) to introduce the notion of conceivable worlds. This theory was further developed in several directions:

- TIL: Transparent Intensional Logic (Tichý, 1988) aimed at further elaboration of the semantics of conceivable worlds
- Description Logic (Donini et al., 1996) focused on the computational aspects of meaning representation.
- DRT: Discourse Representation Theory (Eijk and Kamp, 1996) focused on the treatment of coreferences, quantifiers, and their interplay.
- Hybrid Modal Logic (Areces and Blackburn, 2001, Areces, Blackburn, and Marx, 2004, Blackburn, 2000, Blackburn, 2001) applied the framework of modal logic to natural language semantics.

All these formalisms have been used to represent real-life sentences. There has been a successful attempt to automatically create DRT structures proposed in (Bos, 2005). Hybrid Modal Logic has been investigated from the linguistic viewpoint in (Kruijff, 2001, Novák, 2004, Novák and Hajič, 2006). The TIL has been subject to automatic transduction (Horák, 2001), but not to manual annotation.

How do these systems fit into our criteria? They are very strong in associativity (I.): every concept is represented by one or more variables and these variables can be looked up easily. Inference friendliness (III.) is guaranteed as to the ease of addition of new knowledge: it can be added by simply adding predicates. On the other hand the robustness with respect to contradictions is addressed only in some of these systems and in general requires non-monotonicity of the reasoning.

Local interpretability (II.) is addressed only in DRT, where the relevant contextual embedding should be present only in the current box. Cognitive adequacy (B.) is the most difficult obstacle which prevents these systems from being manually annotated. The model-theoretic way of thinking and use of quantifiers are largely unintuitive. This is not apparent for sentences which are usually addressed in the relevant literature (e.g., "*Every farmer owns a donkey*"). Nevertheless, it emerges when we try to come up with a predicate calculus representation of an ordinary sentence like "*The U.S. trade representative, Carla Hills, announced …*" It seems unintuitive to think about *trade* as a function from possible worlds to a set of objects, which is the typical treatment for nouns.

3.2. Representations Based on Linguistic Structures

The meaning representations based on linguistic structures emerged as an extension of dependency syntax (Tesnière, 1934). There are various formalisms, which all share some common features: they start with the text or speech and transform it into formalized layers of representation, where the last layer should be the most suitable for the knowledge representation tasks. They are:

- Functional Generative Description (Sgall, Hajičová, and Panevová, 1986), where the highest layer of description is the Tectogrammatical Representation (Hajičová, Panevová, and Sgall, 2000)
- Robust Minimal Recursion Semantics (Copestake et al., 2005) as a pluggable layer of the framework of (Callmeier et al., 2004)
- Meaning-Text Theory (Mel'čuk, 1988, Bolshakov and Gelbukh, 2000), which is in many respects similar to the FGD framework (Žabokrtský, 2005).

These approaches have difficulties with respect to the inference friendliness (III.): to include a piece of inferred knowledge, we often have to add a whole new sentence which describes the fact. For example if we are to apply a rule stating the symmetry of a predicate in a logic-based system, we simply add one predicative statement for every instance. In a linguistics-based system, we have to copy the whole statement and transform it into the inverse form.

The next obstacle concerns the cognitive adequacy: the tree constraints force the annotators to choose only one connection where more of them could be applied: in "*They met during the concert on Tuesday.*" the above mentioned systems require the annotator to decide whether *on Tuesday* is connected to *met* or *concert*, although from the knowledge base viewpoint it would be ideal if both *met* and *concert* were connected with the temporal specification under consideration.

3.3. Semantic Networks

Semantic networks, as different from the logic-based systems as they may seem, have much in common with them. The semantic network, being a directed graph, can usually be turned into a set of formulae of predicate calculus. The main difference lies in the fact that the relationship between the predicates and the knowledge is not direct: the predicates encode information about the network. The elements of the network then carry their own meaning.

The main advantage of semantic networks is their concept-centeredness. As noted on page 4 of (Helbig, 2006), the difference is similar to the difference between a logical programming language (e.g., Prolog) and an object oriented programming language (e.g., Java). Every concept should correspond to a cognitive concept and it is assumed that two distinct concepts do not represent the same object, unless there is a piece of information indicating the opposite. On the other hand, in a model-theoretic framework, the model builders tend to create a model as small as possible, therefore collapsing the referents of all variables where possible. This, in effect, often leads to a wrong conclusion.

Individual semantic network formalisms differ in their repertoire of formal means. In prac-

tice, two systems have been used for purposes of natural language processing:

- KL-ONE: knowledge representation system (Brachman and Schmolze, 1985)
- MultiNet: Multilayered Extended Semantic Networks (Helbig, 2006)

They satisfy all the criteria presented in Section 2 and therefore they are discussed in the remaining chapters.

3.4. Semantic Web

A Semantic web is sometimes considered yet another semantic representation. However, it is more a framework allowing us to standardize the representations and exchange the data in a structured format. It is therefore not possible to simply create a semantic web corpus. The technologies being used are the Web Ontology Language (Horrocks and Patel-Schneider, 2004), which allows for standardization and exchange of ontologies, and Resource Description Framework (RDF Core Working Group, 2007), which is an XML-based data format for exchanging predicate-like structures.

4. Evaluation Metrics

Human annotations are usually evaluated against each other to measure the consistency of the annotation. The most common measures of agreement are accuracy (number of correct decisions divided by the number of all decisions) and F-measure (harmonic mean between the recall and the precision). However, these approaches suffer from the fact that some annotation agreement is present simply by chance. This fact was the reason to propose annotation agreement metrics corrected for the agreement by chance. First, Scott's π (Scott, 1955) and Cohen's κ (Cohen, 1960) were introduced. They were later generalized to the K coefficient of agreement (Carletta, 1996).

I do not use any of these corrections for three reasons:

1. The agreement metrics itself is difficult to develop and to obtain the most appropriate agreement score there is still much to do.
2. The agreement by chance is difficult to compute in such a complex situation. The probability that two annotators will produce exactly the same oriented graph with the same size and all the attributes is virtually zero.
3. The measures have no clear probabilistic interpretation (Artstein and Poesio, 2007).

When a stable level of annotator agreement is achieved and maintained, and the agreement measure is robust with respect to equivalent annotations, the metrics extended for hierarchical values should be used. An example is Krippendorf's α (Krippendorff, 1980).

5. Evaluation Data

The initial evaluation presented in this section has been carried out on a portion of *The Wall Street Journal* articles from the Penn Treebank (Marcus, Marcinkiewicz, and Santorini, 1993), which have been annotated on all the FGD layers and are available as the Prague English Dependency Treebank (Hajič et al., est. 2009). Initially, some sentences were used during the

training of annotators. These sentences were removed from the evaluation sample. The evaluation sample contains 67 annotated sentences (1793 words), annotated by two annotators, of which 46 sentences (1236 words) were annotated by three independent annotators. All annotators are native English speakers.

6. Structural Agreement

The structural agreement is measured for every sentence in isolation in two steps. First, the best match between the two annotators' graphs is found. Most of the graph nodes are connected to the tectogrammatical tree and for the remaining nodes, all possible one-to-one mappings are constructed and the optimal mapping w.r.t. the F-measure is selected. Second, the optimal mapping is used to compute the agreement.

Formally, we start with a set of tectogrammatical trees containing a set of nodes N. The annotation is a tuple $G = (V, E, T, A)$, where V are the vertices, $E \subseteq V \times V \times P$ are the directed edges and their labels (e.g., agent of an action: AGT $\in P$), $T \subseteq V \times N$ is the mapping from vertices to the tectogrammatical nodes, and finally A are attributes of the nodes. We simplified the problem by ignoring the mapping from edges to tectogrammatical nodes, the metaedges, and the MultiNet edge attribute *knowledge type*. Analogously, $G' = (V', E', T', A')$ is another annotation of the same sentence and our goal is to measure the similarity $s(G, G') \in [0, 1]$ of G and G'.

To measure the similarity we need a set Φ of admissible one to one mappings between vertices in the two annotations. A mapping is admissible if it connects vertices which are indicated by the annotators as representing the same tectogrammatical node:

$$\begin{aligned} \Phi \quad = \quad & \left\{ \phi \subseteq V \times V' \,\middle|\, \right. \\ & \underset{\substack{n \in N \\ v \in V \\ v' \in V'}}{\forall} \Big(\big((v, n) \in T \wedge (v', n) \in T' \big) \rightarrow (v, v') \in \phi \Big) \\ \wedge \quad & \underset{\substack{v \in V \\ v', w' \in V'}}{\forall} \Big(\big((v, v') \in \phi \wedge (v, w') \in \phi \big) \rightarrow (v' = w') \Big) \\ \wedge \quad & \left. \underset{\substack{v, w \in V \\ v' \in V'}}{\forall} \Big(\big((v, v') \in \phi \wedge (w, v') \in \phi \big) \rightarrow (v = w) \Big) \right\} \end{aligned} \tag{2}$$

In Equation 2, the first condition ensures that Φ is constrained by the mapping induced by the links to the tectogrammatical layer. The remaining two conditions guarantee that Φ is a one-to-one mapping.

Then we can define the annotation agreement s as

$$s_{(G,G',m)} = F_m(G, G', \phi^*) \tag{3}$$

where ϕ^* is the optimal mapping between nodes of alternative annotations:

$$\phi^* = \text{argmax}_{\phi \in \Phi}(F_m(G, G', \phi)) \tag{4}$$

and F_m is the F1-measure:

$$F_m(G, G', \phi) = \frac{2 \cdot m(\phi)}{|E| + |E'|} \tag{5}$$

where $m(\phi)$ is the number of edges that match given the mapping ϕ. We use four variants of m, which gives us four variants of F and consequently four scores for every sentence:
Directed unlabeled:

$$m_{du}(\phi) = \left| \left\{ (v, w, \rho) \in E \middle| \exists_{v', w' \in V', \rho' \in P} \Big((v', w', \rho') \in E' \wedge (v, v') \in \phi \wedge (w, w') \in \phi \Big) \right\} \right| \tag{6}$$

Undirected unlabeled:

$$m_{uu}(\phi) = \left| \left\{ (v, w, \rho) \in E \middle| \exists_{v', w' \in V', \rho' \in P} \Big(\big((v', w', \rho') \in E' \vee (w', v', \rho') \in E'\big) \wedge (v, v') \in \phi \wedge (w, w') \in \phi \Big) \right\} \right| \tag{7}$$

Directed labeled:

$$m_{dl}(\phi) = \left| \left\{ (v, w, \rho) \in E \middle| \exists_{v', w' \in V'} \Big((v', w', \rho) \in E' \wedge (v, v') \in \phi \wedge (w, w') \in \phi \Big) \right\} \right| \tag{8}$$

Undirected labeled:

$$m_{ul}(\phi) = \left| \left\{ (v, w, \rho) \in E \middle| \exists_{v', w' \in V'} \Big(\big((v', w', \rho) \in E' \vee (w', v', \rho) \in E'\big) \wedge (v, v') \in \phi \wedge (w, w') \in \phi \Big) \right\} \right| \tag{9}$$

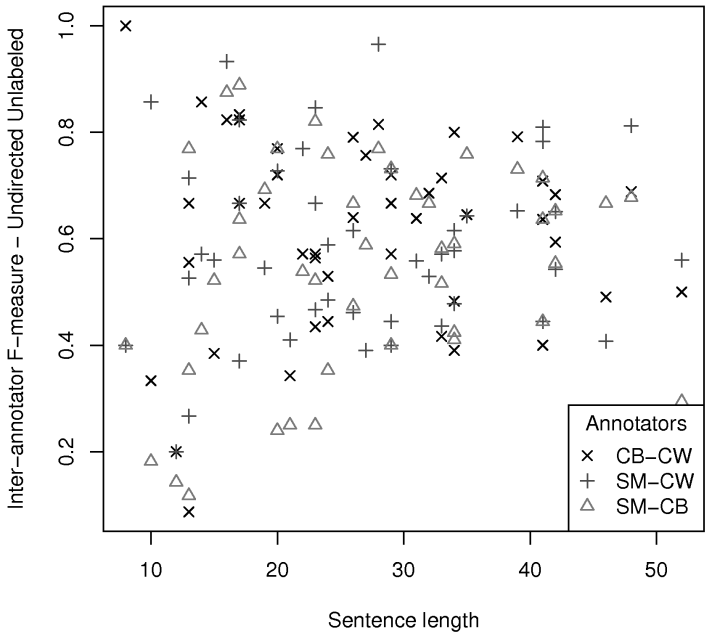


Figure 1. Inter-annotator agreement depending on the sentence length.

These four $m(\phi)$ functions give us four possible F_m measures, which allows us to have four scores for every sentence: s_{du}, s_{uu}, s_{dl} and s_{ul}.

Figure 1 shows that the agreement is not correlated with the sentence length. This means that longer sentences are on average no more difficult than short sentences. The variance decreases with the sentence length as expected.

In Figure 2 I present a comparison of directed and labeled evaluations with the undirected unlabeled case. By definition, the undirected unlabeled score is the upper bound for all the other scores. The directed score is well correlated and not very different from the undirected score, indicating that the annotators did not have much trouble with determining the correct direction of the edges. This might be in part due to support from the formalism and the *cedit* tool: each relation type is specified by a sort signature; a relation that violates its signature is reported immediately to the annotator. On the other hand, labeled score is significantly lower than the unlabeled score, which suggests that the annotators have difficulties in assigning the

Sample	Annotators	Agreement F-measure			
		s_{uu}	s_{du}	s_{ul}	s_{dl}
Smaller	CB-CW	61.0	56.3	37.1	35.0
Smaller	SM-CB	54.9	48.5	27.1	25.7
Smaller	SM-CW	58.5	50.7	31.3	30.2
Smaller	average	58.1	51.8	31.8	30.3
Larger	CB-CW	64.6	59.8	40.1	38.5

Table 1. Inter-annotator agreement in percents. The results come from the two samples described in the Section 5.

correct relation types. The correlation coefficient between s_{uu} and s_{ul} (approx. 0.75) is also much lower than the correlation coefficient between s_{uu} and s_{du} (approx. 0.95).

Figure 3 compares individual annotator pairs. The scores are similar to each other and also have a similar distribution shape.

A more detailed comparison of individual annotator pairs shows that there is a significant positive correlation between scores, i.e., if two annotators can agree on the annotation, the third annotator is also likely to agree, but this correlation is not a very strong one. The actual correlation coefficient varies between 0.34 and 0.56. All the results are summarized in Table 1.

Acknowledgements This work was supported by the Czech Ministry of Education grants LC536 and 0021620838 and by Czech Academy of Sciences grant 1ET201120505.

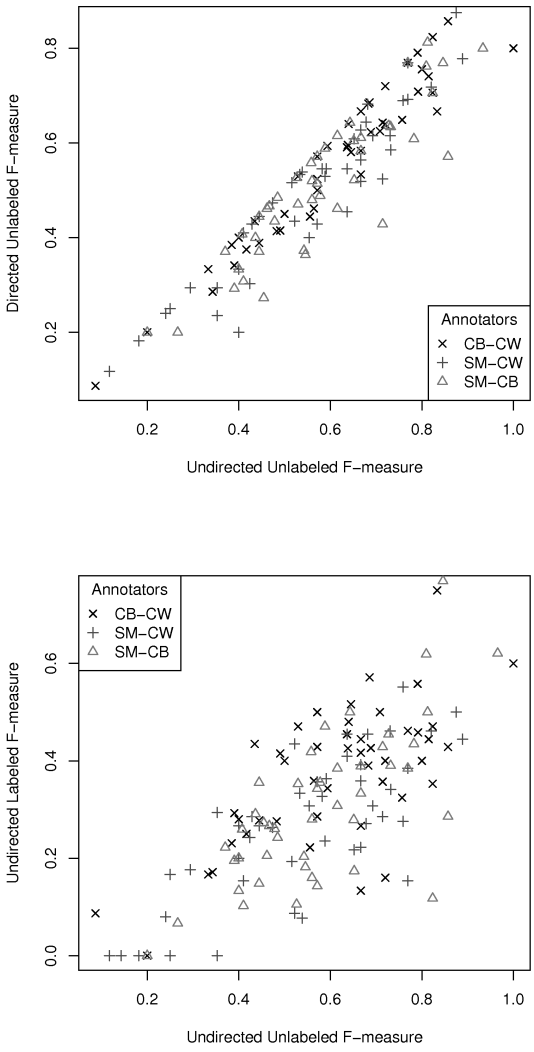


Figure 2. Upper: Directed vs. undirected inter-annotator agreement. Lower: Labeled vs. unlabeled inter-annotator agreement.

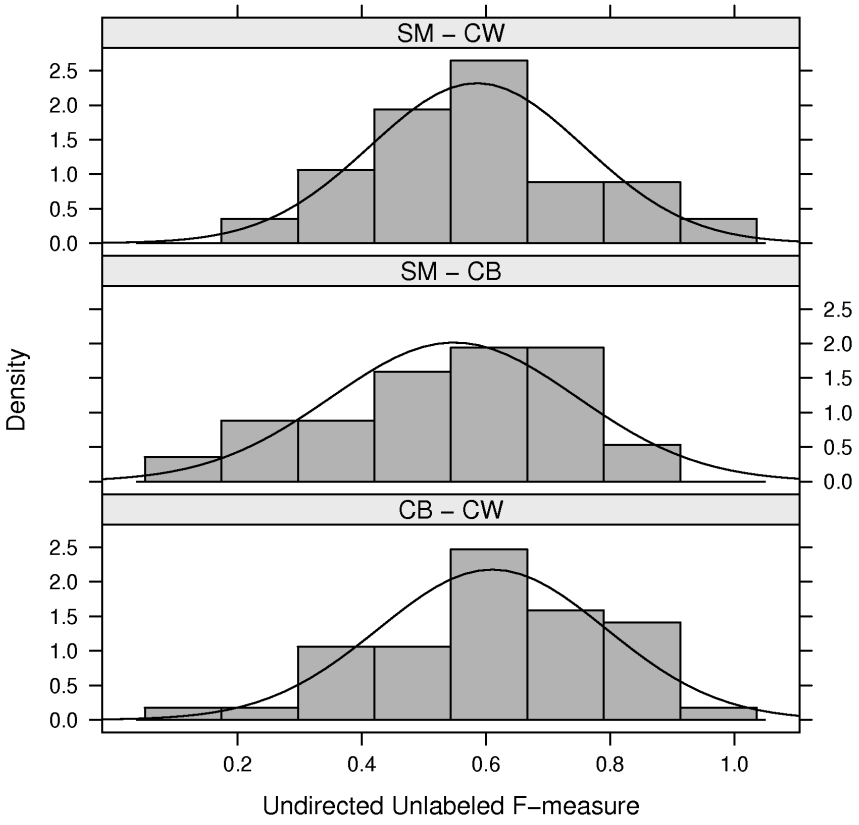


Figure 3. Comparison of individual annotator pairs.

Bibliography

Areces, Carlos and Patrick Blackburn. 2001. Bringing them all Together. *Journal of Logic and Computation*, 11(5):657–669.

Areces, Carlos, Patrick Blackburn, and Maarten Marx. 2004. Hybrid Logics: Characterization, Interpolation and Complexity. To appear in the Journal of Symbolic Logic, http://www.loria.fr/projets/hylo/Papers/jsl.pdf.

Artstein, Ron and Massimo Poesio. 2007. Inter-coder agreement for computational linguistics. *Computational Linguistics*, submitted.

Blackburn, Patrick. 2000. Representation, Reasoning, and Relational Structures: A Hybrid Logic Manifesto. *Logic Journal of the IGPL*, 8(3):339–625.

Blackburn, Patrick. 2001. Modal Logic As Dialogical Logic. *Synthese*, 127(1 - 2):57–93, April.

Bolshakov, Igor and Alexander Gelbukh. 2000. The Meaning-Text Model: Thirty Years After. *International Forum on Information and Documentation*, 1:10–16.

Bos, Johan. 2005. Towards Wide-Coverage Semantic Interpretation. In *Proceedings of Sixth International Workshop on Computational Semantics IWCS-6*, pages 42–53.

Brachman, Ronald and James Schmolze. 1985. An Overview of the KL-ONE Knowledge Representation System. *Cognitive Science*, 9:171–216.

Callmeier, Ulrich, Andreas Eisele, Ulrich Schäfer, and Melanie Siegel. 2004. The DeepThought Core Architecture Framework. In *Proceedings of LREC*, May.

Carletta, Jean. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Copestake, Ann, Dan Flickinger, Carl Pollard, and Ivan Sag. 2005. Minimal Recursion Semantics: An Introduction. *Research on Language and Computation*, 3(4):281–332, December.

Donini, Francesco M., Maurizio Lenzerini, Daniele Nardi, and Andrea Schaerf. 1996. Reasoning in Description Logics. In Gerhard Brewka, editor, *Principles of Knowledge Representation*. CSLI Publications, Stanford, California, pages 191–236.

Eijk, Jan and Hans Kamp. 1996. Representing Discourse in Context. In Johan Benthem and Alice Meulen, editors, *Handbook of Logic and Language*. Elsevier, Amsterdam, pages 179–237.

Frege, Gottlob. 1892. Über Sinn und Bedeutung. In Mark Textor, editor, *Funktion - Begriff - Bedeutung*, volume 4 of *Sammlung Philosophie*. Vandenhoeck & Ruprecht, Göttingen.

Hajič, Jan, Kristýna Čermáková, Lucie Mladová, Anja Nedolužko, Jiří Semecký, Jana Šindlerová, Josef Toman, and Zdeněk Žabokrtský. est. 2009. Prague English Dependency Treebank (in progress).

Hajičová, Eva, Jarmila Panevová, and Petr Sgall. 2000. A Manual for Tectogrammatic Tagging of the Prague Dependency Treebank. Technical Report TR-2000-09, ÚFAL MFF UK, Prague, Czech Republic. in Czech.

Hajičová, Eva and Petr Sgall. 1980. Linguistic meaning and knowledge representation in automatic understanding of natural language. In *Proceedings of the 8th conference on Computational linguistics*, pages 67–75, Morristown, NJ, USA. Association for Computational Linguistics.

Helbig, Hermann. 2006. *Knowledge Representation and the Semantics of Natural Language*. Springer, Berlin, Germany.

Horák, Aleš. 2001. *The Normal Translation Algorithm in Transparent Intensional Logic for Czech*. Ph.D. thesis, Faculty of Informatics, Masaryk University, Brno, Czech Republic.

Horrocks, Ian and Peter F. Patel-Schneider. 2004. Reducing OWL Entailment to Description Logic Satisfiability. *Journal of Web Semantics*, 1(4):345–357.

Karcevskij, Sergei. 1929. Du dualisme asymétrique du signe linguistique. *Travaux du Cercle linguistique de Prague*, 1:88–93.

Krippendorff, Klaus. 1980. *Content analysis: an introduction to its methodology*. Sage Publications, Newbury Park, CA.

Kruijff, Geert Jan. 2001. *A Categorial-Modal Logical Architecture of Informativity*. Ph.D. thesis, Faculty of Mathematics and Physics, Charles University. http://www.coli.uni-sb.de/ gj/dissertation.phtml.

Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19(2):313–330.

Meľčuk, Igor A. 1988. *Dependecy Syntax: Theory and Practice*. State University of New York Press.

Montague, Richard. 1972. Pragmatics and Intensional Logic. *Semantics of Natural Language*.

Novák, Václav. 2004. Towards Logical Representation of Language Structures. *The Prague Bulletin of Mathematical Linguistics*, 82:5–86.

Novák, Václav. 2008. *Semantic Network Manual Annotation and its Evaluation*. Ph.D. thesis, Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic.

Novák, Václav and Jan Hajič. 2006. Perspectives of Turning Prague Dependency Treebank into a Knowledge Base. In *Proceedings of the LREC Conference*, Genova, Italy.

RDF Core Working Group. 2007. Resource Description Framework (http://www.w3.org/RDF/).

Scott, William A. 1955. Reliability of Content Analysis: the Case of Nominal Scale Coding. *Public Opinion Quarterly*, 19:321–325.

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

Tesnière, Lucien. 1934. Comment construire une Syntaxe. *Bulletin de la Faculté des Lettres de Strasbourg*, pages 219–229.

Tichý, Pavel. 1988. *The Foundations of Frege's Logic*. Walter de Gruyter & Co, Berlin/New York.

Žabokrtský, Zdeněk. 2005. Resemblances between Meaning-Text Theory and Functional Generative Description. In *Proceedings of the 2nd International Conference of Meaning-Text Theory*, pages 549–557.

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 90 DECEMBER 2008 83–108

Czech Verbs of Communication with respect to the Types of Dependent Content Clauses

Václava Kettnerová

Abstract

The present paper describes a classification of Czech verbs of communication based on the information on the type of dependent content clauses which these verbs require to be complemented by. We distinguish assertive, interrogative and directive verbs of communication. Furthermore, we propose a method how to treat those verbs of communication which behave 'neutrally' with respect to the type of dependent content clauses.

Introduction

Verbs of communication represent a large group of verbs. They render situations concerning communication in a broad sense: speaking, writing and gestures. Generally, they express situations where a 'Speaker' conveys a 'Message' to a 'Recipient'. Prototypically, the 'Message' may be morphematically realized as a dependent content clause.

In our paper, the information on the type of dependent content clause, which the verbs of communication require to be complemented by, is taken as a key criterion for a classification of these verbs. On this basis, we distinguish assertive, interrogative and directive verbs of communication according to whether they are complemented by assertive, interrogative, or directive dependent content clauses, respectively. The main motivation behind the classification is to create classes of verbs of communication that would be semantically and morphosyntactically more coherent.

The type of the dependent content clauses is determined as a starting point for the classification since the different types of the dependent content clauses are regularly associated with several other morphosyntactic properties of the verbs of communication.

For instance, being complemented by an assertive dependent content clause, the 'Addressee's' valency slot of the verb *říci*^{pf} 'to tell' is optional (ex 1 and ex 2). Furthermore, the splitting of the theme and dictum is allowed in this utterance (ex 2) (Section 3.1.2). On the other hand,

Please cite this article as: Václava Kettnerová, Czech Verbs of Communication with respect to the Types of Dependent Content Clauses. The Prague Bulletin of Mathematical Linguistics No. 90, 2008, 83–108.

the 'Addressee' is semantically obligatory and the splitting of the theme and dictum is not possible if the verb is complemented by an interrogative (ex 3) or a directive dependent content clause (ex 4):

(1) *Řekla, že ji bolí hlava.* (SYN2005)
E. *She said that she had got headache.*

(2) *Řekla nám o sobě, že má upřímnou povahu, je veselá a ráda se baví.* (SYN2006pub)
E. *'Told – us – about – herself – that – has – frank – character – is – cheerful – and – glad – refl – enjoy.'*

(3) *Můžete nám říci, zda počítáte s tím, že v příštím volebním období bude nájemné zcela uvolněno?* (SYN2006pub)
E. *Could you tell us whether you take into account that the rent will not be fixed in the next term of office?*

(4) *Řeknu jí, aby vám napsala a pozvala vás.*(SYN2005)
E. *I will ask her to write to you and invite you.*

The present paper is structured as follows. First, Section 1 describes three above mentioned participants of the verbs of communication, 'Speaker', 'Recipient' and 'Message', with respect to their tectogrammatical counterparts, syntactic behavior and morphemic realizations. A special attention is devoted to the participant 'Message'. Its morphemic realizations are described with regard to its two possible aspects: the theme and the dictum. Second, three types of dependent content clauses – assertive, interrogative and directive – are distinguished on the basis of modality in Section 2.

Section 3 presents the principal issue of this contribution – the classification of the verbs of communication and a description of their morphosyntactic properties. The group of verbs of communication is subdivided into semantically and morphosyntactically more coherent classes – assertive (Section 3.1), interrogative (Section 3.2) and directive verbs of communication (Section 3.3).

In Section 3.4, we propose a method how to treat those verbs of communication which behave 'neutrally' – they exhibit syntactic properties of assertive, interrogative, and directive verbs of communication according to whether they are complemented by an assertive, an interrogative or a directive dependent content clause, as in ex 1–4.

When describing valency, we use the Functional Generative Description (FGD in the sequel) (Sgall, Hajičová, and Panevová, 1986) as the theoretical background. FGD distinguishes between arguments (inner participants, actants) and free modifications (adjuncts). Both types of complementations can be obligatory or optional. First two (verbal) arguments are determined on the basis of syntactic criteria, semantic criteria are considered for the verbs with three or more arguments, see esp. (Panevová, 1974) and (Panevová, 1975). Five types of (verbal) arguments are determined: 'ACTor', 'PATient', 'ADDRessee', 'ORIGin' and 'EFFect', see esp. (Panevová, 1980).

Our description of the verbs of communication is based first of all on the material provided by the valency lexicon of Czech verbs, VALLEX[1], see esp. (Lopatková, Žabokrtský, and Kettnerová, 2008), (Žabortský and Lopatková, 2007) and (Žabokrtský, 2005), and on the material from the Czech National Corpus.[2] Furthermore, we have worked with Czech dictionaries *Slovník spisovného jazyka českého* (SSJČ, 1964) and *Slovník spisovné češtiny pro školu a veřejnost* (SSČ, 2003). The Czech valency dictionaries *Slovesa pro praxi* (Svozilová, Prouzová, and Jirsová, 1997) and *Slovník slovesných, substantivních a adjektivních vazeb a spojení* (Svozilová, Prouzová, and Jirsová, 2005) are taken into account as well.

1. Verbs of communication

Verbs of communication, traditionally called 'verba dicendi', represent a large group of verbs. They involve communication in a broad sense: speaking (e.g., *říci*^{pf}, *říkat*^{impf}'to say', *dodat*^{pf}, *dodávat*^{impf}, 'to add', *volat*^{impf} 'to shout', *ptát se*^{impf} 'to ask', *přikázat*^{pf}, *přikazovat*^{impf} 'to order', etc.), writing (e.g., *napsat*^{pf} 'to write', *zaznamenat*^{pf}, *zaznamenávat*^{impf} 'to record'), and gestures (e.g., *naznačit*^{pf}, *naznačovat*^{impf}, 'to indicate', etc.). They denote such situations where a 'Speaker' conveys a 'Message' to a 'Recipient'.

1.1. 'Speaker', 'Recipient' and 'Message'

In this section, we briefly describe the participants of the verbs of communication – the 'Speaker', 'Recipient' and 'Message' – with respect to their syntactic behavior and morphemic realizations.

1.1.1. 'Speaker' and 'Recipient'

We observe basically two possible realizations of the participant 'Speaker':

- Several verbs of communication represent rather a one-sided communication – the 'Speaker' conveys the 'Message' to the 'Recipient'. In these cases, the semantic participant 'Speaker' occupies the 'Actor's' valency slot and the 'Recipient' the 'Addressee's' one. Examples of such verbs of communication are the following: *doporučit*^{pf}, *doporučovat*^{impf} 'to recommend', *informovat*^{biasp} 'to inform', *lhát*^{impf} 'to lie', *líčit*^{impf} 'to depict', *nahlásit*^{pf}, *nahlašovat*^{impf} 'to report', *nařídit*^{pf}, *nařizovat*^{impf} 'to order', *oznámit*^{pf}, *oznamovat*^{impf} 'to announce', *psát*^{impf} 'to write', *říci*^{pf}, *říkat*^{impf} 'to say', *sdělit*^{pf}, *sdělovat*^{impf} 'to tell', *tázat se*^{impf} 'to ask', *vyprávět / vypravovat*^{impf} 'to tell', *zeptat se*^{pf} 'to ask', etc.
  As for the morphemic forms, the 'Actor' is realized by the nominative case. The 'Addressee' is morphematically expressed by the dative (e.g., *oznámit*^{pf}, *oznamovat*^{impf} 'to announce', *říci*^{pf}, *říkat*^{impf} 'to say', *sdělit*^{pf}, *sdělovat*^{impf} 'to tell', etc.) (ex 5), by the genitive (e.g., *tázat se*^{impf} 'to ask' (ex 6), *zeptat se*^{pf} 'to ask', etc.), or by the accusative case (e.g., *informovat*^{biasp} 'to inform', etc.) (ex 7).

[1] http://ufal.mff.cuni.cz/vallex/

[2] http://ucnk.ff.cuni.cz/

(5) *Charles Haughey.*ACT(= Speaker, nom) *včera sdělil členům.*ADDR(= Recipient, dat) *své strany, že příští týden rezignuje na svou funkci premiéra.* (SYN2006pub)
E. *Yesterday Charles Haughey.*ACT(= Speaker) *told the members.*ADDR(= Recipient) *of his party that he would resign from his post of the Prime Minister.*

(6) *Juli·n.*ACT(= Speaker, nom) *se zeptal průvodčího.*ADDR(= Recipient, gen) *na cestu.* (SYN2005)
E. *Julian.*ACT(= Speaker) *has asked the conductor.*ADDR(= Recipient) *the way.*

(7) *A tehdy Andrew.*ACT(= Speaker, nom) *informoval Johna Rowea.*ADDR(= Recipient, acc), *že stav jeho ženy je vážný.* (SYN2005)
E. *At that time Andrew.*ACT(= Speaker) *informed John Rowe.*ADDR(= Recipient) *that his wife's condition is serious.*

- Some other verbs of communication express a symmetrical process of communication – the 'Speaker' and 'Recipient' change their roles in the process of communication. As a result, 'Actor's' and 'Addressee's' valency slots are occupied by both 'Speaker' and 'Recipient'. These verbs are lexically reciprocal. The examples of such verbs of communication are the following: *bavit se*^{impf} *s někým* 'to negotiate', *diskutovat*^{impf} *s někým* 'to discuss', *dohodnout se*^{pf}, *dohodovat / dohadovat se*^{impf} *s někým* 'to agree', *hádat se*^{impf} *s někým* 'to quarrel', *hovořit*^{impf} *s někým* 'to speak', *jednat*^{impf} *s někým* 'to confer', *komunikovat*^{impf} *s někým* 'to communicate', *mluvit*^{impf} *s někým* 'to talk', *pohádat se*^{pf} *s někým* 'to dispute', *přít se*^{impf} *s někým* 'to argue', etc.
  As for the morphemic forms of the 'Actor' and 'Addressee', the 'Actor' of these verbs is prototypically realized by the nominative case, and the 'Addressee' by the prepositional group *s* 'with' + the instrumental case. See the following examples:

  (8) *Lesníci.*ACT(= Speaker, Recipient, nom) *se hádají s ekology.*ADDR(= Recipient, Speaker, *s* + instr) *o to, jak mají bránit šíření kůrovce.* (SYN2006pub)
  E. *The foresters.*ACT(= Speaker, Recipient) *argue with the environmentalists.*ADDR(= Recipient, Speaker) *over how they should prevent from the spread of the bark beetle.*

Apart from the above mentioned cases of the realizations of the 'Recipient' (see Section 1.1.1), we observe the following cases:

- Still another group of verbs of communication indicates an asymmetrical process of communication – the 'Message' is addressed to the 'Recipient', however, the active participation of the 'Recipient' in the process of communication is weakened. In these cases, the 'Recipient' fills the 'Addressee's' valency slot and is expressed by the prepositional group *k* 'to' + dative (ex 9) or *na* 'at' + accusative case (ex 10). Examples of these verbs of communication are the following: *bručet*^{impf} *na někoho* 'to growl at', *hovořit*^{impf} *k někomu* 'to talk to', *křiknout*^{pf}, *křičet*^{impf} *na někoho* 'to shout at', *mluvit*^{impf} *k někomu, na někoho* 'to speak to', *řvát*^{impf} *na někoho* 'yell at', *volat*^{impf} *na někoho* 'to call at', etc.

(9) *Buzková*.ACT(= Speaker, nom) *pak mluvila k lidem*.ADDR(= Recipient, *k* + dat), *kteří stáli pod pódiem asi půl druhého metru od ní, a překřikovala nádražní hlášení.* (SYN2006pub)
E. *Then Buzková*.ACT(= Speaker) *has spoken to the people*.ADDR(= Recipient), *who were standing under the podium half a meter from her, shouting down the station report.*

(10) *Ten, když viděl, jak couvá, křičel na něj*.ADDR(= Recipient, *na* + accusative), *že by mohl spadnout.* (SYN2006pub)
E. *When he saw him backing, he*.ACT(= Speaker) *was shouting at him*.ADDR(= Recipient) *that he could fall down.*

- Furthermore, several verbs of communication, as e.g. *dodat*^{pf}, *dodávat*^{impf} 'to add', *definovat*^{biasp} 'to define', *komentovat*^{biasp} 'to comment', *konstatovat*^{biasp} 'to state', *prohlásit*^{pf}, *prohlašovat*^{impf} 'to declare', *zveřejnit*^{pf}, *zveřejňovat*^{impf}, specify the 'Recipient' as an audience which can be expressed as an optional free modification with locative characteristics in most cases (ex 11). For more information on such verbs, see Section 3.1.1 below.

(11) *Ruský preziden Boris Jelcin minulý týden v Kremlu prohlásil, že v Čečně se neděje nic bez jeho vědomí.* (SYN2006pub)
E. *The Russian President Boris Jelcin*.ACT(= Speaker) *declared in Kremlin last week (that nothing was happening in Chechnya without his knowledge.)*

1.1.2. 'Message'

The 'Message'[3] represents a complex participant, two aspects of which are sometimes distinguished: who or what is spoken about (the so-called theme) and what is said about the theme (the so-called dictum). However, in many cases, the theme and the dictum are not distinguishable.

Two aspects of the 'Message' are distinguishable, the 'Message' stands for either the theme (ex 12), or the dictum (ex 13), or both theme and dictum (ex 14). Konečná makes an attempt at specifying the theme and dictum (Konečná, 1966). According to her, the dictum is characterized as an object expressed, especially by a direct or indirect speech. Furthermore, some words referring to a part of text, as *věta* 'sentence', *myšlenka* 'idea', *pravda*, 'truth', *nesmysl* 'nonsense', or some demonstrative or indefinite pronouns, as *to* 'this', *něco* 'something', *nic* 'nothing', etc., can realize the dictum as well. The theme is specified as an object expressed by a noun or a dependent clause under the condition that (i) a dependent clause introduced by the conjunction

[3] Remark on terminology: In *Mluvnice češtiny III* (Mluvnice češtiny III, 1987) and in *Větné vzorce v češtině* (Daneš and Hlavsa, 1987), this complementation is referred to as the participant of information. In *Skladba češtiny* (Grepl and Karlík, 1998), these complementions are classified as the so-called situational actants within which the authors distinguish information, instructions, stimuli and purposes. In our view, the 'Message' involves the information (as in *It was announced in the radio that the dangerous prisoner had escaped from the prison*) and instruction (as in *He allowed me to smoke*).

zda 'whether' can be nominalized (e.g., the verb *analyzovat* 'to analyze', *zkoumat* 'to investigate', etc.) or (ii) meaning of an object is similar to the meaning of the object of the verb *mluvit* 'to speak'.

(12) *Petr Eben a rektor Karlovy univerzity prof. Radim Palouš ve svých vystoupeních hovořili (o spirituálním poslání hudby).*(= Message-theme) (SYN2006pub)
E. *Petr Eben and the rector of Charles University Radim Palouš spoke (about spiritual message of music.*(= Message-theme) *in their performances.*

(13) *Caldwell mi říká, (že má stále úzkostné sny).*(= Message-dictum) (SYN2006pub)
E. *Caldwell tells me (that he still has anxious dreams).*(= Message-dictum)

When the 'Message' represents both the theme and the dictum, we observe two cases: (i) the 'Message' occupies a single valency slot – that of 'Patient's' (ex 14), or (ii) it is split into two valency slots – the theme and dictum occupy each its own valency slot. Then the theme fills the slot of 'Patient' and the dictum the one of 'Effect' (ex 15). This case is referred to as 'splitting of the theme and the dictum'. (For more information, see Section 3.1.2 below.)

(14) *Řekli jsme jim o únosu.*(= Message-theme and dictum), … (SYN2005)
E. *We have told them about the kidnapping.*(= Message-theme and dictum), …

(15) *Říká se o hercích.*PAT(= Message-theme), *(že nemají charakter).*EFF(= Message-dictum) (SYN2006pub)
E. *Actors.*PAT(= Message-theme) *are said (not to be persons of good character).*EFF(= Message-dictum)

The 'Message' can have the following morphemic forms:

- **Dependent content clauses**, the prototypical realization of the 'Message', are discussed in detail in Section 2 below.
- **Prepositionless case**, namely the accusative case (e.g., *deklarovat*^{biasp} 'to declare', *diktovat*^{impf} 'to dictate', *dokázat*^{pf}, *dokazovat*^{impf} 'to demonstrate', *formulovat*^{biasp} 'to phrase', *hlásit*^{impf} 'to report', *konstatovat*^{biasp} 'state', *konzultovat*^{impf} 'to consult', *křiknout*^{pf}, *křičet*^{impf} 'to shout', *naznačit*^{pf} *naznačovat*^{impf}, 'to suggest', *oznámit*^{pf}, *oznamovat*^{impf} 'to announce', *poznamenat*^{pf} *poznamenávat* 'to remark', *psát*^{impf} 'to write', *sdělit*^{pf}, *sdělovat*^{impf} 'to tell', *telefonovat*^{impf} 'to phone', *tvrdit*^{impf} 'to assert', *volat*^{impf} 'to call', *vyprávět / vypravovat*^{impf} 'to tell', *vyslovit*^{pf}, *vyslovovat*^{impf}, 'to say', etc.) The accusative usually expresses the 'Message' with the character of the dictum, if it is distinguishable. See the following examples:

  (16) *Musím čtenářům sdělit příjemnou zprávu.*(= Message-dictum) (SYN2006pub)
  E. *I must tell the readers the pleasant message.*(= Message-dictum)

  (17) *Vyslovil jste naprostou lež.*(= Message-dictum) (SYN2006pub)
  E. *You have pronounced the absolute lie.*(Message-dictum)

- **Prepositional groups.** They realize the 'Message' representing (i) the theme, or (ii) both the theme and the dictum, if these aspects of the 'Message' are distinguishable. The group

occupies the 'Patient's' valency slot. The following prepositional groups belong to the most frequent ones.

- *o* + **locative** (e.g., *bavit*^{impf} *se o něčem* 'to speak about', *diskutovat*^{impf} *o něčem* 'to discuss', *domluvit se*^{pf} *domlouvat se*^{impf} *o něčem* 'to agree on', *hovořit*^{impf} *o něčem* 'to talk about', *informovat*^{biasp} *o něčem* 'to inform on', *jednat*^{impf} *o něčem* 'to confer on', *kázat*^{impf} *o něčem* 'to preach about', *komunikovat*^{impf} *o něčem* 'to communicate about', *vyprávět / vypravovat*^{impf} *o něčem* 'to tell', etc.):

  (18) *V Rio de Janeiru se diskutovalo především o problému.*PAT(= Message-theme) *financování ekologie rozvojového Jihu průmyslovým Severem.* (SYN2006pub)
  E. *Especially the problem.*PAT(= Message-theme) *of financing the ecology of the developing South by the industrial North was discussed in Rio de Janeiro.*

- *na* + **accusative** (e.g., *nadávat*^{impf} *na něco* 'to grumble about', *ptát*^{impf} *se na něco* 'to ask about', *tázat se*^{impf} *na něco* 'to ask about', etc.):

  (19) *…ptal se jí na detaily.*PAT(= Message-theme and dictum) *obou pitev.* (SYN2005)
  E. *…he has asked her about the details.*PAT(= Message-theme and dictum) *of both autopsies.*

- *o* + **accusative** (e.g., *hádat se*^{impf} *o něco* 'to quarrel over', *prosit*^{impf} *o něco* 'to beg for', *přít se*^{impf} *o něco* 'to argue over', etc.):

  (20) *Také Česká televize se nejspíš s Radou pro rozhlasové a televizní vysílání začne přít o výklad.*PAT(= Message-theme) *zákona o České televizi.* (SYN2006pub)
  E. *The Czech Television will start to argue with the Czech Radio and Television Broadcasting Council over the interpretation.*PAT(= Message-theme) *of the law on Czech Television .*

- *po* + **locative** (e.g., *ptát se*^{impf} *po něčem* 'to ask after', *tázat se*^{impf} *po něčem* 'to ask after', etc.):

  (21) *Zrovna tak se neptala po souhlasu.*PAT(=Message) *ingušské vlády s průchodem ruských vojsk přes Ingušsko.* (SYN2006pub)
  E. *Even so it did not ask after the Ingush government approval.*PAT(= Message) *to the Russian army transit across the Ingushetia area.*

- *nad* + **instrumental** (e.g., *diskutovat*^{impf} *nad něčím* 'to discuss', etc.):

  (22) *Středověká církev dlouho diskutovala nad otázkou, zda byl Ježíš Kristus na kříži nahý.*PAT(= Message-theme) (SYN2006pub)
  E. *For a long time, the medieval Church discussed the question whether Jesus Christ was naked on the cross.*PAT(= Message-theme)

- *k* + **dative** (e.g., *přiznat se*^{pf}, *přiznávat se*^{impf} *k něčemu* 'to confess to', etc.):

(23) *Nakonec se však při výslechu přiznala ke lži*.PAT(= Message) *a částku 51 000 korun vrátila.* (SYN2006pub)
E. *Finally, she confessed the lie*.PAT(= Message) *during the interrogation and she gave the sum of 51 000 crowns back.*

- *z* + **genitive** (e.g., *obvinit*^{pf}, *obviňovat*^{impf} *z něčeho* 'to blame for', *nařknout*^{pf}, *nařkat*^{impf} *z něčeho* 'to accuse of', etc.):

(24) *Před několika dny se navzájem obvinili ze lži*.PAT(= Message) (SYN2006pub)
E. *They blamed each other for the lie*.PAT(= Message) *several weeks ago.*

- *na* + **locative** (e.g., *domluvit se*^{pf}, *domlouvat se*^{impf} *na něčem* 'to agree on', *dohodnout se*^{pf} *na něčem*, *dohodovat se / dohodovat se*^{impf} *na něčem* 'to arrange', etc.):

(25) *Lidem se možná bude zdát, že jsme se domluvili na společném tématu*.PAT(= Message) (SYN2006pub)
E. *It may seem to the people that we have agreed on the common topic*.PAT(= Message)

- **Infinitives.** With particular verbs of communication, the participant 'Message' may be expressed by an infinitive, see esp. (Panevová, 1996) and (Mikulová et al., 2005). The referential correspondence either between the 'Actor' (ex 26) or the 'Addressee' (ex 27), or between both the 'Actor' and 'Addressee' (ex 28 and 29) on the one hand and the subject of the given infinitive on the other is typical of these verbs.
The 'Message' of the following verbs of communication can be expressed by an infinitive: *dovolit*^{pf}, *dovolovat*^{impf} 'to allow', *doporučit*^{pf}, *doporučovat*^{impf} 'to recommend', *nařídit*^{pf}, *nařizovat*^{impf} 'to order', *navrhnout*^{pf}, *navrhovat*^{impf} 'to suggest', *poručit*^{pf}, *poroučet*^{impf} 'to command', *přikázat*^{pf}, *přikazovat*^{impf}, 'to command', *přísahat*^{impf} 'to swear', *slíbit*^{pf}, *slibovat*^{impf} 'to promise', *uložit*^{pf}, *ukládat*^{impf} 'to oblige', *zakázat*^{pf}, *zakazovat*^{impf} 'to prohibit', etc. See the following examples:

(26) *Faust*.ACT *mu*.ADDR *přece slíbil vše proti Bohu a křesťanství dělat*.PAT …(SYN2005)
E. *Faust*.ACT *has promised him*.ADDR *to do*.PAT *everything against God and religion.*

(27) *A dovolil jim*.ADDR *chodit*.PAT *na nákupy, kdy si jen vzpomněly.* (SYN2005)
E. *And he*.ACT *allowed them*.ADDR *to go*.PAT *shopping whenever they had wanted.*

(28) *…prezident*.ACT *nabídl kancléři*.ADDR *umožnit*.PAT *sudetským Němcům účast na privatizacích …* (SYN2006pub)
E. *…the President*.ACT *has proposed the chancellor*.ADDR *to allow*.PAT *Germans to take part in privatizations …*

(29) *Zřízení*.ACT *ekonomicko-správní fakulty nabídlo i absolventům*.ADDR *jiných škol doplnit*.PAT *si vzdělání v oboru, který v Brně po celá léta nebylo možno studovat.* (SYN2006pub)

E. *The establishment*.ACT *of the Economic-administrative faculty has offered also graduates*.ADDR *of other schools to complete*.PAT *their qualification in the field which had not been possible to study in Brno for a long time.*

2. Dependent Content Clauses

In most cases, the participant 'Message' of the verbs of communication is expressed by dependent content clauses (DCCs in the sequel).[4] However, with several verbs of communication, this participant cannot be expressed by the DCCs. For instance, the verbs *bavit se*^{impf} 'to talk', *definovat*^{biasp} 'to define', *diskutovat*^{impf} 'to discuss', *hovořit*^{impf} 'to talk', *charakterizovat*^{biasp} 'characterize', *komunikovat*^{impf} 'to communicate', *mluvit*^{impf} 'to speak' represent such exceptions.

We distinguish three types of the DCCs according to their modality: assertive, interrogative and directive DCCs. These types are formally characterized by the type of subordinating conjunctions, and by several temporal and modal devices, see esp. (Běličová-Křížková, 1979). These devices stand in the center of our interest.

2.1. Assertive Dependent Content Clauses

The assertive DCCs (assertDCC in the sequel) express the content of what is indicated as a statement by the governing verb. The assertDCCs are typically introduced by the subordinating conjunction *že* 'that', cf. Section 2.1.1. They can be usually paraphrased by a direct speech with declarative sentential modality. See the following example and its paraphrase:

(30) *Řekl jsem jim, že odcestoval do Evropy a že nevím přesně kam.* (SYN2005)
E. *I told them that he had departed to Europe and I did not know precisely where.*
[*Řekl jsem jim: "Odcestoval do Evropy a nevím přesně kam."*][5]
[E. *I told them: 'He departed to Europe and I don't know precisely where.'*]

Relative tenses are characteristic of this type of the DCCs. The use of the relative tenses follows the rules indicated esp. in (Bauer, 1965), (Panevová, Benešová, and Sgall, 1971), (Mluvnice češtiny II, 1986) and (Mluvnice češtiny III, 1987).

As for the **verbal mood**, the indicative mood is typical of the assertDCCs (ex 31). The conditional may indicate desirable (ex 32) or potential events (ex 33), events denied by the

[4] On the other hand, the DCCs do not realize only the participant 'Message' of the verbs of communication, they may be also a morphemic realization of one of valency complementations of the verbs indicating (i) mental actions (e.g., *Komunisté mínili, že ti kteří nemohou do továren jako jiní, mají sedět doma a být zticha.* (SYN2006pub) E. *The communists thought that those who could not work in factories should sit at home and should keep quiet*), (ii) perception (e.g., *Ta námaha ale stojí za to, když vidíte děti, že se jim ze školky nechce domů ...* (SYN2006pub) E. *Seeing children not wanting to go home is worth the effort ...*), or (iii) psychological states (e.g., *Překvapilo ho, že znovu mluví o své operaci.* (SYN2000) E. *It suprised him that he was speaking about his operation again*).

[5] Czech paraphrases are given in square brackets.

'Speaker' (ex 34), etc. The conditional of the verbs of communication in the governing clauses does not interfere with the modality of the DCC.

(31) *Řekl si, (že potřebuje víc informací).*(assertDCC) (SYN2005)
E. *He said to himself (that he needed more information).*(assertDCC)

(32) *Řekla bych, (že bychom se s Chrisem měli už vrátit).*(assertDCC) (SYN2000)
E. *I would say (that it is about time for me and Chris to go back).*(assertDCC)

(33) *Řekla, (že by se Bob naštval).*(assertDCC) (SYN2000)
E. *She said (that Bob would be angry).*(assertDCC)

(34) *Nikdy bych neřekl, (že by se ve mně vzala taková síla).*(assertDCC) (SYN2000)
E. *I would never say (that such strength would gather in me).*(assertDCC)

2.1.1. Assertive Dependent Content Clauses introduced by *zda* 'whether'

The assertDCCs may be connected by the conjunctions *zda, zdali, -li,* or *jestli* as well. In such cases, a 'Speaker' conveys only incomplete information to a 'Recipient' (ex 35) in contrast to the assertDCCs connected by the conjunction *že* 'that' (ex 36) which convey complete information:

(35) *Ministerstvo žadateli sdělí, (zda byl na něj vůbec nějaký spis veden a zda se zachoval.)* (assertDCC with incomplete information) (SYN2006pub)
E. *The Ministry told an applicant (whether any file about him was kept at all and whether the file is preserved.)* (assertDCC with incomplete information)

(36) *Za okamžik se vrátil a sdělil nám, (že pan Baker je v zahradě.)* (SYN2005) (assertDCC with complete information)
E. *He came back in a moment and he told us (that Mr Baker is in the garden.)* (assertDCC with complete information)

However, the conjunctions *zda, zdali, -li,* or *jestli* introduce the interrogative DCCs (interDCCs in the sequel) as well (Section 2.2). They express the 'Speaker's' uncertainty whether the content of the DCCs holds or not (ex 37):

(37) *Sdělte mi, prosím, (zda s diskriminací vašeho listu souhlasíte …)* (interDCC) (SYN2005)
E. *Please, tell me (whether you agree with the discrimination against your newspaper …)* (interDCC)

In contrast to the interDCCs, the assertDCCs introduced by the conjunctions *zda, zdali, -li,* or *jestli* do not exhibit the interrogative characteristic – they do not express the 'Speaker's' uncertainty or lack of knowledge, see (Daneš and Hlavsa, 1987). In ex 38 with the assertDCCs, the possibility that *the member of the presidium* knows whether *Izetbegović will take part in the peace talks in New York or not* is not excluded in contrast to ex 39 with the interDCC where

the 'Speaker' asks the 'Recipient' to answer his question because he does not know whether *the coalition satisfies him*.[6]

(38) *Člen předsednictva neřekl, (zda se Izetbegović zúčastní mírových rozhovorů v New Yorku.)* (assertDCC with incomplete information) (SYN2006pub)
E. *The member of the presidium did not tell (whether Izetbegović would take part in the peace talks in New York.)* (assertDCC with incomplete information)

(39) *Řekněte, (zda vám vyhovuje koalice.)* (interDCC) (SYN2006pub)
E. *Tell (whether the coalition satisfies you.)* (interDCC)

Lastly, the assertDCCs introduced by *zda, zdali, -li,* or *jestli* usually indicate mutually excluding alternatives and they are characterized by the possibility of having a positive or a negative form without any change of meaning. See the following examples:

(40) *Soudci však neřekli, zda útočníci se svým jednáním provinili proti tehdejším zákonům.* (SYN2006pub)
E. *However, the judges have not said whether the attackers had violated the laws of that time by their actions.*

(41) *Příští pátek a sobotu občané v referendu řeknou, zda si přejí vstup do NATO …* (SYN2006pub)
E. *The next Friday and Saturday the citizens are going to say whether they want to join NATO …*

2.2. Interrogative Dependent Content Clauses

InterDCCs indicate the content of what is indicated by the governing verb as the question – they express 'Speaker's' uncertainty or lack of knowledge, etc. They are usually connected by the conjunctions *zda*, *zdali*, *-li* and *jestli* 'if', 'whether'. (For more information on the difference between interDCCs and assertDCCs, see Section 2.1.1 above). The interDCCs can be paraphrased by a direct speech with interrogative sentential modality. These direct speeches have the form of a yes / no question. See the following example of the interDCCs and their paraphrase by the direct speech:

(42) *Král si dal dcery zavolat a ptal se jich, (zda voják mluví pravdu).*(interDCC) (SYN2000)
E. *The king had his daughters called and he asked them (whether the soldier told the truth)*
(*Král si dal dcery zavolat a ptal se jich: "Mluví voják pravdu?"*)
(E. *The king had his daughters called and he asked them: 'Does the soldier tell the truth?'*)

The **relative tenses** are characteristic of the interDCCs, similarly as in the case of the asertDCCs, see (Panevová, Benešová, and Sgall, 1971), (Bauer, 1965), (Mluvnice češtiny III, 1987),

[6] Apparently, the imperative mood of the governing verb influences the choice of the following DCC. However, we leave aside the interplay between the grammatical categories of the governing verbs and the type of DCC as this issue requires a further investigation.

and (Mluvnice češtiny II, 1986). As for the **verb mood**, the indicative mood expresses a question (ex 43). The conditional mood prevails in the interDCCs expressing a proposal (ex 44), a polite request (ex 45), etc.

(43) *Povězte mi, jestli o něm něco víte.* (SYN2005)
E. *Tell me whether you know anything about him.*

(44) *Bydlel v téměř prázdném domě a jednou řekl "pár lidem", zda by se do volných bytů nechtěli nastěhovat.* (SYN2006pub)
E. *He lived in an almost empty flat and once he told 'few people' if they wouldn't move in the vacant flats.*

(45) *Řekl jsem hlavnímu sudímu, zda by ho nemohl vyměnit.* (SYN2006pub)
E. *I told the chief referee if he could replace him.*

2.3. Directive Dependent Content Clauses

Directive DCCs (directDCC in the sequel) express the content of what is indicated by the governing verb as a command, appeal, request, etc. These DCCs denote events which have not been realized yet but the realization of which is desirable for the 'Speaker' – they generally refer to the future. They are typically introduced by subordinating conjunctions *aby* 'so that' and *ať* 'to let'. The conjunction *aby* implies the conditional mood of verbs.

The directDCCs can be paraphrased by direct speeches with the imperative sentential modality. See the following example and its paraphrase by the direct speech:

(46) *Poté zákazník přikázal taxikáři, (aby jej odvezl ke stanici Budějovická).*(directDCC) (SYN2006pub)
E. *Then the client has ordered the taxi driver to take him to the station Budějovická.*
(*Poté zákazník přikázal taxikáři: "Odvezte mě ke stanici Budějovická!"*)
(E. *Then the client has ordered the taxi driver: 'Take me to the station Budějovická!'*)

3. Subclasses of the Verbs of Communication

The information on the type of the DCCs, which the verbs of communication require to be complemented by, is taken as a key criterion for subdividing this group of verbs into semantically and morphosyntactically more coherent classes.

The type of the dependent content clauses serves as the basis of the classification for the following reasons: (i) the type of the dependent content clauses reveals the semantic properties of the governing verb to some extent. For instance, the interDCCs do not realize the 'Message' of the verbs of communication expressing an order (e.g., **Nařídil mu, jestli přijde večer brzy.* E.**He ordered him whether he comes early*). Vice versa, if a verb of communication expresses a question, it cannot be complemented by a directDCC (e.g., **Ptal se ho, aby něco udělal / ať něco udělá.* E. **He questioned him to do something*), see (Mluvnice češtiny III, 1987) and (Běličová and Sedláček, 1990).

(ii) Each type of the dependent content clauses is regularly associated with several other morphosyntactic properties. For instance, the splitting of the theme and the dictum is realized only when a particular verb of communication is complemented by the assertDCC. Furthermore, if a particular verb of communication governs the interDCC or directDCC, then its valency frame contains the obligatory 'Addressee'. On the other hand, being complemented by the assertDCC, the verbs of communication have an obligatory or an optional 'Addressee' or it is not present in their valency frames at all.

As a result of this classification, the verbs of communication are divided into three subtler classes – assertive (Section 3.1), interrogative (Section 3.2) and directive verbs of communication (Section 3.3). Furthermore, in Section 3.4, we propose a method how to treat the 'neutral verbs' – these verbs can be complemented by all three types of the DCCs.

3.1. Assertive Verbs of Communication

This subclass contains the verbs of communication which require to be complemented by the **assertDCCs**. The verbs of this subclass denote such events of speaking in which the content of the 'Message' is conveyed by the 'Speaker' as a fact. This subclass contains the following verbs: *líčit*^{impf} 'to depict', *lhát*^{impf} 'to lie', *vyprávět / vypravovat*^{impf} 'to tell', *žalovat*^{impf} 'to complain', etc. The 'Speaker' can express his attitude to the truthfulness of the content of the assertDCC. See the following examples:

(47) *Někdo mi vyprávěl, že zde snad ještě můžeme dostat vízum na Haiti nebo do San Dominga.* (SYN2000)
E. *Somebody told me that maybe it was possible for us to get visa for Haiti and San Domingo here.*

Two issues concerning the assertive verbs of communication will be discussed in more detail: (i) the participant 'Recipient' (Section 3.1.1) and (ii) the splitting of the theme and dictum (Section 3.1.2).

3.1.1. 'Addressee' of Assertive Verbs of Communication

The participant 'Recipient' is realized in the 'Addressee's' valency slot. This slot can be obligatory (as in the cases of the verbs *konzultovat*^{biasp} 'to consult', *svěřit se*^{pf}, *svěřovat se*^{impf} 'to confide', *vyprávět / vypravovat*^{impf} 'to tell', *žalovat*^{impf} 'to complain', etc.) or optional (as in the cases of the verbs *číst*^{impf} 'to read', *chlubit se*^{impf} 'to boast', *lhát*^{impf} 'to lie', *líčit*^{impf} 'to depict', *zmínit se*^{pf}, *zmiňovat se*^{impf} 'to mention', etc.).

In case that the valency frames of assertive verbs do not contain an 'Addressee's' slot, an audience, which does not actively participate in the event of speaking, can be morphematically expressed especially by the prepositional group *před* 'in front of' + instrumental representing an optional free modification with locative meaning.

(48) *Mnohé firmy nejsou vůbec schopny definovat před svými zaměstnanci, co je obchodní tajemství.* (SYN2006pub)

E. *Many companies are not able to define in front of their employees what the trade secret is.*

The examples of the assertive verbs of communication without 'Addressee's' valency slot are the following: *definovat*^{biasp} 'to define', *deklarovat*^{biasp} 'to declare', *komentovat*^{biasp} 'to comment', *konstatovat*^{biasp} 'to claim', etc.

3.1.2. Splitting of the Theme and the Dictum

The splitting of the theme and the dictum – i.e. the participant 'Message' occupies two valency slots: 'Patient' and 'Effect' – is typical of several verbs of communication of this subclass, e.g., *číst*^{impf} 'to read', *líčit*^{impf} 'to depict', *konstatovat*^{biasp} 'to claim', *vyprávět / vypravovat*^{impf} 'to tell', *zmínit se*^{pf}, *zmiňovat se*^{impf} 'to mention', *žalovat*^{impf} 'to complain', see (Daneš and Hlavsa, 1987), (Mluvnice češtiny III, 1987) and (Součková, 2005).[7]

Furthermore, this property is characteristic of most 'neutral' verbs of communication (Section 3.4), e.g., *hlásat*^{impf} 'to propagate', *hlásit*^{impf} 'to report', *oznámit*^{pf}, *oznamovat*^{impf} 'to announce', *povědět*^{pf}, *povídat*^{impf} 'to tell', *psát*^{impf} 'to write', *říci*^{pf}, *říkat*^{impf} 'to say', *sdělit*^{pf}, *sdělovat*^{impf} 'to tell', *šeptnout*^{pf}, *šeptat*^{impf} 'to whisper'. See the following example:

(49) *Řekla o mně, že jsem línej jako veš.* (SYN2000)
E. *'Said – about – me – that – (I-)am – lazy – as – louse.'*

As for the morphemic form of the 'Patient', it can be expressed by the following prepositional groups *o* 'about' + locative, *na* 'about' + accusative and *k* 'on' + dative (for more information, see Section 1.1.2 above). The 'Effect' is realized by the assertDCCs and by the accusative in some cases.

The separated part of the 'Message' realized in the governing clause is always in the relation of coreference with an expression or with a whole segment of the DCC, see (Hajičová, Panevová, and Sgall, 1985-1987). We observe the cases of (i) textual coreference – the separated part is referentially identical with a personal pronoun (ex 50) – and the cases of (ii) a more complicated relation between the separated part and the anaphoric element. To a great extent, this relation is based on the shared knowledge. For instance, the separated part and the anaphoric element can be in the relation of metonymy (as *mother* and *her tongue* in ex 51), synonymy (as *Agassi's ability of returning services* and *his returns* in ex 52), hyponymy and hyperonymy (as *pub* and *facility* in ex 53), see esp. (Cruse, 1986), or (Filipec and Čermák, 1985). If a whole segment or even a whole assertDCC represent the anaphoric device, the content relationship between them may be very loose (as *Milevina's limping* and *he would have never had courage to get married to a wife who would not be absolutely healthy* ex 54).

(50) *Ramos o návštěvě řekl, že významně uvolňuje napětí mezi oběma zeměmi.* (SYN2005)
E. *'Ramos – about – visit – said – that – significantly – (it-)eases – tension – between – both – countries.'*

[7] Some verbs expressing mental activity allow for the splitting of the theme and dictum as well.

(51) *Když jsem však řekla o matce, …, že jí pusa jede jako dítěti …* (SYN2005)
E. *'When – however – (I-)said – about – mother – …– that – her – tongue – never gives a rest – …'*

(52) *Poražený první hráč řekl o schopnosti Agassiho vracet podání, že jeho returny byly jako laserové paprsky.* (SYN2000)
E. *'Beaten – first – player – said – about – ability – Agassi's – return – services – that – his – returns – were – as – laser – beams.'*

(53) *…tvrdí Pavel Doležal o své hospodě U Andyho, že jej podnik dobře uživí.* (SYN2006pub)
E. *'…claims – Pavel – Doležal – about – his – pub – at – Andy's – that – him – facility – well – maintains.'*

(54) *Uvádí se, že jistý Einsteinův kolega jednou řekl o Milevině kulhání, že by nikdy neměl odvahu oženit se s ženou, která by nebyla absolutně zdravá.* (SYN2005)
E. *It is stated that an Einstein's colleague has said about Milevina's limping that he would never have courage to get married to a wife who would not be absolutely healthy.*

The anaphoric element may occur in different **syntactic positions**: in the position of the subject (ex 50), the direct object (ex 55 and 56), the indirect object (ex 56) or in the adverbial position (ex 57):

(55) *Miloš Zeman prohlásil o Wagnerovi, že ho do svých řad nechtěli ani komunisté.* (SYN2006pub)
E. *'Miloš – Zeman – declared – about – Wagner – that – him – in – their – ranks – had not wanted – even – the communists.'*

(56) *…a psát o ní, že ji vlastně vzývá a očekává od ní pomoc a požehnání v nejnesmyslnějších věcech.* (SYN2005)
E. *…and – write – about – her – that – in fact – invokes – and – expects – from – her – help – and – blessing – in – the most unreasonable – situations.*

(57) *Je nepřesné říci o Marxovi, že technický pokrok znamená podle něj vždy úsporu práce.* (SYN2005)
E. *'It – is – not exact – say – about – Marx – that – technical – progress – implies – according to – him – always – the saving – work.'*

The splitting of the theme and the dictum represents a difficulty in the description of the valency structure as the verbs allowing the splitting of the theme and the dictum are regularly used without such a splitting in other contexts as well. As a result, two separated valency frames have to be postulated despite an apparent similarity in their meanings.

For the purpose of an explicit description of the valency structures of the verbs of communication, we propose to exploit the alternation model according to which the alternations are taken as regular changes in the valency structure. (This model was outlined for the purpose of *VALLEX, Valency Lexicon of Czech Verbs*, see (Lopatková, Žabokrtský, and Kettnerová, 2008), (Markéta Lopatková, 2006) and (Žabokrtský, 2005)).

Under such a treatment, the splitting of the theme and the dictum represents a syntactic alternation (SplTD in the sequel), applicable to **some assertive verbs of communication** or **'neutral' cases when complemented by the assertDCCs** (Section 3.4). SplTD is characterized by the changes in the valency frame – in the number of valency complementations and their morphemic forms. However, this alternation is not accompanied by a substantial change in the lexical meaning – the separated part of the DCC is only emphasized. For illustration, the rules of the SplTD applicable to the verb *říci*^{pf}, *říkat*^{impf} 'to tell' can be formulated as follows:[8]

ACT^{1}	ACT^{1}
[ADDR]^{3}	[ADDR]^{3}
PAT^{4,assertDCC}	$\Rightarrow$ PAT$^{k+3,na+4,o+6}$ EFF^{4,assertDCC}

Table 1. The SplTD alternation applicable to the verb říci^{pf}*, říkat*^{impf} *'to tell': (i) PAT is split into PAT and EFF and (ii) the morphemic forms of PAT are changed.*

3.1.3. Valency Frames of the Assertive Verbs

In summary, the participant **'Speaker'** occupies the 'Actor's' valency slot which is obligatory. The participant **'Recipient'** fills the 'Addressee's valency slot which can be obligatory or optional; some assertive verbs do not contain the 'Addressee's' slot in their valency frames at all, see Section 3.1.1 above.

The participant **'Message'** can fill a single valency slot, then it is expressed as the 'Patient', or it can be realized in two valency slots: its theme is realized as the 'Patient' and its dictum as the 'Effect', see Section 1.1.2 above. In conclusion, we introduce a list of all the assertive verbs of communication enumerated at the beginning of this section and their valency frames (Table 2). The valency frames involving the 'Effect' are the ones derived by the SplTD.

3.2. Interrogative Verbs of Communication

Interrogative verbs of communication represent a relatively restricted set. Their participant 'Message' is prototypically expressed by the **interDCCs** (Section 2.2). The verbs of this subclass express getting knowledge or verifying particular information – they denote those events of speaking in which the 'Speaker' urges the 'Recipient' to provide him with particular information which is unknown to him, or to confirm or disprove particular information. See the following example:

(58) *Příští den jsem se ho.*ADDR *otázal, (zda bych si mohl u něj ještě den odpočinout).*PAT (SYN2006pub)

[8] The square brackets indicate that the given valency complementation is optional.

assertive verb of communication	valency frame
číst$^{impf}{}_1$ 'to read'	ACT$_1$ [ADDR]$_3$ PAT$_{4,o+6,assertDCC}$
číst$^{impf}{}_2$ 'to read'	ACT$_1$ [ADDR]$_3$ PAT$_{o+6}$ EFF$_{4,assertDCC}$
definovat^{biasp} 'to define'	ACT$_1$ PAT$_{4,assertDCC}$
deklarovat^{biasp} 'to declare'	ACT$_1$ PAT$_{4,assertDCC}$
chlubit se^{impf} 'to boast'	ACT$_1$ [ADDR]$_3$ [PAT]$_{7,s+7,assertDCC}$
líčit^{impf} 'to depict'	ACT$_1$ [ADDR]$_3$ PAT$_{4,assertDCC}$
lhát^{impf} 'to lie'	ACT$_1$ [ADDR]$_3$ PAT$_{o+6,assertDCC}$
komentovat^{biasp} 'to comment'	ACT$_1$ PAT$_{4,asertDCC}$
konstatovat$^{biasp}{}_1$ 'to claim'	ACT$_1$ PAT$_{4,asertDCC}$
konstatovat$^{biasp}{}_2$ 'to claim'	ACT$_1$ PAT^{o+6} EFF$_{4,asertDCC}$
konzultovat^{biasp} 'to consult'	ACT$_1$ ADDR$_{s+7}$ PAT$_{4,o+6,assertDCC}$
svěřit se^{pf}, *svěřovat se*^{impf} 'to confide'	ACT$_1$ ADDR$_3$ PAT$_{s+7,assertDCC}$
vyprávět / vypravovat$^{impf}{}_1$ 'to tell'	ACT$_1$ ADDR$_3$ PAT$_{4,o+6,assertDCC}$
vyprávět / vypravovat$^{impf}{}_2$ 'to tell'	ACT$_1$ ADDR$_3$ PAT$_{o+6}$ EFF$_{4,assertDCC}$
zmínit se$^{pf}{}_1$, *zmiňovat se*$^{impf}{}_1$ 'to mention'	ACT$_1$ [ADDR]$_3$ PAT$_{o+6,assertDCC}$
žalovat$^{impf}{}_1$ 'to complain'	ACT$_1$ ADDR$_3$ PAT$_{4,na+4,assertDCC}$
žalovat$^{impf}{}_2$ 'to complain'	ACT$_1$ ADDR$_3$ PAT$_{na+4}$ EFF$_{4,assertDCC}$

Table 2. The list of the assertive verbs of communication and their valency frames.

E. *I.*ACT *asked him.*ADDR *the next day (whether I could have a rest by him for one more day).*PAT

The following verbs represent the examples of the interrogative verbs of communication: *otázat se*^{pf} 'to ask', *ptát se*^{impf}, *tázat se*^{impf}, *vyptat se*^{pf}, *vyptávat se*^{impf}, *zeptat se*^{pf}, etc.

3.2.1. Valency Frame of the Interrogative Verbs

The valency frame of the interrogative verbs of communication contains the 'Actor's', 'Addressee's' and 'Patient's' obligatory valency slots. The participant 'Speaker' occupies the 'Actor's' valency slot, the 'Recipient' fills the 'Addressee's' one and the 'Message' occurs in the 'Patient's' slot. These verbs do not allow the splitting of the theme and the dictum. The list summarizing the valency characteristics of the interrogative verbs enumerated in this section is given in Table 3.

3.3. Directive Verbs of Communication

The participant 'Message' of these verbs of communication is expressed by the **directDCCs** and under conditions discussed in Section 3.3.1 below also by the **assertDCCs**.

The 'Speaker' represents an external stimulus expressing the volition to (non-)realize the

interrogative verb of communication	valency frame
otázat se^{pf} 'to ask'	ACT$_1$ ADDR$_2$ PAT$_{na+4,interDCC}$
ptát se^{impf} 'to ask'	ACT$_1$ ADDR$_2$ PAT$_{na+4,interDCC}$
tázat se^{impf}, 'to ask'	ACT$_1$ ADDR$_2$ PAT$_{na+4,interDCC}$
vyptat se^{pf}, *vyptávat se*^{impf}, 'inquire'	ACT$_1$ ADDR$_2$ PAT$_{na+4,interDCC}$
zeptat se^{pf}, 'to ask'	ACT$_1$ ADDR$_2$ PAT$_{na+4,interDCC}$

Table 3. The list of the interrogative verbs of communication and their valency frames.

action expressed in the DCCs (as *taking on retirees* in ex 59). The actual performer of this action is situated in the 'Addressee's' slot (*employers* here):

(59) *Nemůžeme nařídit zaměstnavatelům, aby důchodce zaměstnávali či naopak.* (SYN20006pub)
E. *We cannot order employers to take on retirees or not.*

The 'Speaker's' volition can be expressed by verbs denoting a command (e.g., *nakázat*^{pf}, *nakazovat*^{impf} 'to enjoin', *nařídit*^{pf}, *nařizovat*^{impf} 'to order', *poručit*^{pf}, *poroučet*^{impf} 'to dictate', *přikázat*^{pf}, *přikazovat*^{impf}, 'to command', *uložit*^{pf}, *ukládat*^{impf} 'to oblige', etc.), a request (*požádat*^{pf}, *požadovat*^{impf} 'to ask', etc.), a prohibition (e.g., *zakázat*^{pf}, *zakazovat*^{impf} 'to prohibit', etc.), a recommendation (e.g., *doporučit*^{pf}, *doporučovat*^{impf} 'to recommend', etc.), a permission (e.g., *dovolit*^{pf}, *dovolovat*^{impf} 'to allow', etc.), a proposal (e.g., *nabídnout*, *nabízet* 'to offer', *navrhnout*^{pf}, *navrhovat*^{impf} 'to suggest', etc.),[9] a challenge (*vyzvat*^{pf}, *vyzývat*^{impf} 'to challenge'), etc.

3.3.1. Assertive DCCs Dependent on the Directive Verbs of Communication

According to (Mluvnice češtiny III, 1987), the assertDCCs (introduced only by the subordinating conjunction *že* 'that') can realize the participant 'Message' of the directive verbs of communication under the condition that a modal verb is present there. However, the corpus evidence does not support this assumption: in a considerable portion of the assertDCCs dependent on a directive verb, no modal verbs are found. For instance, in SYN2006pub 17.5% of assertDCCs dependent on the verb *nařídit*^{pf} 'to order' do not contain any modal verb. Similarly, no modal verb occurs in 10.5% and even 66% of assertDCCs governed by the verb *dovolit*^{pf} 'to allow' and *navrhnout*^{pf} 'to suggest', respectively. On the other hand, these assertDCCs have the same temporal perspective referring to the future as the directDCCs. Similarly, they express desirable events which have not yet been realized. See the following examples:

(60) *Navrhli jsme mu, že mu vrátíme peníze.* (SYN2005)
E.*We suggested that we will give him money back.*

[9] These verbs can be complemented by the interDCCs as well. However, being complemented by the interDCCs, they express a polite proposal. See the following examples: *Navrhl mi, abych se přestěhoval.* E. *He has suggested that I moved.* and *Navrhl mi, zda bych nechtěl jít do kina.* E. *He has suggested going to the cinema, if I liked.*

(61) *Policie ale nařídila, že od února tady budou auta opět jezdit v obou směrech.* (SYN2005)
E. *However, the police has ordered that cars would go here in both directions from February.*

As for modal verbs, their range is limited in these constructions: on the basis of corpus evidence, only modal verbs relating to the modal categories (i) **necessity** (expressed by *muset* 'must', 'have to', *nemoci* 'not be allowed', *mít* 'ought', 'should', *nemít* 'ought not', 'should not', and *nesmět* 'must not' and (ii) **possibility** (expressed by *moci* 'can' , *nemuset* 'need not', and *smět* 'be allowed') occur in the assertDCCs dependent on the directive verbs of communication, see esp. (Kettnerová-Benešová, 2007). More information on modal categories can be found in (Mluvnice češtiny III, 1987). See the following examples:

(62) *Nařídili mi, že se musím do pěti dnů dostavit na urgentní poradu.* (SYN2000)
E. *'Ordered – me – that – refl – must – in – five – days – come – to – urgent – meeting.'*

(63) *Doporučili jí, že by měla odejít.* (SYN2006pub)
E. *'Recommended – her – that – should – resign.'*

(64) *Navrhl jsem mu, že královnin portrét by mohl být součástí jeho výstavy v Holandsku …*(SYN2006pub)
E. *'Suggested – him – that – queen's – portrait – could – be – a part – his – exhibition – in – the Netherlands …'*

The DCCs of the mentioned type do not contain modal verbs relating to the modality of intention (expressed by *chtít* 'to want' and *hodlat* 'to intend') and the modal meaning of ability (expressed by *umět* 'be able' and *dovést* 'be able'). This restriction follows from the fact that the intention and ability are in competence of the actor of the action himself, so they cannot be affected by the volition of the 'Speaker' as an external stimulus (Section 3.3), see (Kettnerová-Benešová, 2007).

3.3.2. Valency Frame of the Directive Verbs

The valency structure of the directive verbs of communication consists of three obligatory slots: the 'Actor', 'Addressee' and 'Patient'. The participant 'Speaker' occupies the 'Actor's' valency slot, the 'Recipient' and the 'Message' fill the slots of the 'Addressee' and the 'Patient', respectively. They do not allow the splitting of the theme and the dictum. Table 4 presents a list summarizing the directive verbs enumerated in this section and their valency frames:

3.4. 'Neutral' Cases of Verbs of Communication

Some verbs of communication allow for being complemented by all three types of the DCCs. In connection with a particular type of the DCCs, these verbs may express:

1. a **statement** when complemented by an assertDCC:

(65) *Řekla, že ji bolí hlava.* (SYN2005)
E. *She said that she had got headache.*

directive verb of communication	valency frame
nakázat^{pf}, *nakazovat*^{impf} 'to enjoin'	ACT$_1$ ADDR$_3$ PAT$_{4,inf,directDCC,assertDCC}$
nařídit^{pf}, *nařizovat*^{impf} 'to order'	ACT$_1$ ADDR$_3$ PAT$_{4,inf,directDCC,assertDCC}$
poručit^{pf}, *poroučet*^{impf} 'to dictate'	ACT$_1$ ADDR$_3$ PAT$_{4,inf,directDCC,assertDCC}$
přikázat^{pf}, *přikazovat*^{impf}, 'to command'	ACT$_1$ ADDR$_3$ PAT$_{4,inf,directDCC,assertDCC}$
uložit^{pf}, *ukládat*^{impf} 'to oblige'	ACT$_1$ ADDR$_3$ PAT$_{4,inf,directDCC,assertDCC}$
požádat^{pf} 'to ask'	ACT$_1$ ADDR$_4$ PAT$_{o+4,inf,directDCC,assertDCC}$
zakázat^{pf}, *zakazovat*^{impf} 'to prohibit'	ACT$_1$ ADDR$_3$ PAT$_{4,inf,directDCC,assertDCC}$
doporučit^{pf}, *doporučovat*^{impf} 'to recommend'	ACT$_1$ ADDR$_3$ PAT$_{4,inf,directDCC,assertDCC}$
dovolit^{pf}, *dovolovat*^{impf} 'to allow'	ACT$_1$ ADDR$_3$ PAT$_{4,inf,directDCC,assertDCC}$
navrhnout^{pf}, *navrhovat*^{impf} 'to suggest'	ACT$_1$ ADDR$_3$ PAT$_{4,inf,directDCC,assertDCC}$
vyzvat^{pf}, *vyzývat*^{impf} 'to challenge'	ACT$_1$ ADDR$_4$ PAT$_{k+3,inf,directDCC,assertDCC}$

Table 4. The list of the directive verbs of communication and their valency frames.

2. a **question** when complemented by an interDCC:

   (66) *Můžete říct, zda opustil během noci kupé?* (SYN2005)
   E. *Could you say whether he left the compartment during the night?*

3. an **order** when complemented by a directDCC:

   (67) *Řeknu jí, aby vám napsala a pozvala vás.*(SYN2005)
   E. *I will ask her to write to you and invite you.*

The following verbs of communication behave 'neutrally' with regard to the types of the DCCs: *informovat*^{biasp} 'to inform', *křiknout*^{pf}, *křičet*^{impf} 'to shout', *oznámit*^{pf}, *oznamovat*^{impf} 'to announce', *podotknout*^{pf}, *podotýkat*^{impf} 'to remark', *povědět*^{pf}, *povídat*^{impf} 'to tell', *psát*^{impf} 'to write', *poznamenat*^{pf}, *poznamenávat*^{impf} 'to remark', *říci*^{pf}, *říkat*^{impf} 'to say', *sdělit*^{pf}, *sdělovat*^{impf} 'to tell', *šeptnout*^{pf}, *šeptat*^{impf} 'to whisper', *telefonovat*^{biasp} 'to telephone', etc.

In a similar vein as the assertive verbs of communication, the 'Addressee' of these verbs can be obligatory (e.g., *informovat*^{biasp} 'to inform'), optional (e.g., *říci*^{pf}, *říkat*^{impf} 'to say'), or it is not present in the valency frame at all (e.g., *podotknout*^{pf}, *podotýkat*^{impf} 'to remark', *poznamenat*^{pf}, *poznamenávat*^{impf} 'to mention').

On the other hand, the syntactic properties of these verbs of communication vary according to the types of the DCCs which they are complemented by: (i) if they are complemented by an assertDCC, their syntactic behavior corresponds to that of the assertive verbs of communication (see Section 3.1 above), (ii) when they govern an interDCC, they share the syntactic properties with the interrogative verbs of communication (see Section 3.2 above), and (iii) if they are complemented by a directDCC, they exhibit the same syntactic behavior as the directive verbs of communication (see Section 3.3 above).

(i) 'Neutral' Verbs Complemented by an Assertive DCC. If these verbs of communication are complemented by the **assertDCC**, then they allow the splitting of the theme and the dictum,

which is reflected in their valency frames (see Section 3.1.2). See the following examples:

(68) *Řekla nám*.ADDR *o sobě*.PAT, *(že má upřímnou povahu, je veselá a ráda se baví)*.EFF (SYN2006pub)
E. *'(She-)told – us – about – herself – that – (she-)has – frank – character – is – cheerful – and – glad – refl – enjoys.'*

(69) *K boji*.PAT *o následnictví poznamenal, (že nebude "žádný slet supů a shluk hyen")*.EFF (SYN2006pub)
E. *On – contest – for – succession – (he-)remarked – that – won't – no – meeting – vultures – and – riot – hyenas*

(ii) 'Neutral' Verbs Complemented by an Interrogative DCC. Being complemented by an **interDCC**, these verbs have the same syntactic properties as the interrogative verbs of communication. If the 'Addressee's' slot is present in the valency frame, then it is obligatory as in the case of the interrogative verbs of communication. The splitting of the theme and the dictum is not possible in these cases. See the following examples:

(70) *Řekni mi*.ADDR, *(zda je to všechno pravda)*.PAT (SYN2005)
E. *Tell me*.ADDR *(whether it is all true)*.PAT

(71) *Kdosi*.ACT *poznamenal, (zda je to vůbec legální …)*.PAT (SYN2006pub)
E. *Somebody*.ACT *has remarked (whether it is legal …)*.PAT

(iii) 'Neutral' Verbs Complemented by a Directive DCC. When complemented by a **direct-DCC**, they have the similar properties as the directive verbs of communication. If their valency structure contains the 'Addressee's' valency slot, it is obligatory (ex 72). In contrast to the directive verbs of communication, the 'Message' cannot be expressed by an infinitive (ex 73).

(72) *Lupiči*.ACT *řekli prodavačce*.ADDR, *(aby jim vydala peníze)*.PAT (SYN2006pub)
E. *The robbers*.ACT *told the shop assistant*.ADDR *(to give them money out)*.PAT

(73) *Já jí řekl, aby si vzala prášek …* (SYN2000)
E. *'I – her – told – to – refl – took – pill …'*
(**Já jí řekl vzít si prášek …*)
(E. *'I – her – told – take – refl – pill'*)

3.4.1. Valency Frames of the 'Neutral' Cases of Verbs of Communication

Tables 5, 6 and 7 summarize the possible valency frames of the verbs of communication which exhibit 'neutral' behavior with regard to the types of the DCCs. The following three types are distinguished with respect to the 'Addressee' slot: it can be obligatory (Table 5) or optional (Table 6) or it can be missing in the valency frames at all (Table 7).

verb of communication	valency frame
sdělit$^{pf}{}_1$, *sdělovat*$^{impf}{}_1$ 'to tell'	ACT$_1$ ADDR$_3$ PAT$_{4,assertDCC}$
sdělit$^{pf}{}_2$, *sdělovat*$^{impf}{}_2$ 'to tell'	ACT$_1$ ADDR$_3$ PAT$_{k+3,o+6}$ EFF$_{4,assertDCC}$
sdělit$^{pf}{}_3$, *sdělovat*$^{impf}{}_3$ 'to tell'	ACT$_1$ ADDR$_3$ PAT$_{interDCC}$
sdělit$^{pf}{}_4$, *sdělovat*$^{impf}{}_4$ 'to tell'	ACT$_1$ ADDR$_3$ PAT$_{directDCCDCC}$
informovat$^{biasp}{}_1$ 'to inform'	ACT$_1$ ADDR$_4$ PAT$_{o+6,assertDCC}$
informovat$^{biasp}{}_2$ 'to inform'	ACT$_1$ ADDR$_4$ PAT$_{o+6}$ EFF$_{assertDCC}$
informovat$^{biasp}{}_3$ 'to inform'	ACT$_1$ ADDR$_4$ PAT$_{interDCCDCC}$
informovat$^{biasp}{}_4$ 'to inform'	ACT$_1$ ADDR$_4$ PAT$_{directDCC}$
oznámit$^{pf}{}_1$, *oznamovat*$^{impf}{}_1$ 'to announce'	ACT$_1$ ADDR$_3$ PAT$_{4,assertDCC}$
oznámit$^{pf}{}_2$, *oznamovat*$^{impf}{}_2$ 'to announce'	ACT$_1$ ADDR$_3$ PAT$_{o+6,na+4}$ EFF$_{4,assertDCC}$
oznámit$^{pf}{}_3$, *oznamovat*$^{impf}{}_3$ 'to announce'	ACT$_1$ ADDR$_3$ PAT$_{interDCC}$
oznámit$^{pf}{}_4$, *oznamovat*$^{impf}{}_4$ 'to announce'	ACT$_1$ ADDR$_3$ PAT$_{directDCC}$
telefonovat$^{biasp}{}_1$ 'to telephone'	ACT$_1$ ADDR$_3$ PAT$_{4,o+6,assertDCC}$
telefonovat$^{biasp}{}_2$ 'to telephone'	ACT$_1$ ADDR$_3$ PAT$_{o+6}$ EFF$_{4,assertDCC}$
telefonovat$^{biasp}{}_3$ 'to telephone'	ACT$_1$ ADDR$_3$ PAT$_{interDCC}$
telefonovat$^{biasp}{}_4$ 'to telephone'	ACT$_1$ ADDR$_3$ PAT$_{directDCC}$

Table 5. The valency frames of the 'neutral' verbs of communication with an obligatory 'Addressee'.

4. Conclusion

We have described syntactic properties of the Czech verbs of communication. We have given the characteristics of three participants ('Speaker', 'Recipient' and 'Message') of the events that these verbs render. We have provided a description of tectogrammatical counterparts of these participants and their morphemic realizations. A special attention has been devoted to the dependent content clauses. Three types of them are distinguished on the basis of their modality: assertive, interrogative and directive. We have proposed a further subdivision of the group of the verbs of communication with respect to which type of the dependent content clauses these verbs require to be complemented by. These classes are referred to as assertive, interrogative and directive verbs of communication and syntactic properties of the verbs of these three subclasses are described in detail. We have focused on their valency frames and the splitting of the theme and dictum. Furthermore, the verbs of communication which behave 'neutrally' with regard to the types of dependent content clauses, i.e., they can be complemented by more than one type of the dependent content clauses, are debated. As their syntactic properties vary depending on the type of the dependent content clause which they govern, we propose to distinguish four types of valency frames for these verbs: "assertive" without the splitting of the theme and dictum, "assertive" with the splitting of the theme and dictum, "interrogative" and "directive".

verb of communication	valency frame
říci$^{pf}{}_1$, *říkat*$^{impf}{}_1$ 'to say'	ACT$_1$ [ADDR]$_3$ PAT$_{4,o+6,assertDCC}$
říci$^{pf}{}_2$, *říkat*$^{impf}{}_2$ 'to say'	ACT$_1$ [ADDR]$_3$ PAT$_{k+3,na+4,o+6}$ EFF$_{4,assertDCC}$
říci$^{pf}{}_3$, *říkat*$^{impf}{}_3$ 'to say'	ACT$_1$ ADDR$_3$ PAT$_{interDCC}$
říci$^{pf}{}_4$, *říkat*$^{impf}{}_4$ 'to say'	ACT$_1$ ADDR$_3$ PAT$_{o+4,directDCCDCC}$
křiknout$^{pf}{}_1$, *křičet*$^{impf}{}_1$ 'to shout'	ACT$_1$ [ADDR]$_{na+4}$ PAT$_{4,assertDCC}$
křiknout$^{pf}{}_2$, *křičet*$^{impf}{}_2$ 'to shout'	ACT$_1$ [ADDR]$_{na+4}$ PAT$_{o+6}$ EFF$_{4,assertDCC}$
křiknout$^{pf}{}_3$, *křičet*$^{impf}{}_3$ 'to shout'	ACT$_1$ ADDR$_{na+4}$ PAT$_{interDCC}$
křiknout$^{pf}{}_4$, *křičet*$^{impf}{}_4$ 'to shout'	ACT$_1$ ADDR$_{na+4}$ PAT$_{directDCC}$
povědět$^{pf}{}_1$, *povídat*$^{impf}{}_1$ 'to tell'	ACT$_1$ [ADDR]$_3$ PAT$_{4,o+6,assertDCC}$
povědět$^{pf}{}_2$, *povídat*$^{impf}{}_2$ 'to tell'	ACT$_1$ [ADDR]$_3$ PAT$_{k+3,na+4,o+6}$ EFF$_{4,assertDCC}$
povědět$^{pf}{}_3$, *povídat*$^{impf}{}_3$ 'to tell'	ACT$_1$ ADDR$_3$ PAT$_{interDCC}$
povědět$^{pf}{}_4$, *povídat*$^{impf}{}_4$ 'to tell'	ACT$_1$ ADDR$_3$ PAT$_{directDCC}$
psát$^{impf}{}_1$ 'to write'	ACT$_1$ [ADDR]$_3$ PAT$_{4,o+6,assertDCC}$
psát$^{impf}{}_2$ 'to write'	ACT$_1$ [ADDR]$_3$ PAT$_{o+6}$ EFF$_{4,assertDCC}$
psát$^{impf}{}_3$ 'to write'	ACT$_1$ ADDR$_3$ PAT$_{interDCC}$
psát$^{impf}{}_4$ 'to write'	ACT$_1$ ADDR$_3$ PAT$_{directDCC}$
šeptnout$^{pf}{}_1$, *šeptat*$^{impf}{}_1$ 'to whisper'	ACT$_1$ [ADDR]$_3$ PAT$_{4,o+6,assertDCC}$
šeptnout$^{pf}{}_2$, *šeptat*$^{impf}{}_2$ 'to whisper'	ACT$_1$ [ADDR]$_3$ PAT$_{o+6}$ EFF$_{4,assertDCC}$
šeptnout$^{pf}{}_3$, *šeptat*$^{impf}{}_3$ 'to whisper'	ACT$_1$ ADDR$_3$ PAT$_{intertDCC}$
šeptnout$^{pf}{}_4$, *šeptat*$^{impf}{}_4$ 'to whisper'	ACT$_1$ ADDR$_3$ PAT$_{directDCC}$

Table 6. The valency frames of the 'neutral' verbs of communication with an optional 'Addressee'.

Acknowledgments The research reported in this paper is carried under the grants LC536 (Center for Computational Linguistics II) and GA UK 7982/2007.

verb of communication	valency frame
podotknout$^{pf}{}_1$, *podotýkat*$^{impf}{}_1$ 'to remark'	ACT$_1$ PAT$_{4,assertDCC}$
podotknout$^{pf}{}_2$, *podotýkat*$^{impf}{}_2$ 'to remark'	ACT$_1$ PAT$_{k+3,o+6}$ EFF$_{4,assertDCC}$
podotknout$^{pf}{}_3$, *podotýkat*$^{impf}{}_3$ 'to remark'	ACT$_1$ PAT$_{interDCC}$
podotknout$^{pf}{}_4$, *podotýkat*$^{impf}{}_4$ 'to remark'	ACT$_1$ PAT$_{directDCC}$
poznamenat$^{pf}{}_1$, *poznamenávat*$^{impf}{}_1$ 'to remark'	ACT$_1$ PAT$_{4,assertDCC}$
poznamenat$^{pf}{}_2$, *poznamenávat*$^{impf}{}_2$ 'to remark'	ACT$_1$ PAT$_{k+3,o+6}$ EFF$_{4,assertDCC}$
poznamenat$^{pf}{}_3$, *poznamenávat*$^{impf}{}_3$ 'to remark'	ACT$_1$ PAT$_{interDCC}$
poznamenat$^{pf}{}_4$, *poznamenávat*$^{impf}{}_4$ 'to remark'	ACT$_1$ PAT$_{directDCC}$

Table 7. The valency frames of the 'neutral' verbs of communication without an 'Addressee'.

Bibliography

Bauer, Jaroslav. 1965. Souvětí s větami obsahovými. In *SPFFBU*, volume XIV, A 13, pages 55–66.

Běličová, Helena and Jan Sedláček. 1990. *Slovanské souvětí*. Academia, Praha.

Běličová-Křížková, Hana. 1979. Větná modalita a podřadné souvětí. *SaS*, XL, 3:218–231.

Cruse, D.A. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge.

Daneš, František and Zdeněk Hlavsa. 1987. *Větné vzorce v češtině*. Academia, Praha.

Filipec, Jaroslav and František Čermák. 1985. *Česká lexikologie*. Academia, Praha.

Grepl, Miroslav and Petr Karlík. 1998. *Skladba češtiny*. Votobia, Olomouc.

Hajičová, Eva, Jarmila Panevová, and Petr Sgall. 1985-1987. Coreference in the grammar and in the text. *Prague Bulletin of Mathematical Linguistics*, 44,46,48.

Kettnerová-Benešová, Václava. 2007. Modality in dependent content clauses by verbs with imperative features in czech. In *Proceedings of Grammar & Corpora 2007*, Prague, Czech Republic (in print).

Konečná, Dana. 1966. K otázce druhů objektu podle významu. *SlPrag*, 8:311–316.

Lopatková, Markéta, Zdeněk Žabokrtský, and Václava Kettnerová. 2008. *Valenční slovník českých sloves*. Karolinum, Prague.

Markéta Lopatková, Zdeněk Žabokrtský, Karolina Skwarska. 2006. Valency Lexicon of Czech Verbs: Alternation-Based Model. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1728–1733, Genova, Italy.

Mikulová, Marie, Allevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk Žabokrtský. 2005. Annotation on the tectogrammatical layer in the prague dependency treebank, annotation manual. Technical report, Prague.

1986. *Mluvnice češtiny II*. Academia, Praha.

1987. *Mluvnice češtiny III*. Academia, Praha.

Panevová, Jarmila. 1974. On verbal Frames in Functional Generative Description. *Prague Bulletin of Mathematical Linguistics*, 22:3–40.

Panevová, Jarmila. 1975. On verbal Frames in Functional Generative Description. *Prague Bulletin of Mathematical Linguistics*, 23:17–52.

Panevová, Jarmila. 1980. *Formy a funkce ve stavbě české věty*. Academia, Praha.

Panevová, Jarmila. 1996. More Remarks on Control. *Prague Linguistic Circle Papers, John Benjamins*, 2:101–120.

Panevová, Jarmila, Eva Benešová, and Petr Sgall. 1971. *Čas a modalita v češtině*. Universita Karlova, Praha.

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

Součková, Kateřina. 2005. Valence sloves mluvení. (diploma work).

1964. *Slovník spisovného jazyka českého*. Academia, Praha.

2003. *Slovník spisovné češtiny pro školu a veřejnost.* Academia, Praha.

Svozilová, Naďa, Hana Prouzová, and Anna Jirsová. 1997. *Slovesa pro praxi: Valenční slovník nejčastějších českých sloves.* Academia, Praha.

Svozilová, Naďa, Hana Prouzová, and Anna Jirsová. 2005. *Slovník slovesných, substantivních a adjektivních vazeb a spojení.* Academia, Praha.

Žabokrtský, Zdeněk. 2005. *Valency Lexicon of Czech Verbs. (PhD thesis).* Ph.D. thesis, Charles University, Prague, Czech Republic.

Žabortský, Zdeněk and Markéta Lopatková. 2007. Valency information in vallex 2.0: Logical structure of the lexicon. *Prague Bulletin of Mathematical Linguistics*, 87:41–60.

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 90 DECEMBER 2008 109–122

A Case Study in Treebank Collaboration and Comparison: *Accusativus cum Infinitivo* and Subordination in Latin

David Bamman, Marco Passarotti, Gregory Crane

Abstract

We describe here a collaboration between two separate treebank projects annotating data for the same language (Latin). By working together to create a common standard for the annotation of Latin syntax and sharing our annotated data as it is created, we are each able to rely on the resources and expertise of the other while also ensuring that our data will be compatible in the future. This compatibility allows us to conduct diachronic studies involving both datasets, and we add our results to an ongoing discussion of one such issue, the gradual replacement of the *Accusativus cum Infinitivo* construction in Latin with subordinate clauses headed by conjunctions such as *quod* and *quia*.

1. Introduction

Latin has been used as a productive language for over two thousand years. The duration of this lifetime has created enough distinguishable areas of scholarship that a single project is unlikely to build a treebank containing both Vergil's *Aeneid* (written in the first century BCE) and Johannes Kepler's *Astronomia nova* (published in 1609). One reason for this is the unique role that treebanks play within the humanities: while NLP-oriented researchers may build a treebank from newswire for such tasks as training automatic parsers and inducing grammars, traditional humanists are interested in the texts themselves, and will build a treebank consisting entirely of the Bible (for instance) in order to study the specific use of syntax within. We must expect and encourage different research groups to create individual treebanks containing texts from these different eras.

The development of more than one treebank for any given language, however, has the potential to lead to balkanization, with each individual project working independently and pursuing its own research agenda. This diversity is of course necessary for scientific progress, but it can also lead to a proliferation of annotation styles and datasets that are ultimately incompatible. The adoption of common structural standards such as XCES (Ide, Bonhomme, and

Please cite this article as: David Bamman, Marco Passarotti, Gregory Crane, A Case Study in Treebank Collaboration and Comparison: *Accusativus cum Infinitivo* and Subordination in Latin. The Prague Bulletin of Mathematical Linguistics No. 90, 2008, 109–122.

Romary, 2000) and infrastructure (CLARIN, 2007) mitigates this to a certain extent, but true dataset compatibility also extends to the level of the individual syntactic decisions themselves. While such compatibility is not always possible, the benefits of working together are significant. We here present a case study of such a collaboration.

2. The Treebanks

Our two groups are each independently creating a treebank for Latin – the Latin Dependency Treebank (LDT) (Bamman and Crane, 2006, Bamman and Crane, 2007) on works from the Classical era, and the *Index Thomisticus* (IT-TB) (Busa, 1974–1980, Passarotti, 2007) on the works of Thomas Aquinas. The composition of both treebanks is given in Tables 1 and 2.

Date	Author	Words	Sentences
1st c. BCE	Caesar	1,488	71
1st c. BCE	Cicero	5,663	295
1st c. BCE	Sallust	12,391	703
1st c. BCE	Vergil	2,613	178
4th-5th c. CE	Jerome	8,382	405
	Total	30,537	1,652

Table 1. LDT composition.

Date	Author	Words	Sentences
13th c. CE	Aquinas	22,116	1,009
	Total	22,116	1,009

Table 2. IT-TB composition.

These projects are the first of their kind for Latin, so we do not have prior established guidelines to rely on for syntactic annotation. Since we are both working within the theoretical framework of Dependency Grammar, we have each independently based our annotations on that used by the Prague Dependency Treebank (PDT) (Hajič et al., 1999) while tailoring it for Latin via the grammar of Pinkster (Pinkster, 1990). Adopting an annotation style wholesale, however, is easier said than done. Since nearly all Latin available to us is highly stylized, we are constantly confronted with idiosyncratic constructions that could be syntactically annotated in several different ways. These constructions (such as the ablative absolute or the passive periphrastic) are common to Latin of all eras. Rather than have each project decide upon and record each decision for annotating them, we decided to pool our resources and create a single

annotation manual (Bamman et al., 2007) that would govern both treebanks.

3. Annotation Standards

The creation of this common standard has been vital for the evolution of both of our projects. First and most importantly, it ensures that the treebanks we each create will be annotated in the same way. Both of our individual annotation styles have undergone significant revisions in order to converge on a common ground. Early in our collaboration this involved large-scale reassessments – dropping syntactic functions (the LDT, for instance, once had dedicated tags for indirect objects, ablative absolutes, and complements) or changing the representation of entire constructions (e.g., object complements or accusative + infinitives in the IT-TB). Its effects, however, extend well beyond compatibility. Since we are working with dialects of Latin separated by thirteen centuries, this collaboration has allowed us to base our syntactic decisions on a variety of examples from a wider range of texts. Our individual workflows are each independent of the other, but as both projects annotate more data, we each come across sentences that push the limits of our existing annotation standards: here our collaboration begins. After one group identifies a syntactic construction in its data for which the current annotation standards are insufficient, we both search our respective corpora for similar constructions and then come to a common solution by consulting with each other and with outside advisors. Once we come to an agreement on annotation, we include it as part of the guidelines.

The diversity in our projects allows different annotation problems to surface with our individual texts. Two examples can illustrate this.

Ex. 1: Diverse syntactic constructions. Reflexive passives (in which an action is expressed without specifying the agent responsible for it) are much more common in later Latin (Medieval and beyond) than in Classical Latin, but are still present in all eras. In the course of annotating, the IT-TB uncovered eight examples of the reflexive passive in its data, while there were no examples in the LDT. By using the data from the IT-TB, we were able to revise our guidelines in order to codify the annotation and can now refer to that decision whenever we encounter it in our Classical texts.

Ex. 2: Diverse annotator errors. Since our individual annotators are working with different texts, they make different kinds of errors. By expanding our common guidelines to include more detailed descriptions of how to avoid such errors in the future, both groups benefit. For example: early in our development, the annotators for the LDT would frequently vary in their annotation of indirect questions. By focusing especially on this problem and including it in the guidelines' appendix,[1] we are able to refer annotators from both projects to its solution.

Figure 1 presents two sentences annotated under these guidelines, one from each project.

[1] The final section of the annotation guidelines ("How To Annotate Specific Constructions") specifically addresses syntactic problems as they are known in traditional Latin grammars – e.g., "relative clauses," "indirect questions," "the ablative absolute," "accusative + infinitive constructions," etc.

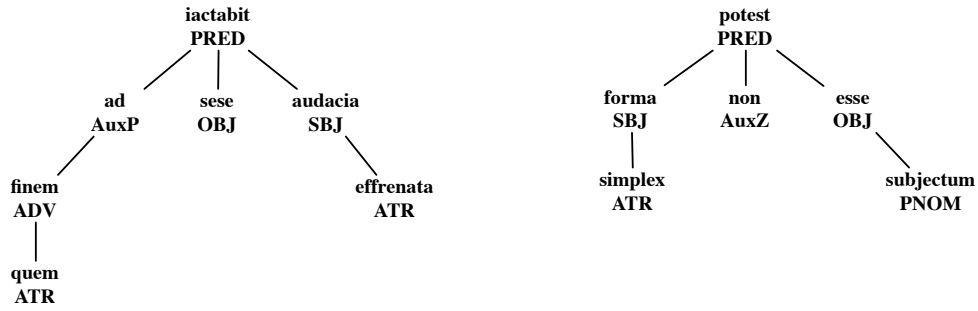


Figure 1. Left: Dependency tree of quem ad finem sese effrenata iactabit audacia *("to what end does your unbridled audacity throw itself?"), Cicero,* Cat. *1.1, from the LDT. Right: Dependency tree of* simplex forma subjectum esse non potest *("the simple form cannot be the subject"), Aquinas,* Super Sententiis Petri Lombardi*, Liber I, Qu. 1, Art. 4, Arg. 1, from the IT-TB.*

4. Differences

While we both adhere to these common standards in all other respects, we do differ in the annotation of a single construction: ellipsis. Since its inception, the LDT has annotated ellipsis in a manner that attempts to preserve the structure of the underlying sentence with a complex syntactic tag, while the IT-TB has followed the PDT convention of attaching an orphan to its head with the relation ExD. This difference can be seen in the differing annotations provided in figure 2.

While the edge labels we assign to these orphans are different, the structure of the tree is not, and our data is still compatible since the formalism used by the LDT can always be reduced to that used by the IT-TB.

5. Data

The data that each of our projects produces plays an important role in our future development, since it can supply the training data we need for automatic syntactic parsing. By at least partially parsing our texts automatically, we can increase the efficiency of our annotators, but statistical dependency parsers such as MaltParser (Nivre et al., 2007) and MSTParser (McDonald et al., 2005) generally perform best with larger amounts of data. By combining our datasets – both annotated under the same general guidelines – we are able to double the size of our training data for such parsers.

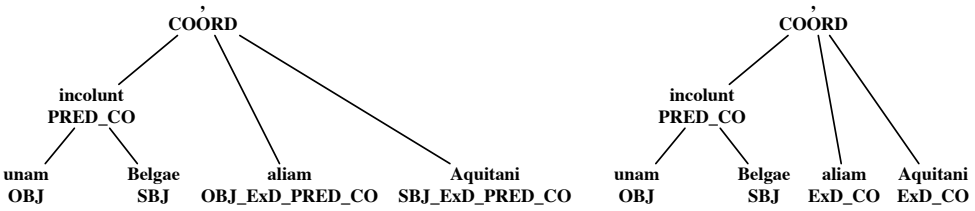


Figure 2. Dependency tree of unam incolunt Belgae, aliam Aquitani *("one the Belgae inhabit, another the Aquintani") (Caes.* B.G. *1.1): on the left is the annotation by the LDT, on the right that by the IT-TB.*

6. Comparison

A sizable body of research has accumulated on the gradual replacement of the Accusativus cum Infinitivo construction in Latin with subordinate clauses headed by the conjunctions *quod* and *quia*. Several studies, such as Mayen (1889), Herman (1963), Wirth-Poelchau (1977) and Cuzzolin (1994), among others, include statistical data gathered by hand about the relative preponderance of one construction over the other in a given time period or within a specific work. Since the texts in our two treebanks are separated in time by thirteen centuries, we are in an excellent position to add our data to this discussion.

6.1. Accusativus cum Infinitivo (ACI)

The Accusativus cum Infinitivo (ACI) in Classical Latin is the primary engine by which indirect discourse is expressed following verbs of saying or thinking (in traditional terms, *verba dicendi vel sentiendi*).[2] While the nominative case is required for subjects of tensed verbs (e.g., sentence 1), in the ACI the subject is expressed in the accusative case and is dependent on an infinitive verb (sentence 2).

(1) tu es contentus ("you are content").

(2) contentum te esse dicebas[3] ("you said that you were content").

In our common manner of annotation, we annotate the ACI (headed by its infinitive verb) as an argument of the verb that introduces it. When that verb is active, the ACI usually depends on it as its object (OBJ), as in figure 3.

The ACI is also found as the subject of impersonal verbs like *oportet* (sentence 3) or with *sum* (sentence 4), in a manner similar to other substantival infinitives.[4]

[2]While the ACI is used most frequently with these two verb classes, it is also found with *verba affectuum* (verbs of feeling) and *verba voluntatis* (verbs of wishing) as well.

[3]Cic. Cat. 1.3 (Perseus:text:1999.02.0010;text=Catil.:Speech=1:chapter=3;num1:dicebas0).

[4]e.g., *Pulchrum est bene facere rei publicae*, Sal. Cat. 3 ("To do well for the republic is good").

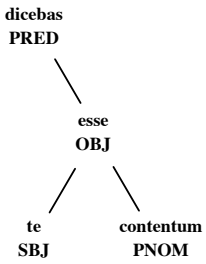


Figure 3. Dependency tree of "contentum te esse dicebas" (Cic. Cat. 1.3)

(3) ergo oportuit materiam illam esse sub forma alicujus quatuor elementorum[5] ("Therefore, that the matter was under the form of some one of the four elements was fitting").

(4) Vos quoque Pergameae iam fas est parcere genti[6] ("That you should also spare the Trojan race is right").

As Pinkster (1990) and Schoof (2003) have pointed out, the term *ACI* is also commonly applied to the arguments of *iubeo* ("to order") and *moneo* ("to warn"), both of which have at least two distinct argument structures containing an accusative noun and an infinitive verb. In the first of these, the accusative noun also fulfills the semantic function of Addressee:

(5) reliquos cum custodibus in aedem Concordiae venire iubet[7] ("he orders the rest to come with the guards into the temple of Concord").

This is not, strictly speaking, an ACI construction because the phrase does not function as a unit if the head verb is made passive: the accusative noun becomes the subject of the passivized verb and assumes the nominative case (resulting in a Nominativus cum Infinitivo construction):

(6) tum pendere poenas Cecropidae iussi[8] ("the Cecrops' children were then ordered to pay the penalties").

In these cases, verbs like *iubeo* and *moneo* require three distinct arguments: a subject, a direct object (semantically the Addressee) and an infinitive complement. We can, however, identify a distinct argument structure involving the ACI when there is no Addressee: here the force is in commanding that a situation come about rather than ordering a specific person to do something:

(7) Caesar portas claudi ... iussit[9] ("Caesar ordered that the gates be closed").

[5]Thomas Aquinas, *Super Sententiis Petri Lombardi* II, Dist. 12, Qu. 1, Art. 4, Arg. 4, 8-8, 10-2.

[6]Verg. Aen. 6.63 (Perseus:text:1999.02.0055;Book=6:card=42;vos0:dardaniae0).

[7]Sal. Cat. 46 (Perseus:text:1999.02.0123;chapter=46;consul0:iubet1).

[8]Verg. Aen. 6.20-21 (Perseus:text:1999.02.0055;Book=6:card=14;tum0:natorum0).

[9]Caes. B.G. 2.32 (Perseus:text:1999.02.0002;Book=2:chapter=32;Sub0:acciperent0).

(8) te interfici iussero[10] ("I will have ordered that you be killed").

This "true" ACI comes about with inanimate objects that cannot be commanded (a door, for instance, cannot be ordered to close) or with passive infinitives, where the order must have a declarative rather than imperative force. Note, however, that all examples of the former variety (e.g., sentence 5) are technically ambiguous since the accusative noun need not always be seen as the Addressee.

6.2. From ACI to the quod/quia clause

While the ACI was the primary method of expressing indirect discourse in Classical Latin, it was gradually replaced over several centuries by subordinate clauses with overt conjunctions (such as *quod* and *quia*), as in sentences 9 and 10.

(9) et vidi quod aperuisset agnus unum de septem signaculis[11] ("And I saw that the lamb had opened one of the seven seals").

(10) quidam enim dicunt, quod anima est composita ex materia et forma[12] ("For some say that the soul is composed out of matter and form").

In this subordinate clause, the subject is in the nominative case rather than accusative and the subordinate verb is inflected, unlike the infinitive found in the ACI. The reason for this movement can be seen as a combination of several other contemporaneous changes in the evolution of the language, such as the movement from SOV word order to SVO and the emergence of the article (Calboli, 1978, Lehmann, 1989, Cuzzolin, 1994) or the loss of case markings, notably the accusative – since an accusative subject is the hallmark of the ACI construction, its absence would favor the use of a different means of expression (Herman, 1989).

Another major explanation for this movement can also be found in the resolution of ambiguity. As Cuzzolin (1991b, 1994) points out, the ACI's use of the infinitive instead of a tensed verb with a mood blocks its communicative modality – whether it represents a statement of fact (indicative) or one of opinion/possibility (subjunctive). The use of the accusative case for both the subject and the direct object of the ACI infinitive verb can also easily give rise to ambiguity. As Herman (1989) notes, while authors would avoid the use of completely ambiguous sentences such as *Petrum Paulum diligere scio* (whose ambiguity borders on ungrammaticality), they would still have to take pains to ensure the meaning is clear in ACI constructions they do employ (e.g., by avoiding the use of two noun phrases of the same semantic category or by providing enough contextual disambiguating information). Subordinate clauses do not contain this ambiguity and are therefore less awkward to use.

These changes led to the gradual replacement of the ACI by subordinate clauses headed by *quod* and *quia* (and eventually the *que* and *che* of modern romance languages). Statistical studies reveal this gradual progression. Mayen (1889), for instance, charts the replacement in

[10] Cic. Cat. 1.5 (Perseus:text:1999.02.0010;text=Catil.:Speech=1:chapter=5;nam0:manus0).

[11] Rev. 6.1 (Perseus:text:1999.02.0060 book=Apocalypse:chapter=6 et0:veni0).

[12] Thomas Aquinas, *Super Sententiis Petri Lombardi* I, Dist. 8, Qu. 5, Art. 2, Solutio, 2-3, 3-6.

terms of the ratio of ACI to *quod*-clauses within various authors: 33:1 in Tertullian (d. ca. 235 CE), 12:1 in Cyprian (d. ca. 258 CE), and 6:1 in Lucifero di Cagliari (d. ca. 370 CE). Herman (1989) notes generally that in the five or six centuries after Petronius, *quod*-clauses are found in about 10% of the places where one could also find an ACI; this number spikes to 15% with Lucifero di Cagliari and 20% in the *Peregrinatio Aetheriae* (ca. 400 CE).

6.3. Methodology

As mentioned above, we annotate the ACI in our treebanks as a self-contained phrase dependent on its introducing verb via SBJ or OBJ depending on that verb's voice (see figure 4). *Quod* and *quia* clauses that function as verbal arguments (as opposed to adverbial clauses translated as "because" or "since") are annotated similarly (see figure 5). Following the PDT, however, we treat the subordinating conjunction as a "bridge" between the embedded and matrix verbs.

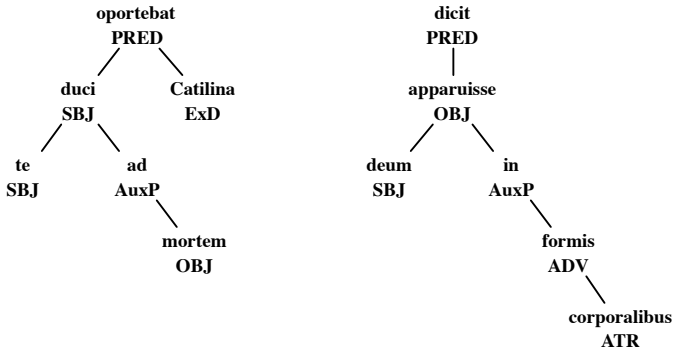


Figure 4. Annotation of ACI constructions. Left: ad mortem te, Catilina, duci ... oportebat *("That you be led to death, Catiline, was fitting"), Cicero,* Cat. *1.1. Right: "dicit deum apparuisse in corporalibus formis" ("he says that god had appeared in bodily forms"), Aquinas,* Super Sententiis Petri Lombardi *II, Dist. 8, Qu. 1, Prologus, 14-1, 14-6.*

The clear value of a treebank is the ease with which we can locate all instances of a particular syntactic phenomenon. Given these tree structures, we can find all instances of the ACI by searching for all infinitive verbs and accusative participles (optionally governing an infinitive of "sum" as an auxiliary in compound verbs) dependent on their heads via an argument relationship (SBJ or OBJ). Since Latin is a pro-drop language, an accusative subject is not required of the infinitive verb in the ACI and so cannot be a necessary criterion for finding it. This search of course also results in a number of prolative infinitives such as those dependent on modals like *possum*, as well as non-ACI "accusative and infinitive" constructions such as those found

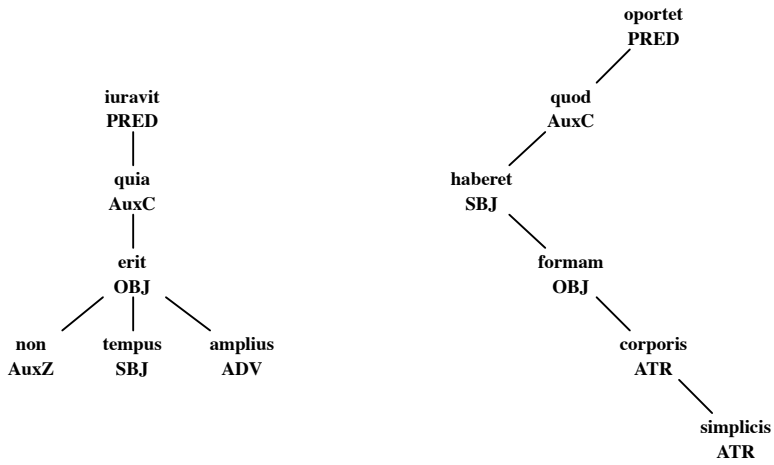


Figure 5. Annotation of quod/quia clauses. Left: iuravit ... quia tempus amplius non erit *("he swore ... that time will not be any longer"), Rev. 10:6. Right:* oportet quod haberet formam corporis simplicis *("That it should have the form of a simple body is fitting"), Aquinas,* Super Sententiis Petri Lombardi *II, Dist. 21, Qu. 1, Art. 4, Arg. 3, 5-6, 6-5.*

with *iubeo* (see sentence 5 above). These are pruned either en masse by head verb (*possum* and *coepio* for instance, never allow the ACI as an object) or by individual inspection.

For *quod* and *quia* clauses, we simply search for all verbs or participles dependent via an argument relation (SBJ or OBJ) on a head that is itself dependent on its head via AuxC (the bridge relationship between embedded and matrix verbs).

6.4. Results

We conducted these searches on three subsections of our treebanks: one for authors of the Classical era of the first century BCE (Caesar, Cicero, Sallust and Vergil), one for Jerome (ca. 400 CE) and one for Thomas Aquinas (ca. 1200 CE). We then grouped the results into two categories, one for verba dicendi and sentiendi[13] and one for impersonal verbs.[14] The results

[13]Since the distinction between a verb of "saying" and "thinking" is often blurry (given the cognitive similarity between the two), we group them into a single class for evaluation. Verba dicendi and sentiendi in our texts include: aio, audio, cerno, certus, cognosco, comperio, conclamo, confido, confirmo, conjuro, constituo, credo, decerno, demonstro, dico, dictito, doceo, dubito, edoceo, existimo, fateor, fero, habeo, hortor, imagino, induco, infitior, instituo, intellego, invenio, judico, juro, loquor, memini, nego, nescio, nuntio, oro, ostendo, polliceor, pono, praedico, propono, puto, respondeo, scio, scribo, sentio, statuo, testor and video.

[14]Impersonal verbs include: accedo, consto, contingo, convenio, debeo, decet, do, intersum, juvo, licet, oportet, placeo, praesto, refero, relinquo, sequor and sum.

are listed in tables 3 and 4.

Author	ACI	Quod/quia clause	Ratio
Classical authors	182	1	99.5%
Jerome	3	9	25.0%
Aquinas	35	80	30.4%

Table 3. verba dicendi and sentiendi.

Author	ACI	Quod/quia clause	Ratio
Classical authors	33	1	97.1%
Jerome	15	0	100%
Aquinas	27	72	27.3%

Table 4. impersonal verbs.

We can see here the process of language change in action. As other authors have noted, the replacement of the ACI construction by *quod* and *quia* subordinate clauses is progressive. While Cuzzolin (1991b, 1994) suggests that the progress within verba dicendi and sentiendi was tied with the assertiveness of the introducing verb (whether it is strongly or weakly assertive), we can see here that the progress applies to other ACI constructions as well. In the 5th century (with Jerome), the ACI construction following verba dicendi and sentiendi was in the process of being replaced by *quod* and *quia*,[15] but it is still dominant in impersonal constructions – it is only with Aquinas much later that we see a strong indication of tensed subordinate clauses being used here as well.

Our results also confirm Herman's (1989) observation concerning the placement of *quod* and *quia* clauses with respect to their governor. Herman notes that in four Christian authors of the 3rd to 5th centuries CE, the ACI construction has much more freedom of placement than tensed subordinate clauses, occuring with relatively equal frequency to the left or right of its head verb.[16] *Quod* and *quia* clauses, however, are much less free, occurring in almost all instances after their head verb.[17] When considering the same instances that provided the figures in tables 3 and 4, we find the following distribution.

[15]It is also interesting to note that the two of the three uses of the ACI following verba sentiendi in Jerome (Rev. 2:9 and Rev. 3:9) are identical – *qui dicunt se Iudaeos esse et non sunt* ("who say that they are Jews and are not"), which may suggest a common source.

[16]Herman reports 55 instances of the ACI after the verb in Cyprian compared to 45 before, 44/56 in Lucifero di Cagliari, 56/44 in the *Peregrinatio* and 40/60 in Salvien.

[17]98 instances after the verb in Cyprian compared to 2 before, 95/5 in Lucifero di Cagliari, 100/0 in the *Peregrinatio* and 100/0 in Salvien.

Author	Before verb	After verb	Ratio
Classical authors	100	115	46.5%
Jerome	0	18	0%
Aquinas	2	60	3.2%

Table 5. Position of ACI constructions with respect to their head verb.

Author	Before verb	After verb	Ratio
Classical authors	1	1	50.0%
Jerome	0	9	0%
Aquinas	1	151	0.7%

Table 6. Position of quod and quia clauses with respect to their head verb.

In Classical authors, the ACI occurs with relatively equal frequency before and after its head verb. With Jerome and Aquinas, however, we can see a movement toward a post-verbal position for both types of subordination: not only do *quod* and *quia* clauses almost always occur after the verb that governs them (as in the case in the four Christian authors studied by Herman), but the ACI construction now also does as well. Given the late period in which both of these authors are writing, we can likely attribute this not only to a stylistic avoidance of *quod* and *quia* clauses before the verb (which, as Herman notes, would be understood as causal), but to a typological difference between SOV word order in Classical Latin and the later SVO.

7. Transparency

The reproducibility of experiments lies at the cornerstone of the scientific method, but philological studies often leave out the information that allows others to investigate their claims – not only the specific works (and textual editions) on which they are based, but the sentence-level annotations themselves that give rise to reported statistics. In his study of the ratio of ACI to *quod* clauses following *verba affectuum*, P. Cuzzolin (1991a) summarizes Raphael Kühner's work on the subject in the great Kühner-Stegmann reference grammar (1914):

> Kühner himself reported the number of passages he counted: "So hat nach meiner Zählung bei *doleo* 57 Stellen mit *Acc. c. Inf.* gegen 4 *quod*, bei *miror* 110 gegen 8, bei *glorior* 19 gegen 2, bei *queror* 71 gegen 15, bei *gaudeo* 84 gegen 9 usw." (1914:77), although it is difficult to say what he meant by the word "Stelle" and impossible to say which texts his counting is based upon.

Both treebanks used in this study are publicly available.[18] The impact of this transparency is twofold: first, it allows others to verify our results (and also conduct their own inquiries to consider or eliminate other variables not examined here); and second, it lets others make use of the results of our labor in whatever ways they see fit (thereby avoiding duplicated efforts in the future). Our data is not simply a tally of ACI constructions and *quod/quia* clauses in our authors, but a corpus in which the syntactic relationship for every word in a sentence is annotated (and from which these constructions – as well as many others – can be extracted). By sharing this data, we hope to pave the way for a number of future inquiries (both by ourselves and others), well beyond the scope of this single research question.

8. Future

Collaborating has allowed both of our projects to accomplish more than if we each worked alone, both in terms of creating our respective treebanks and in the varieties of research we can subsequently pursue with them. This type of collaboration lays the foundation for a more distributed method of treebank building, with contributions from a decentralized audience around the world. By creating a communal standard for the annotation of Latin syntax and making our data freely available, we hope to encourage other research groups working in different eras of Latin to collaborate with us. Classical philology has long been a science of counting; by annotating our texts only once and sharing our data, we avoid unnecessarily duplicating our efforts and simultaneously promote a level of transparency that can only be healthy for the discipline as a whole.

9. Acknowledgments

Grants from the Digital Library Initiative Phrase 2 (IIS-9817484) and the National Science Foundation (BCS-0616521) provided support for this work.

Bibliography

Bamman, David and Gregory Crane. 2006. The design and use of a Latin dependency treebank. In *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT2006)*, pages 67–78, Prague. ÚFAL MFF UK.

Bamman, David and Gregory Crane. 2007. The Latin Dependency Treebank in a cultural heritage digital library. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, pages 33–40, Prague. Association for Computational Linguistics.

Bamman, David, Marco Passarotti, Gregory Crane, and Savina Raynaud. 2007. Guidelines for the syntactic annotation of Latin treebanks, version 1.3. Technical report, Tufts Digital Library, Medford, http://nlp.perseus.tufts.edu/syntax/treebank/1.3/docs/guidelines.pdf.

[18]The LDT data can be found online at http://nlp.perseus.tufts.edu/syntax/treebank, and the IT-TB data can be found at http://gircse.marginalia.it/~passarotti.

Busa, Roberto. 1974–1980. *Index Thomisticus : sancti Thomae Aquinatis operum omnium indices et concordantiae, in quibus verborum omnium et singulorum formae et lemmata cum suis frequentiis et contextibus variis modis referuntur quaeque / consociata plurium opera atque electronico IBM automato usus digessit Robertus Busa SI.* Frommann-Holzboog, Stuttgart-Bad Cannstatt.

Calboli, Gualtiero. 1978. Die Entwicklung der klassischen Sprachen und die Beziehung zwischen Satzbau, Wortstellung und Artikel. *Indogermanische Forschungen*, 83:197–261.

CLARIN. 2007. http://www.mpi.nl/clarin/.

Cuzzolin, Pierluigi. 1991a. On sentential complementation after *verba affectuum*. In Jozsef Herman, editor, *Linguistic Studies on Latin*. Benjamins, Amsterdam-Philadelphia, pages 167–178.

Cuzzolin, Pierluigi. 1991b. Sulle prime attestazioni del tipo sintattico "dicere quod". *Archivio Glottologico Italiano*, 76(1):26–78.

Cuzzolin, Pierluigi. 1994. *Sull'origine della construzione dicere quod: aspetti sintattici e semantici.* La Nuove Italia, Florence.

Hajič, Jan, Jarmila. Panevová, Eva Buráňová, Zdenka Urešová, and Alla Bémová. 1999. Annotations at analytical level: Instructions for annotators (English translation by Z. Kirschner). Technical report, ÚFAL MFF UK, Prague, Czech Republic.

Herman, Jozsef. 1963. *La formation du système roman des conjonctions de subordination.* Akademie Verlag, Berlin.

Herman, Jozsef. 1989. Accusativus cum infinitivo et subordonée à quod, quia en latin tardif. In Gualtiero Calboli, editor, *Subordination and Other Topics in Latin. Proceedings of the Third Colloquium on Latin Linguistics, Bologna, 1-5 April 1985*. Benjamins, Amsterdam-Philadelphia, pages 133–152.

Ide, Nancy, Patrice Bonhomme, and Laurent Romary. 2000. XCES: An XML-based encoding standard for linguistic corpora. In *Proceedings of the Second Language Resources and Evaluation Conference (LREC)*, pages 825–830, Athens.

Kühner, R. and C. Stegmann. 1914. *Ausführliche Grammatik der lateinsichen Sprache II. Satzlehre. I. Teile Zweite Auflage.* Hahnsche Buchhandlung, Hannover.

Lehmann, C. 1989. Latin subordination in typological perspective. In Gualtiero Calboli, editor, *Subordination and Other Topics in Latin. Proceedings of the Third Colloquium on Latin Linguistics, Bologna, 1-5 April 1985*. Benjamins, Amsterdam-Philadelphia, pages 153–179.

Mayen, Georg. 1889. *De particulis quod, quia, quoniam, quomodo ut pro acc. cum infinitivo post verba sentiendi et declarandi positis.* H. Fiencke, Kiel.

McDonald, Ryan, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 523–530.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Passarotti, Marco. 2007. Verso il Lessico Tomistico Biculturale. La treebank dell'Index Thomisticus. In Petrilli Raffaella and Femia Diego, editors, *Il filo del discorso. Intrecci testuali, articolazioni linguistiche, composizioni logiche. Atti del XIII Congresso Nazionale della Società di Filosofia del Linguaggio, Viterbo, Settembre 2006*, pages 187–205. Roma, Aracne Editrice, Pubblicazioni della Società di Filosofia del Linguaggio.

Pinkster, Harm. 1990. *Latin Syntax and Semantics*. Routledge, London.

Schoof, Susanne. 2003. Impersonal and personal passiviation of Latin infinitive constructions: A scrutiny of the structures called AcI. In Jong-Bok Kim and Stephen Wechsler, editors, *Proceedings of the Ninth International Conference on HPSG*. CSLI Publications, Stanford, pages 293–312.

Wirth-Poelchau, L. 1977. *AcI und quod-Satz im lateinischen Sprachgebrauch mitteralterlicher und humanistischer Autoren*. Erlangen-Nürnberg.

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 90 DECEMBER 2008 123–128

REVIEWS

Key Thinkers in Linguistics and the Philosophy of Language

Siobhan Chapman, Christopher Routledge (eds.)

Oxford: Oxford University Press, 2005, xii+282pp.
ISBN-13:978-0-19-518767-0

Reviewed by Jun Qian, Peking University

This dictionary-like book comes as a useful reference book for people interested in languages sciences. There are eighty entries alphabetically arranged in terms of the linguists' or philosophers' surnames. Each entry consists of three parts, with an introduction to the thinker's essential ideas as the main body, followed by "Primary works" by that thinker, which is in turn followed by "Further reading". The eighty articles come from thirty contributors (the two editors included). And the eighty 'key thinkers" are as follows:

Aristotle (384–322 BC), Antoine Arnauld (1612–1694), John Langshaw Austin (1911–1960), Alfred Jules Ayer (1910–1989), Mikhail Mikhailovich Bakhtin (1895–1975), Roland Barthes (1915–1980), Émile Benveniste (1902–1976), George Berkeley (1685–1753), Basil Bernstein (1924–2000), Leonard Bloomfiled (1887–1949), Franz Boas (1858–1942), Franz Bopp (1791–1867), Pierre Bourdieu (1930–2002), Karl Brugmann (1849–1919), Deborah Cameron (b. 1958), Rudolf Carnap (1891–1970), Noam Chomsky (b. 1928), Donald Davidson (1917–2003), Jacques Derrida (1930–2004), Rene Descartes (1596–1650), Michael Dummett (b. 1925), John Rupert Firth (1890–1960), Jerry Fodor (b. 1935), Gottlob Frege (1848–1925), Peter Geach (b. 1916), Nelson Goodman (1906–1998), Joseph Greenberg (1915–2002), Algirdas Greimas (1917–1992), Herbert Paul Grice (1913–1988), Jacob Grimm (1785–1863), Michael Halliday (b. 1925), Georg Hegel (1770–1831), Louis Hjelmslev (1899–1965), Charles Hockett (1916–2000), Wilhelm von Humboldt (1767–1835), David Hume (1711–1777), Edmund Husserl (1859–1938), Roman Jakobson (1896–1982), Daniel Jones (1881–1967), Immanuel Kant (1724–1804), Jerrold Jakob Katz (1932–2002), Saul Kripke (b. 1940), Julia Kristeva (b. 1941), William Labov (b. 1927), Jacques Lacan (1901–1981), Gott-

fried Wilhelm Leibniz (1646–1716), David Lewis (1941–2001), John Locke (1632–1704), Bronislaw Malinowski (1884–1942), Andre Martinet (1908–1999), Karl Marx (1818–1883), John Stuart Mill (1806–1873), Lesley Milroy (b. 1944), Richard Montague (1930–1971), George Edward Moore (1873–1958), Charles Morris (1901–1979), Charles Santiago Sanders Peirce (1839–1914), Jean Piaget (1896–1980), Kenneth Pike (1912–2000), Plato (427–347 BC), Karl Popper (1902–1994), Hilary Putnam (b. 1926), Willard Van Orman Quine (1908–2000), Frank Plumpton Ramsey (1903–1930), Rasmus Rask (1787–1832), Bertrand Russell (1872–1970), Gilbert Ryle (1900–1976), Harvey Sacks (1935–1975), Edward Sapir (1884–1939), Ferdinand de Saussure (1857–1913), John Searle (b. 1932), John Sinclair (b. 1933), Burrhus Frederic Skinner (1904–1990), Peter Frederick Strawson (b. 1918), Deborah Tannen (b. 1945), Alfred Tarski (1902–1983), Tzvetan Todorov (b. 1939), Nikolai Sergeevich Trubetzkoy (1890–1938), Benjamin Lee Whorf (1897–1941), Ludwig Wittgenstein (1889–1951). It is not the intention of this review to validate or invalidate the inclusion of any one of those eighty scholars as a noteworthy thinker because given the same opportunity we probably will come up with quite varied name lists of key thinkers, as determined by our academic background, research interests, knowledge of the vast field of language sciences, perspectives and foci, orientations and purposes, among other factors (cf. Sebeok 1966). However, there are certain issues that this book leads to that cannot go unnoticed, for example, what are the criteria of a "key thinker"? How is a philosopher's influence on and relevance to linguistics determined and evaluated?

Linguists themselves, either theoretical or applied, are thinkers, as attested to somehow by Johann Wolfgang von Goethe's (1749–1832) remarks "Jeder Mensch, weil er spricht, glaubt über die Sprache sprechen zu konnen."(Language Vol. 15, 1939:123) The distinction between them, therefore, is not necessarily that between a thinker-linguist and a non-thinker-linguist. A key thinker-linguist is presumably a linguist who distinguishes himself/herself in terms of an original idea, a novel approach, or a unique theoretical framework and this idea, or approach, or framework of his/hers has either local (i.e. in one subfield, say, morphology or phonology) or global influence in the field of language sciences. However, if this standard is applied or adopted, many contemporary linguists missing from the above list become qualified, e.g. what about Jan Firbas (1920–2000), whose concept of communicative dynamism and the theory of functional sentence perspective are well-known (Firbas 1992)? What about František Daneš's concept of thematic progression and his theory of three-level approach to syntax (Daneš 1964, 1974)? What about Petr Sgall and his group's work on the theory of functional general description (Sgall 1967, 2006; Sgall et al. 1986)? And what about Susumu Kuno's concept of empathy and his theory of functional syntax (1976, 1987, Kuno & Etsuko Kaburaki 1977)?

Besides, one could not but wonder why Jan Baudouin de Courtenay (1845–1929), Henry Sweet (1845–1912), Otto Jespersen (1860–1943), and Vilém Mathesius (1882–1946) are not counted as "key thinkers" and missing from the list (all of them are included in Sebeok 1966; cf. Jakobson 1929, 1966; Hjelmslev 1942–1943; Haislund 1943; Trnka 1946; Wrenn 1946). Jan Baudouin de Courtenay, founder of Kazan School of Linguistics, a forerunner of structural linguistics, is known for his theory of the phoneme and phonetic alternations (1895; Stankiewicz 1972). Henry Sweet was representative of "The English School of Phonetics" as well as a distinguished grammarian (Sweet 1877, 1892, 1898, 1906, 1913; Firth 1946). Otto Jespersen's

contributions touch almost every area of linguistics one can think of at the time, phonetics, morphology, syntax, history of language, philosophy of grammar, language teaching, and artificial language (Jespesen 1904, 1909–1949, 1913, 1924, 1928). Vilém Mathesius, founder of the Prague School of Linguistics, a forerunner of structural-functional linguistics, is noted for his theory and practice of linguistic characterology (1928, 1975).

The missing of these great names from this book makes one wonder if the development of linguistics is really that fast that the predecessors' work quickly becomes outdated, or rather if it is characteristic of present-day linguistic practitioners to be oblivious of their predecessors' work.

There are two suggestions for the improvement. First, historical accuracy should be attended to. There is an inaccurate account on page 140, "Jakobson moved to Prague in 1920, where he and Trubetzkoy co-founded the Prague School of Linguistics in 1926." The fact was it was Mathesius and Jakobson who co-founded the Prague Linguistic Circle (Cercle Linguistique de Prague, cf. Mathesius 1936;Vachek 1966; Toman 1995). Second, English translations, if any, should be referred to for writings in languages other than English. For example, on page 227, *Undersögelse om det gamle Nordiske eller Islandske Sprogs Oprindelse* (1818) is listed as one of the two primary works by Rasmus Rask. It would be much more preferable to include its English translation right after the Danish title, i.e. *Investigation of the Origin of the Old Norse or Icelandic Language*, by Rasmus Kristian Rask; translated by Niels Ege. Copenhagen: The Linguistic Circle of Copenhagen, 1993. And for those who can read Danish but have no access to the original, e-resources, if any, should be referred to, e.g. the digitized Rask 1818 edition is available at `http://books.google.com/`. The same holds true for Franz Bopp's 1833–1852 (p. 43), among others.

References

Baudouin de Courtenay, Jan. 1895. *Versuch einer theorie phonetischer alternationen: Ein Capital aus der Psychophonetik.* Strassburg: K.J. Trübner.

Bopp, Franz. 1833–1852. *Vergleichende Grammatik des Sanskrit, Zend, Griechischen, Lateinischen, Litthauischen, Gothischen und Deutschen.* 2 volumes. Berlin: F. Dümmler. (*A Comparative Grammar of the Sanskrit, Zend, Greek, Latin, Lithuanian, Gothic, German, and Sclavonic Languages*, by Franz Bopp; translated by Edward Backhouse Eastwick. Vol. I, second edition, London: John Murray, 1854. Vol. II, second edition, third edition, London: Williams and Morgate, 1856, 1862. Reprinted by Hildesheim; New York: G. Olms, 1985. Digitized English translation of Volume I, second edition 1854, Volume II, second edition 1856, third edition 1862, at http://books.google.com/)

Daneš, František. 1964. A Three-Level Approach to Syntax. *Travaux Linguistiques de Prague* 1, 225–240.

Daneš, František. 1974. Functional Sentence Perspective and the Organization of the Text. In Daneš, František (ed.), *Papers on Functional Sentence Perspective.* Prague: Academia.1974, 106–128.

Firbas, Jan. 1992. *Functional Sentence Perspective in Written and Spoken Communication.* Cambridge: Cambridge University Press.

Firth, John R. 1946. The English School of Phonetics. *Transactions of the Philological Society* 1946, 92–132. In Firth, John, *Papers in Linguistics 1934–1951*. London: Oxford University Press. 1957, 92–120.

Haislund, Niels. 1943. Otto Jespersen. *Englische Studien* 75, 273–283. Reprinted in Sebeok, Thomas A. (ed.), *Portraits of Linguists. Volume Two.* 1966, 148–157.

Hjelmslev, Louis. 1942–1943. Nécrologie Otto Jespersen. *Acta Linguistica* 3, 119–130. Reprinted in Sebeok, Thomas A. (ed.), *Portraits of Linguists. Volume Two.* 1966, 158–173.

Jakobson, Roman. 1929. Jan Baudouin de Courtenay. *Slavische Rundschau* 1, 809–812. Included in Sebeok, Thomas A. (ed.), *Portraits of Linguists. Volume One.* 1966, 533–537. Reprinted in *Selected Writings II: Word and Language*, 389–393.

Jakobson, Roman. 1966. Henry Sweet's Path toward Phonemics. In Bazell, C. E., J. C. Catford, M. A. K. Halliday, R. H. Robins (eds.), *In Memory of J. R. Firth.* London: Longmans. 1966, 242–254. Reprinted in *Selected Writings II: Word and Language*, 456–467.

Jespersen, Otto. 1904. *How to Teach a Foreign Language.* Tr. from the Danish original by Sophia Yhlen-Olsen Bertelsen. New York: Macmillan.

Jespersen, Otto. 1909–1949. *Modern English Grammar on Historical Principles* (7 vols.) London: Allen & Unwin, Copenhagen: Einar Munksgaard.

Jespersen, Otto. 1913. *Lehrbuch der Phonetik.* 2. Aufl. Leipzig, Berlin: B. G. Teubner. (digitized version at http://nrs.harvard.edu/urn-3:HUL.FIG:003161311; 1st edition 1904)

Jespersen, Otto. 1924. *The Philosophy of Grammar.* London: Allen and Unwin.

Jespersen, Otto. 1928. *An International Language.* London, Allen and Unwin.

Kuno, Susumu. 1976. Three Perspectives in the Functional Approach to Syntax. In Matejka, Ladislav (ed.), *Sound, Sign and Meaning: Qinquagenary of the Prague Linguistic Circle.* Ann Arbor: Department of Slavic Languages and Literatures, University of Michigan. 1976, 119–190.

Kuno, Susumu. 1987. *Functional Syntax: Anaphora, Discourse and Empathy.* Chicago: University of Chicago Press.

Kuno, Susumu and Etsuko Kaburaki. 1977. Empathy and Syntax. *Linguistic Inquiry* 8, 627–672.

Mathesius, Vilém. 1928. On Linguistic Characterology of Modern English. *Actes du Premier Congrès International de Linguists à la Haye du 10–15 avril 1928.* Leiden: A.W. Sijthoff, 56–63. Reprinted in Vachek, Josef (ed.), *A Prague School Reader in Linguistics.* Bloomington: Indiana University Press, 1964, 59–67.

Mathesius, Vilém. 1936. Deset let Pražského lingvistického kroužku (Ten Years of the Prague Linguistic Circle). *Slovo a Slovesnost* (The Word and Verbal Art) 2, 137–145. English translation in Vachek, Josef, *The Linguistic School of Prague.* Bloomington, Indiana: Indiana University Press. 1966, 137–151.

Mathesius, Vilém. 1975. *A Functional Analysis of Present-Day English on a General Linguistic Basis.* Edited by Josef Vachek; translated by Libuše Dušková. The Hague: Mouton; Prague: Academia.

Sebeok, Thomas A. (ed.) 1966. *Portraits of Linguists: A Biographical Source Book for the History of Western Linguistics, 1746–1963. Volume One: From Sir William Jones to Karl Brugmann.* Bloomington and London: Indiana University Press.

Sebeok, Thomas A. 1966. *Portraits of Linguists: A Biographical Source Book for the History of Western Linguistics, 1746–1963. Volume Two: From Eduard Sievers to Benjamin Lee Whorf.* Bloomington and London: Indiana University Press.

Sgall, Petr. 1967. Functional Sentence Perspective in a Generative Description. *Prague Bulletin of Mathematical Linguistic* 2, 203–225.

Sgall, Petr. 2006. *Language in its Multifarious Aspects.* Charles University in Prague: The Karolinum Press.

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

Stankiewicz, Edward. 1972. *A Baudouin de Courtenay Anthology: The Beginnings of Structural Linguistics.* Translated and edited with an introduction by Edward Stankiewicz. Indiana: Indiana University Press.

Sweet, Henry. 1877. *A Handbook of Phonetics.* Oxford: Clarendon Press. (Digitized 1877 edition at http://books.google.com/)

Sweet, Henry. 1892. *A New English Grammar, Logical and Historical. Part I Introduction, Phonology, and Accidence.* Oxford: The Clarendon Press.

Sweet, Henry. 1898. *A New English Grammar, Logical and Historical. Part II — Syntax.* Oxford: The Clarendon Press. (Digitized edition at http://books.google.com/)

Sweet, Henry. 1906. *A Primer of Phonetics.* 3rd edition, revised. Oxford: Clarendon Press. (1st edition 1890, 2nd edition 1902. Digitized 1892 edition at http://books.google.com/)

Sweet, Henry. 1913. *Collected Papers.* Arranged by H. C. Wyld. Oxford: The Clarendon Press.

Toman, Jindřich. 1995. *The Magic of a Common Language: Jakobson, Mathesius, Trubetzkoy, and the Prague Linguistic Circle.* Cambridge, Mass.: MIT Press.

Trnka, Bohumil. 1946. Vilém Mathesius. *Časopis pro moderni filologii* (Journal for Modern Philology) 29, 3–13. Translated by Vladimir Honsa. Included in Sebeok, Thomas A. (ed.), *Portraits of Linguists. Volume Two.*1966, 474–489.

Vachek, Josef. 1966. *The Linguistic School of Prague.* Bloomington, Indiana: Indiana University Press.

Wrenn, Charles L. 1946. Henry Sweet. *Transactions of the Philological Society* 46, 177–201. Reprinted in Sebeok Thomas A. (ed.), *Portraits of Linguists. Volume One.* 1966, 512–532.

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 90 DECEMBER 2008 129–130

BOOK NOTICES

Speech and Language Processing (2nd Edition)

Daniel Jurafsky, James H. Martin

Prentice Hall Series in Artificial Intelligence, New Jersey, 2008, xxxiii+988pp.
ISBN-13:978-0-13-187321-6

Notice by Pavel Schlesinger

The second edition of the presented book was released after 8 years after the first one. Since that time the title itself has earned respect and appreciation within the comunity of NLP researchers and students. For those with serious interesest in the field the book has become famous and it's a must.

The question for this notice is a bit different: *Is the second edition just a standard reprint or do we really obtain something new comparing with the first edition?* The number of main chapters has increased from four (I Words, II Syntax, III Semantics, IV Pragmatics) to five (I Words, II Speech, III Syntax, IV Semantics and Pragmatics, V Applications). Beyond changes in a structure there is a major benefit in adding the fifth Applications chapter, where you can find a guided introduction to topics and task solved nowadays in NLP, e.g. Information extraction (specially Named Entity Recognition), Question answering, Summarization, Dialog systems and Machine translation. The second major improvement is concerned with the text of sections retained from the first edition. It was totally revised and references were updated up to 2008. The combination of additions and updates makes me answer *YES* to my question from above. The second edition of the text is worthy even for already-owners. The subtitle of the book *An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition* is completely true, if understated. There is no better general introduction to NLP and stat-of-art reference book, except perhaps for Chris Manning's and Hinrich Schütze's *Foundations of Statistical Natural Language Processing (MIT Press, 1999).*

We are planning to release a full book review in the near future.

The Prague Bulletin of Mathematical Linguistics
NUMBER 90 DECEMBER 2008

INSTRUCTIONS FOR AUTHORS

Manuscripts are welcome provided that they have not yet been published elsewhere and that they bring some interesting and new insights contributing to the broad field of computational linguistics in any of its aspects, or of linguistic theory. The submitted articles may be:

- long articles with completed, wide-impact research results both theoretical and practical, and/or new formalisms for linguistic analysis and their implementation and application on linguistic data sets, or
- short or long articles that are abstracts or extracts of Master's and PhD thesis, with the most intersting and/or promising results described. Also
- short or long articles looking forward that base their views on proper and deep analysis of the current situation in various subjects within the field are invited, as well as
- short articles about current advanced research of both theoretical and applied nature, with very specific (and perhaps narrow, but well-defined) target goal in all areas of language and speech processing, to give the opportunity to junior researchers to publish as soon as possible;
- short articles that contain contraversing, polemic or otherwise unusual views, supported but some experimental evidence but not necessarily evaluated in the usual sense are also welcome.

The recommended length of long article is 12–30 pages and of short paper is 6-15 pages.

The copyright of papers accepted for publication remains with the author. The editors reserve the right to make editorial revisions but these revisions and changes have to be approved by the author(s). Book reviews and short book notices are also appreciated.

The manuscripts are reviewed by 2 independent reviewers, at least one of them being a member of the international Editorial Board.

Authors receive two copies of the relevant issue of the PBML together with 10 offprints of their article.

The guidelines for the technical shape of the contributions are found on the web site `http://ufal.mff.cuni.cz/pbml.html`. If there are any technical problems, please contact the editorial staff at `pbml@ufal.mff.cuni.cz`.

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 90 DECEMBER 2008

LIST OF AUTHORS

David Bamman
Tufts University
The Perseus Project
Medford, MA 02155
USA
david.bamman@tufts.edu

Ondřej Bojar
Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
bojar@ufal.mff.cuni.cz

Silvie Cinková
Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
cinkova@ufal.mff.cuni.cz

Gregory Crane
Tufts University
The Perseus Project
Medford, MA 02155
USA
gregory.crane@tufts.edu

Václava Kettnerová
Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
kettnerova@ufal.mff.cuni.cz

Jiří Mírovský
Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
mirovsky@ufal.mff.cuni.cz

Václav Novák
Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
novak@ufal.mff.cuni.cz

Marco Passarotti
Catholic University of the Sacred Heart
Department of Philosophy
Milan
Italy

Jan Ptáček
Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
ptacek@ufal.mff.cuni.cz

Jun Qian
English Department
Peking University
Beijing 100871, P.R. China
junqian@pku.edu.cn

Pavel Schlesinger
Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
schlesinger@ufal.mff.cuni.cz

Ivan Šmilauer
LALIC-CERTAL
Institut National des Langues et Civilisations Orientales
104 Quai de Clichy
92110 Clichy-sur-Seine, France
smilauer@cetlef.fr