The Prague Bulletin of Mathematical Linguistics
NUMBER 88 DECEMBER 2007

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 88 DECEMBER 2007

CONTENTS

Articles

Notes

Reviews

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 88 DECEMBER 2007 5–30

Functional Arabic Morphology

Dissertation Summary

Otakar Smrž

Abstract

This is a summary of the author's PhD dissertation defended on September 17, 2007 at the Faculty of Mathematics and Physics, Charles University in Prague. The results comprised in the thesis were obtained within the author's doctoral studies in Mathematical Linguistics during the years 2001–2007. The complete dissertation is available via `http://sourceforge.net/projects/elixir-fm/`.

1. Introduction

Functional Arabic Morphology is a formulation of the Arabic inflectional system seeking the working interface between morphology and syntax. ElixirFM is its high-level implementation that reuses and extends the Functional Morphology library for Haskell (Forsberg and Ranta, 2004), yet the treatment of the language-specific issues constitutes our original work.

In the thesis (Smrž, 2007), we develop a computational model of the morphological processes in Arabic. With this system, we are able to derive and inflect words, as well as to analyze the structure of word forms and to recognize their grammatical functions.

The approach to building our morphological model strives to be comprehensive with respect to linguistic generalization, and high-level and modern with respect to the programming techniques that we employ. We describe the linguistic concept and try to implement it in a very similar, yet abstract way, using the declarative functional programming language Haskell. We emphasize the flexibility of our system, its reusability and extensibility.

1.1. Morphological Models

One can observe several different streams both in the computational and the purely linguistic modeling of morphology. Some are motivated by the need to analyze word forms as to

Please cite this article as: Otakar Smrž, Functional Arabic Morphology: Dissertation Summary. The Prague Bulletin of Mathematical Linguistics No. 88, 2007, 5–30.

their compositional structure, others consider word inflection as being driven by the underlying system of the language and the formal requirements of its grammar.

There are substantial discrepancies between the grammatical descriptions of Arabic represented e.g. by (Fischer, 2001) or (Holes, 2004), and the information that the available morphological computational systems provide. One of the reasons is that there is never a complete consensus on what the grammatical description should be. The other source of the incompatibility lies in the observation that many implementations overlook the principal difference between the function and the form of a linguistic symbol.

Many of the computational models of Arabic morphology, including in particular (Beesley, 2001), (Ramsay and Mansur, 2001) or (Buckwalter, 2002), are *lexical* in nature, i.e. they tend to treat inflectional affixes just like full-fledged lexical words. As they are not designed in connection with any syntax–morphology interface, their interpretations are destined to be *incremental*. That means that the only clue for discovering the morphosyntactic properties of a word is through the explicit affixes and their prototypical functions.

Some signs of a *lexical–realizational* system can be found in (Habash, 2004). The author mentions and fixes the problem of underdetermination of inherent number with plurals, when developing a generative counterpart to (Buckwalter, 2002).

The computational models in (Cavalli-Sforza, Soudi, and Mitamura, 2000) and (Habash, Rambow, and Kiraz, 2005) attempt at the *inferential–realizational* direction. Unfortunately, they implement only sections of the Arabic morphological system. The Arabic resource grammar in the Grammatical Framework (El Dada and Ranta, 2006) is perhaps the most complete inferential–realizational implementation to date. Its style is compatible with the linguistic description in e.g. (Fischer, 2001) or (Badawi, Carter, and Gully, 2004), but the lexicon is now very limited and some other extensions for data-oriented computational applications are still needed.

ElixirFM, the implementation of the system developed in this thesis, is inspired by the methodology in (Forsberg and Ranta, 2004) and by functional programming, just like the Arabic GF is (El Dada and Ranta, 2006). Nonetheless, ElixirFM reuses the Buckwalter lexicon (Buckwalter, 2002) and the annotations in the Prague Arabic Dependency Treebank (Hajič et al., 2004b), and implements a yet more refined linguistic model.

In our view, influenced by the Prague linguistic school and the theory of Functional Generative Description (Sgall, 1967, Sgall, Hajičová, and Panevová, 1986, Panevová, 1980, Hajičová and Sgall, 2003), the task of morphology should be to analyze word forms of a language not only by finding their internal structure, i.e. recognizing morphs, but even by *strictly* discriminating their functions, i.e. providing the true morphemes. Conceived in such a way, it should be *completely* sufficient to generate the word form that represents a lexical unit and features all grammatical categories (and structural components) required by context, purely from the information comprised in the analyses.

It appears from the literature on most other implementations (many summarized in (Al-Sughaiyer and Al-Kharashi, 2004)) that the Arabic computational morphology has understood its role in the sense of operations with morphs rather than morphemes (cf. (El-Sadany and Hashish, 1989)), and has not concerned itself systematically and to the necessary extent with

the role of morphology for syntax. In other terms, the syntax–morphology interface has not been clearly established and respected.

The outline of formal grammar in (Ditters, 2001), for instance, works with grammatical categories like number, gender, humanness, definiteness, but one cannot see which of the existing systems could provide for this information correctly, as they misinterpret some morphs for bearing a category, and underdetermine lexical morphemes in general as to their intrinsic morphological functions. Nowadays, the only exception is the Arabic Grammatical Framework (El Dada and Ranta, 2006, Dada, 2007), which implements its own morphological and syntactic model.

Certain syntactic parsers, like (Othman, Shaalan, and Rafea, 2003), may resort to their own morphological analyzers, but still, they do not get rid of the form of an expression and only incidentally introduce truly functional categories. In syntactic considerations they often call for discriminative extra-linguistic features instead. Commercial systems, e.g. (Chalabi, 2004), do not seem to overcome this interference either.

1.2. Reused Software

The ElixirFM implementation of Functional Arabic Morphology would not have come into existence were it not for many open-source software projects that we could use during our work, or by which we got inspired.

ElixirFM and its lexicons are licensed under GNU GPL and are available on `http://sourceforge.net/projects/elixir-fm/`, along with the other accompanying software (MorphoTrees, Encode Arabic) and the source code of the thesis (ArabTEX extensions, TreeX).

ElixirFM 1.0 is intended for use with the Hugs interactive interpreter of Haskell, available for a number of platforms via `http://haskell.org/hugs/`.

Buckwalter Arabic Morphological Analyzer The bulk of lexical entries in ElixirFM is extracted from the data in the Buckwalter lexicon (Buckwalter, 2002). We devised an algorithm in Perl using the morphophonemic patterns of ElixirFM that finds the roots and templates of the lexical items, as they are available only partially in the original, and produces the ElixirFM lexicon in customizable formats for Haskell and for Perl.

Functional Morphology Library Functional Morphology (Forsberg and Ranta, 2004) is both a methodology for modeling morphology in a paradigmatic manner, and a library of purposely language-independent but customizable modules and functions for Haskell. It partly builds on the Zen computational toolkit for Sanskrit (Huet, 2002). Functional Morphology is also related to the Grammatical Framework, cf. (El Dada and Ranta, 2006) and `http://www.cs.chalmers.se/~markus/FM/`.

TrEd Tree Editor TrEd `http://ufal.mff.cuni.cz/~pajas/tred/` is a general-purpose graphical editor for trees and tree-like graphs written by Petr Pajas. It is implemented in Perl

and is designed to enable powerful customization and macro programming. We have extended TrEd with the annotation mode for MorphoTrees.

1.3. Original Contributions

The following are the original contributions and proposals of the present study:

(i) Recognition of functional versus illusory morphological categories, definition of a minimal but complete system of inflectional parameters in Arabic
(ii) Morphophonemic patterns and their significance for the simplification of the model of morphological alternations
(iii) Inflectional invariant and its consequence for the efficiency of morphological recognition in Arabic
(iv) Intuitive notation for the structural components of words
(v) Conversion of the Buckwalter lexicon into a functional format resembling printed dictionaries
(vi) ElixirFM as a general-purpose model of morphological inflection and derivation in Arabic, implemented with high-level declarative programming
(vii) Abstraction from one particular orthography affecting the clarity of the model and extending its applicability to other written representations of the language
(viii) MorphoTrees as a hierarchization of the process of morphological disambiguation
(ix) Expandable morphological positional tags, restrictions on features, their inheritance
(x) Open-source implementations of ElixirFM, Encode Arabic, MorphoTrees, and extensions for ArabTEX

2. Writing & Reading Arabic

In the context of linguistics, morphology is the study of word forms. In formal language theory, the symbols for representing words are an inseparable part of the definition of the language. In natural languages, the concept is a little different—an utterance can have multiple representations, depending on the means of communication and the conventions for recording it. An abstract computational morphological model should not be limited to texts written in one customary orthography.

This chapter explores the interplay between the genuine writing system and the transcriptions of Arabic. We introduce in detail the ArabTEX notation, a morphophonemic transliteration scheme adopted as the representation of choice for our general-purpose morphological model. We then discuss the problem of recognizing the internal structure of words given the various possible types of their record.

2.1. ArabTEX Notation

The ArabTEX typesetting system (Lagally, 2004) defines its own Arabic script meta-encoding that covers both contemporary and historical orthography. The notation is human-readable

and very natural to write with. Its design is inspired by the standard phonetic transcription of Arabic, which it mimics, yet some distinctions are introduced to make the conversion to the original script or the transcription unambiguous.

Unlike other transliteration concepts based on the strict one-to-one substitution of graphemes, ArabTEX interprets the input characters in context in order to get their proper meaning. Deciding the glyphs of letters (initial, medial, final, isolated) and their ligatures is not the issue of encoding, but of visualizing of the script. Nonetheless, definite article assimilation, inference of *hamza* carriers and silent *ʾalif*s, treatment of auxiliary vowels, optional quoting of diacritics or capitalization, resolution of notational variants, and mode-dependent processing remain the challenges for parsing the notation successfully.

ArabTEX's implementation is documented in (Lagally, 1992), but the parsing algorithm for the notation has not been published except in the form of the source code. The TEX code is organized into deterministic-parsing macros, yet the complexity of the whole system makes consistent modifications or extensions by other users quite difficult.

We describe our own implementations of the interpreter in Chapter 9, where we show how to decode the notation and its proposed extensions. To encode the Arabic script or its phonetic transcription into the ArabTEX notation requires heuristic methods, if we want to achieve linguistically appropriate results.

2.2. Recognition Issues

Arabic is a language of rich morphology, both derivational and inflectional. Due to the fact that the Arabic script usually does not encode short vowels and omits some other important phonological distinctions, the degree of morphological ambiguity is very high.

Besides this complexity, Arabic orthography prescribes to concatenate certain word forms with the preceding or the following ones, possibly changing their spelling and not just leaving out the whitespace in between them. This convention makes the boundaries of lexical or syntactic units, which need to be retrieved as tokens for any deeper linguistic processing, obscure, for they may combine into one compact string of letters and be no more the distinct 'words'.

Thus, the problem of disambiguation of Arabic encompasses not only diacritization (discussed in (Nelken and Shieber, 2005)), but even tokenization, lemmatization, restoration of the structural components of words, and the discovery of their actual morphosyntactic properties, i.e. morphological tagging (cf. (Hajič et al., 2005), plus references therein). These subproblems, of course, can come in many variants, and are partially coupled.

3. Morphological Theory

This chapter defines lexical words as the tokens on which morphological inflection proper will operate. We explore what morphosyntactic properties should be included in the functional model. We discuss the linguistic and computational views on inflectional morphology. Further, we are concerned with Arabic morphology from the structural perspective, designing original morphophonemic patterns and presenting roots as convenient inflectional invariants.

3.1. Functional and Illusory Categories

Functional Arabic Morphology endorses the inferential–realizational principles in the morphological theory (cf. (Stump, 2001)). It re-establishes the system of inflectional and inherent morphosyntactic properties (or grammatical categories or features, in the alternative naming) and discriminates precisely the senses of their use in the grammar. It also deals with syncretism of forms (cf. (Baerman, Brown, and Corbett, 2006)) that seems to prevent the resolution of the underlying categories in some morphological analyzers.

In the thesis, we offer examples of morphological analyses disclosing that grammatical descriptions cannot dispense with a single category for number or for gender, but rather, that we should always specify the sense in which we mean these:

functional category is introduced as the morphosyntactic property that is involved in grammatical considerations; we further divide functional categories into
  logical categories on which agreement with numerals and quantifiers is based
  formal categories controlling other kinds of agreement or pronominal reference
illusory category denotes the value derived merely from the morphs of an expression

Does the classification of the senses of categories actually bring new quality to the linguistic description? We explore in detail the extent of the differences in the values assigned. It may, of course, happen that the values for a given category coincide in all the senses. However, promoting the illusory values to the functional ones is in principle conflicting:

(i) Illusory categories are set only by a presence of some 'characteristic' morph, irrespective of the functional categories of the whole expression.

(ii) If no morph 'characteristic' of a value surrounds the word stem and the stem's morpheme does not have the right information in the lexicon, then the illusory category remains unset. It is the particular issue with the internal/broken plural in Arabic, for which the illusory analyses do not reveal any values of number or gender.

The problem concerns every nominal expression individually and pertains to some verbal forms, too. Functional Arabic Morphology makes the functional gender and number information available thanks to the lexicon that can stipulate some properties as inherent to some lexemes, and thanks to the paradigm-driven generation that associates the inflected forms with the desired functions directly.

3.2. The Pattern Alchemy

In Functional Arabic Morphology, patterns constitute the inventory of phonological melodies of words, regardless of the other functions of the words. Morphophonemic patterns abstract from the consonantal root, which is often recognized or postulated on etymological grounds. Other types of patterns, like the decomposition into separate CV patterns and vocalisms, can be derived from the morphophonemic patterns.

Fischer (2001) uses patterns that abstract away from the root, but can include even inflectional affixes or occasionally restore weak root consonants. For instance, we can find references

to patterns like *ʾafʿala* for *ʾaḥsana* أَحْسَنَ ‘he did right’ or *ʾahdā* أَهْدَى ‘he gave’, but *ʾafʿalu* for *ʾaʿlā* أَعْلَى ‘higher’. In our model, the morphophonemic pattern pertains to the morphological stem and reflects its phonological qualities. Thus, our patterns become `HaFCaL` for *ʾaḥsana* أَحْسَنَ, while `HaFCY` for both *ʾahdā* أَهْدَى and *ʾaʿlā* أَعْلَى.

Beesley (1998) uses the term ‘morphophonemic’ as ‘underlying’, denoting the patterns like `CuCiC` or `staCCaC` or `maCCuuC`. Yet, he also uses the term for anything but the surface form, cf. “an interdigitated but still morphophonemic stem” or “there may be many phonological or orthographical variations between these morphophonemic strings and their ultimate surface pronunciation or spelling” (Beesley, 1998).

Kay (1987) gives an account of finite-state modeling of the nonconcatenative morphological operations. He calls CV patterns ‘prosodic templates’, both terms following (McCarthy, 1981). For further terminological explanations, cf. ((Kiraz, 2001), pages 27–46).

We build on morphophonemic patterns rather than on CV patterns and vocalisms. Words like *istağāb* اِسْتَجَاب ‘to respond’ and *istağwab* اِسْتَجْوَب ‘to interrogate’ have the same underlying VstVCCVC pattern, thus the information on CV patterns alone would not be enough to reconstruct the differences in the surface forms. Morphophonemic patterns, in this case `IstaFAL` and `IstaFCaL`, can easily be mapped to the hypothetical CV patterns and vocalisms, or linked with each other according to their relationship. Morphophonemic patterns deliver more information in a more compact way.

With this approach, we also get a more precise control over the actual word forms—we explicitly confirm that the ‘word’ the pattern should create does undergo the implied transformations. One can therefore speak of ‘weak patterns’ rather than of ‘weak roots’.

The idea of pre-computing the phonological constraints within CV patterns into the ‘morphophonemic’ ones is present in (Yaghi and Yagi, 2004), but is applied to verbs only and is perhaps not understood in the sense of a primary or full-fledged representation ((Yaghi and Yagi, 2004), sec. 5):

> The transformation may be made on the morphological pattern itself, thus producing a sound surface form template. … A coding scheme is adopted that continues to retain letter origins and radical positions in the template so that this will not affect [the author’s model of] affixation. … The surface form template can be rewritten as اِh_F2تّh_M0ˊh_L2ى `AiF2t~aM0aL2Y`. This can be used to form stems such as اِتَّدَى `Ait~adaY` by slotting the root ودي `wdy`.

Yaghi’s templates are not void of root-consonant ‘placeholders’ that actually change under inflection, cf. h_F2 h_L2 indexed by the auxiliary integers to denote their ‘substitutability’. The template, on the other hand, reflects some of the orthographic details and includes Form VIII assimilations that can be abstracted from, cf. esp. the تّ `t~a` group.

With Functional Arabic Morphology, the morphophonemic pattern of *ittadā* اِتَّدَى is simply `IFtaCY`, the root being *wdy* ودي. One of its inflected forms is `IFtaCY |<< "tumA"` *ittadaytumā* اِتَّدَيْتُمَا ‘the two of you accepted compensation’, to follow again the example in (Yaghi and Yagi, 2004). We describe the essence of this notation in Chapter 5.

CV templates are viewed from the perspective of moraic templates in the Prosodic Morphology (McCarthy and Prince, 1990), later discussed by (Kiraz, 2001) within his development of a multitier nonlinear morphological model. Given that we can define a mapping from morphophonemic templates into prosodic or moraic templates, which we easily can, we claim that the prosodic study of the templates is separable from the modeling of morphology.

3.3. The Inflectional Invariant

In our approach, we define roots as sequences of consonants. In most cases, roots are triliteral, such as *k t b* كتب, *q w m* قوم, *d s s* دسس, *r ʾ y* رأي, or quadriliteral, like *d ḥ r ǧ* دحرج, *ṭ m ʾ n* طمأن, *z l z l* زلزل.

Roots in Arabic are, somewhat by definition, inflectional invariants. Unless a root consonant is weak, i.e. one of *y*, *w* or *ʾ*, and unless it assimilates inside a Form VIII pattern, then this consonant will be part of the inflected word form. This becomes apparent when we consider the repertoire and the nature of morphophonemic patterns.

The corollary is that we can effectively exploit the invariant during recognition of word forms. We can check the derivations and inflections of the identified or hypothesized roots only, and need not inflect the whole lexicon before analyzing the given inflected forms in question.

While this seems the obvious way in which learners of Arabic analyze unknown words to look them up in the dictionary, it contrasts strongly with the practice in the design of computational analyzers, where finite-state transducers (Beesley and Karttunen, 2003), or analogously tries (Forsberg and Ranta, 2004, Huet, 2002), are most often used. Of course, other languages than Arabic need not have such convenient invariants.

4. Impressive Haskell

Haskell is a purely functional programming language based on typed λ-calculus, with lazy evaluation of expressions and many impressive higher-order features.

It is beyond the scope of our study to give any general, yet accurate account of the language. We only overview some of its characteristics and point to Haskell's website `http://haskell.org/` for the most appropriate introduction and further references.

In Chapter 5, we exemplify and illustrate the features of Haskell step by step while developing ElixirFM. In Chapter 9, we present the implementation of a grammar of rewrite rules for Encode Arabic.

5. ElixirFM Design

ElixirFM is a high-level implementation of Functional Arabic Morphology. It reuses and extends the Functional Morphology for Haskell (Forsberg and Ranta, 2004), yet the treatment of the language-specific issues constitutes our original contribution.

Inflection and derivation are modeled in terms of paradigms, grammatical categories, lexemes and word classes. The functional and the structural aspects of morphology are clearly

separated. The computation of analysis or generation is conceptually distinguished from the general-purpose linguistic model.

The lexicon of ElixirFM is designed with respect to abstraction, yet is no more complicated than printed dictionaries. It is derived from the open-source Buckwalter lexicon (Buckwalter, 2002) and is enhanced with other unique information.

In Section 5.1, we survey some of the categories of the syntax–morphology interface in Modern Written Arabic, described by Functional Arabic Morphology. In passing, we introduce the basic concepts of programming in Haskell, a modern purely functional language that is an excellent choice for declarative generative modeling of morphologies, as Forsberg and Ranta (2004) have shown.

Section 5.2 is devoted to describing the lexicon of ElixirFM. We develop a so-called domain-specific language embedded in Haskell with which we achieve lexical definitions that are simultaneously a source code that can be checked for consistency, a data structure ready for rather independent processing, and still an easy-to-read-and-edit document resembling the printed dictionaries.

In Section 5.3, we illustrate how rules of inflection and derivation interact with the parameters of the grammar and the lexical information. We claim that the system is reusable in many applications, including computational analysis and generation in various modes, exploring and exporting of the lexicon, printing of the inflectional paradigms, etc.

5.1. Morphosyntactic Categories

Different morphological models categorize individual inflected word forms differently. Some models introduce features and values that are not always complete with respect to the inflectional system, nor mutually orthogonal. We explain what we mean by revisiting the notions of state and definiteness in contemporary written Arabic.

Functional Arabic Morphology refactors the category of state, also denoted as formal definiteness, depending on two parameters. The first controls prefixation of the (virtual) definite article, the other reduces some suffixes if the word is a head of an annexation. In ElixirFM, we define these parameters as type synonyms to standard Haskell types:

```
type Definite = Maybe Bool
type Annexing = Bool
```

The `Definite` values include `Just True` for forms with the definite article, `Just False` for forms in some compounds or after *lā* لَا or *yā* يَا (absolute negatives or vocatives), and `Nothing` for forms that reject the definite article for other reasons. The `State` category is in our view considered as a result of coupling the two independent parameters:

```
type State = Couple Definite Annexing
```

Thus, the indefinite state describes a word void of the definite article(s) and not heading an annexation, i.e. `Nothing :-: False`. Conversely, *ar-rafīʿū* اَلرَّفِيعُو 'the-highs-of' is in the state `Just True :-: True`. The classical construct state is `Nothing :-: True`. The definite state is `Just _ :-: False`, where _ is `True` for (El Dada and Ranta, 2006) and `False` for (Fischer, 2001).

```
|> "k t b" <| [

    FaCaL          `verb`  [ "write", "be destined" ]       `imperf`  FCuL,

    FiCAL          `noun`  [ "book" ]                       `plural`  FuCuL,

    FiCAL |< aT    `noun`  [ "writing" ],

    FiCAL |< aT    `noun`  [ "essay", "piece of writing" ]  `plural`  FiCAL |< At,

    FACiL          `noun`  [ "writer", "author", "clerk" ]  `plural`  FaCaL |< aT
                                                            `plural`  FuCCAL,

    FuCCAL         `noun`  [ "kuttab", "Quran school" ]     `plural`  FaCACIL,

    MaFCaL         `noun`  [ "office", "department" ]       `plural`  MaFACiL,

    MaFCaL |< Iy   `adj`   [ "office" ],

    MaFCaL |< aT   `noun`  [ "library", "bookstore" ]       `plural`  MaFACiL    ]
```

Figure 1. Entries of the ElixirFM lexicon nested under the root k t b كتب using morphophonemic templates.

5.2. ElixirFM Lexicon

Unstructured text is just a list of characters, i.e. a string. Yet words do have structure, particularly in Arabic. We work with strings as the superficial word forms, but the internal representations are more abstract (and computationally more efficient, too).

The definition of *lexemes* can include the derivational *root and pattern* information if appropriate, cf. (Habash, Rambow, and Kiraz, 2005), and our model does encourage this. The surface word *kitāb* كِتَاب 'book' can decompose to the triconsonantal root *k t b* كتب and the morphophonemic pattern `FiCAL`:

```
data PatternT = FaCaL             | FAL      | FaCY     | FaCL
              | HaFCAL    | HACAL            | HaFCA'               | HACA'
              | FiCAL                         | FiCA'
              | FuCCAL    | FUCAL
              | TaFACuL                       | TaFACI
              | IFtiCAL           | IFtiyAL   | IFtiCA'
              | MustaFCaL         | MustaFAL  | MustaFCY | MustaFaCL

              | {- ... -}                          deriving (Eq, Enum, Show)
```

The `deriving` clause associates the type `PatternT` with methods for testing equality, enumerating all the values, and turning the names of the values into strings. Of course, ElixirFM provides functions for properly interlocking the patterns with the roots:

```
merge "k t b"  FiCAL      ⟶   "kitAb"
merge "^g w b" IstaFAL    ⟶   "ista^gAb"
merge "^g w b" IstaFCaL   ⟶   "ista^gwab"
merge "s ' l"  MaFCUL     ⟶   "mas'Ul"
merge "r ' y"  HAFA'      ⟶   "'ArA'"
merge "z h r"  IFtaCaL    ⟶   "izdahar"
```

The *izdahar* اِزدَهَر ‘to flourish’ case exemplifies that exceptionless assimilations need not be encoded in the patterns, but can instead be hidden in rules.

The whole generative model adopts the notation of ArabTEX (Lagally, 2004) as a meta-encoding of both the orthography and phonology. Therefore, instantiation of the `"'"` *hamza* carriers or other merely orthographic conventions do not obscure the morphological model. With Encode Arabic interpreting the notation, ElixirFM can at the surface level process the original Arabic script (non-)vocalized to any degree or work with some kind of transliteration or even transcription thereof.

Morphophonemic patterns represent the stems of words. The various kinds of abstract prefixes and suffixes can be expressed either as atomic values, or as literal strings wrapped into extra constructors:

```
data Prefix = Al | LA | Prefix String

data Suffix = Iy | AT | At | An | Ayn | Un | In | Suffix String

al = Al; lA = LA                        -- function synonyms

aT = AT; ayn = Ayn; aN = Suffix "aN"
```

Affixes and patterns are put together via the `Morphs a` data type, where `a` is a triliteral pattern `PatternT` or a quadriliteral `PatternQ` or a non-templatic word stem `Identity` of type `PatternL`:

```
data PatternL = Identity
data PatternQ = KaRDaS | KaRADiS {- ... -}

data Morphs a = Morphs a [Prefix] [Suffix]
```

The word *lā-silkīy* لَاسِلكِيّ ‘wireless’ can thus be decomposed as the root *s l k* سلك and the value `Morphs FiCL [LA] [Iy]`. Shunning such concrete representations, we define new operators `>|` and `|<` that denote prefixes, resp. suffixes, inside `Morphs a`. If it is strings that need to be prefixed or suffixed, the shorthand `>>|` and `|<<` can also be used:

```
lA >| FiCL |< Iy    ⟶    Morphs FiCL [LA] [Iy]

al >| lA >| FiCL |< Iy |<< "u"   ⟶   Morphs FiCL [Al, LA] [Suffix "u", Iy]
```

With the introduction of patterns and templates, some synonymous functions and the intuitive operators, we start developing what can be viewed as a domain-specific language embedded in the general-purpose programming language. Encouraged by the flexibility of many other domain-specific languages in Haskell, we design the lexicon to look as shown in Figure 1, yet be a verifiable source code defining a directly interpretable data structure.

```
data Mood = Indicative | Subjunctive | Jussive | Energetic  deriving (Eq, Enum)
data Gender = Masculine | Feminine                          deriving (Eq, Enum)

data ParaVerb = VerbP      Voice Person Gender Number
              | VerbI Mood Voice Person Gender Number
              | VerbC                   Gender Number       deriving Eq


paraVerbC :: Morphing a b => Gender -> Number -> [Char] -> a -> Morphs b
paraVerbC g n i = case n of

            Singular    -> case g of    Masculine ->  prefix i . suffix ""
                                        Feminine  ->  prefix i . suffix "I"

            Plural      -> case g of    Masculine ->  prefix i . suffix "UW"
                                        Feminine  ->  prefix i . suffix "na"

            _           ->                            prefix i . suffix "A"
```

Figure 2. Excerpt from the implementation of verbal inflectional features and paradigms in ElixirFM.

The lexicon of ElixirFM is derived from the open-source Buckwalter lexicon (Buckwalter, 2002). We devised an algorithm in Perl using the morphophonemic patterns of ElixirFM that finds the roots and templates of the lexical items, as they are available only partially in the original, and produces the lexicon in formats for Perl and for Haskell.

5.3. Morphological Rules

Inferential–realizational morphology is modeled in terms of paradigms, grammatical categories, lexemes and word classes. ElixirFM implements the comprehensive rules that draw the information from the lexicon and generate the word forms given the appropriate morphosyntactic parameters. The whole is invoked through a convenient `inflect` method.

The lexicon and the parameters determine the choice of paradigms. The template selection mechanism differs for nominals (providing plurals) and for verbs (providing all needed stem alternations in the extent of the entry specifications of e.g. Hans Wehr's dictionary).

In Figure 2, the algebraic data type `ParaVerb` restricts the space in which verbs are inflected by defining three Cartesian products of the elementary categories: a verb can have `VerbP` perfective forms inflected in voice, person, gender, number, `VerbI` imperfective forms inflected also in mood, and `VerbC` imperatives inflected in gender and number only.

The paradigm for inflecting imperatives, the only such paradigm in ElixirFM, is implemented as `paraVerbC`. It is a function parametrized by some particular value of gender `g` and number `n`, as well as the initial imperative prefix `i` and the verbal stem (both inferred from the morphophonemic patterns in the lexical entry) and yielding the inflected form.

The definition of `paraVerbC` is simple and concise due to the chance to compose with . the partially applied `prefix` and `suffix` functions and to virtually omit the next argument. This advanced formulation perhaps does not seem as minimal as the mere specification of the literal endings or prefixes, but we present it here to illustrate the options that there are. An abstract paradigm can be used on more abstract types than just strings. Inflected forms need not be merged with roots yet, and can retain the internal structure:

```
paraVerbC Feminine Plural "u" FCuL    ⟶    "u" >>| FCuL |<< "na"

merge "k t b" (Prefix "u" >| FCuL |< Suffix "na")  ⟶

                                        "uktubna" uktubna اُكْتُبْنَ fem.pl. ‘write!’
```

The highlight of the Arabic morphology is that the ‘irregular’ inflection actually rests in strictly observing some additional rules, the nature of which is phonological. Therefore, surprisingly, ElixirFM does not even distinguish between verbal and nominal word formation when enforcing these rules. This reduces the number of paradigms to the prototypical 3 verbal and 5 nominal. Yet, the model is efficient.

Nominal inflection is also driven by the information from the lexicon and by phonology. Note that the morphophonemic patterns and the `Morphs a` templates are actually extremely informative. We can use them as determining the inflectional class and the paradigm function, and thus we can almost avoid other unintuitive or excessive indicators of the kind of weak morphology, diptotic inflection, and the like.

5.4. Applications

The ElixirFM linguistic model and the data of the lexicon can be integrated into larger applications or used as standalone libraries and resources.

The language-independent part of the system could rest in the Functional Morphology library (Forsberg and Ranta, 2004). Among other useful things, it implements the compilation of the inflected word forms and their associated morphosyntactic categories into morphological analyzers and generators. The method used for analysis is deterministic parsing with tries, cf. also (Huet, 2002, Ljunglöf, 2002).

Nonetheless, ElixirFM provides an original analysis method exploiting the inflectional invariant defined in Chapter 3. We can, at least in the present version of the implementation, dispense with the compilation into tries, and we use rather minimal computational resources.

We define a class of types that can be `Resolved`, which introduces one rather general method `resolveBy` and one more specific method `resolve` saying that the form in question should be resolved by equality with the inflected forms in the model. The generic `resolveBy` method can be used esp. for recognition of partially vocalized or completely non-vocalized representations of Arabic, or allow in fact arbitrary kinds of omissions, cf. Chapter 6.

Reusing and extending the original Functional Morphology library, ElixirFM also provides functions for exporting and pretty-printing the linguistic model into XML, LaTeX, Perl, SQL, and other custom formats.

6. Other Listings

This chapter is a non-systematic overview of the features of ElixirFM. It can serve as a tutorial for the first sessions with ElixirFM in the environment of the Hugs interpreter. Here, we present just a couple of examples.

```
ElixirFM> inflect (FiCAL `noun` []) "--------2-"

[("N------S2I",[("f ` l",FiCAL |<< "iN")]),("N------S2R",[(......
,("N------D2L",[("f ` l",FiCAL |<< "ay")]),...,("N------P2L",[])]

ElixirFM> pretty $ inflect (RE "k t b" $ FiCAL `noun` []) "-------S2[IDR]"

("N------S2I",[("k t b",FiCAL |<< "iN")])
("N------S2R",[("k t b",FiCAL |<< "i")])
("N------S2D",[("k t b",al >| FiCAL |<< "i")])

ElixirFM> uncurry merge ("k t b", FiCAL |<< "iN")

"kitAbiN"

ElixirFM> pretty $ inflect (RE "k t b" $ FiCAL `noun` [] `plural` FuCuL)
                           "-------P2[IDR]"

("N------P2I",[("k t b",FuCuL |<< "iN")])
("N------P2R",[("k t b",FuCuL |<< "i")])
("N------P2D",[("k t b",al >| FuCuL |<< "i")])

ElixirFM> pretty $ resolveBy (omitting "aiuAUI") "ktbuN"

N------S1I kitAbuN "k t b" FiCAL ["book"]
N------P1I kutubuN "k t b" FiCAL ["book"]
N------S1I kAtibuN "k t b" FACiL ["writer","author","clerk"]
A-----MS1I kAtibuN "k t b" FACiL ["writing"]

ElixirFM> pretty $ resolveBy (omitting $ (encode UCS . decode Tim) "~aiuKNF")
                             (decode Tim "ktAb")

N------S1I kitAbuN "k t b" FiCAL ["book"]
N------S1R kitAbu "k t b" FiCAL ["book"]
N------S1A kitAbu "k t b" FiCAL ["book"]
N------S1L kitAbu "k t b" FiCAL ["book"]
N------S2I kitAbiN "k t b" FiCAL ["book"]
N------S2R kitAbi "k t b" FiCAL ["book"]
N------S2A kitAbi "k t b" FiCAL ["book"]
N------S2L kitAbi "k t b" FiCAL ["book"]
N------S4R kitAba "k t b" FiCAL ["book"]
N------S4A kitAba "k t b" FiCAL ["book"]
N------S4L kitAba "k t b" FiCAL ["book"]
N------P1I kuttAbuN "k t b" FACiL ["writer","author","clerk"]
N------P1R kuttAbu "k t b" FACiL ["writer","author","clerk"]
N------P1A kuttAbu "k t b" FACiL ["writer","author","clerk"]
```

```
N------P1L kuttAbu "k t b" FACiL ["writer","author","clerk"]
N------P2I kuttAbiN "k t b" FACiL ["writer","author","clerk"]
N------P2R kuttAbi "k t b" FACiL ["writer","author","clerk"]
N------P2A kuttAbi "k t b" FACiL ["writer","author","clerk"]
N------P2L kuttAbi "k t b" FACiL ["writer","author","clerk"]
N------P4R kuttAba "k t b" FACiL ["writer","author","clerk"]
N------P4A kuttAba "k t b" FACiL ["writer","author","clerk"]
N------P4L kuttAba "k t b" FACiL ["writer","author","clerk"]
```

7. MorphoTrees

MorphoTrees (Smrž and Pajas, 2004) evolved as an idea of building effective and intuitive hierarchies over the information presented by morphological systems. Such a concept is especially interesting for Arabic and the Functional Arabic Morphology, yet, it is not limited to the language, nor to the formalism, and various extensions are imaginable.

7.1. The MorphoTrees Hierarchy

As an inspiration for the design of the hierarchies, consider the following analyses of the string `fhm` فهم. Some readings will interpret it as just one token related to the notion of 'understanding', but homonymous for several lexical units, each giving many inflected forms, distinct phonologically despite their identical spelling in the ordinary non-vocalized text. Other readings will decompose the string into two co-occurring tokens, the first one, in its non-vocalized form `f` ف, standing for an unambiguous conjunction, and the other one, `hm` هم, analyzed as a verb, noun, or pronoun, each again ambiguous in its functions.

Clearly, this type of concise and 'structured' description does not come ready-made—we have to construct it on top of the overall morphological knowledge. We can take the output solutions of morphological analyzers and process them according to our requirements on tokenization and 'functionality' stated above. Then, we can merge the analyses and their elements into a five-level hierarchy similar to that of Figure 3. The leaves of it are the full forms of the tokens plus their tags as the atomic units. The root of the hierarchy represents the input string, or generally the input entity (some linear or structured subpart of the text). Rising from the leaves up to the root, there is the level of lemmas of the lexical units, the level of non-vocalized canonical forms of the tokens, and the level of decomposition of the entity into a sequence of such forms, which implies the number of tokens and their spelling.

Note that the MorphoTrees hierarchy itself might serve as a framework for evaluating morphological taggers, lemmatizers and stemmers of Arabic, since it allows for resolution of their performance on the different levels, which does matter with respect to the variety of applications.

7.2. MorphoTrees Disambiguation

The annotation of MorphoTrees rests in selecting the applicable sequence of tokens that analyze the entity in the context of the discourse. In a naive setting, an annotator would be left

to search the trees by sight, decoding the information for every possible analysis before coming across the right one. If not understood properly, the supplementary levels of the hierarchy would rather tend to be a nuisance …

Instead, MorphoTrees in TrEd take great advantage of the hierarchy and offer the option to restrict one's choice to subtrees and hide those leaves or branches that do not conform to the criteria of the annotation. Moreover, many restrictions are applied automatically, and the decisions about the tree can be controlled in a very rapid and elegant way.

The MorphoTrees of the entity `fhm` فهم in Figure 3 are in fact already annotated. The annotator was expecting, from the context, the reading involving a conjunction. By pressing the shortcut `c` at the root node, he restricted the tree accordingly, and the only one eligible leaf satisfying the `C---------` tag restriction was selected at that moment. Nonetheless, the *fa-* فَ 'so' conjunction is part of a two-token entity, and some annotation of the second token must also be performed. Automatically, all inherited restrictions were removed from the `hm` هم subtree (notice the empty tag in the flag over it), and the subtree unfolded again. The annotator moved the node cursor to the lemma for the pronoun, and restricted its readings to the nominative `--------1-` by pressing another mnemonic shortcut `1`, upon which the single conforming leaf *hum* هُم 'they' was selected automatically. There were no more decisions to make and the annotation proceeded to the next entity of the discourse.

Alternatively, the annotation could be achieved merely by typing `s1`. The restrictions would unambiguously lead to the nominative pronoun, and then, without human intervention, to the other token, the unambiguous conjunction. These automatic decisions need no linguistic model, yet are very effective.

7.3. Further Discussion

Hierarchization of the selection task seems to be the most important contribution of the idea. The suggested meaning of the levels of the hierarchy mirrors the linguistic theory and also one particular strategy for decision-making, neither of which are universal. If we adapt MorphoTrees to other languages or hierarchies, the power of trees remains, though—efficient top-down search or bottom-up restrictions, gradual focusing on the solution, refinement, inheritance and sharing of information, etc.

The levels of MorphoTrees are extensible internally as well as externally in both directions, and the concept incites new views on some issues encompassed by morphological analysis and disambiguation.

In PADT, whose MorphoTrees average roughly 8–10 leaves per entity depending on the data set while the result of annotation is 1.16–1.18 tokens per entity, restrictions as a means of direct access to the solutions improve the speed of annotation significantly.

We propose to evaluate tokenizations in terms of the Longest Common Subsequence problem (cf. (Crochemore et al., 2000)). The tokens that are the members of the LCS with some referential tokenization, are considered correctly recognized. Dividing the length of the LCS by the length of one of the sequences, we get recall, doing it for the other of the sequences, we get precision. The harmonic mean of both is $F_{\beta=1}$-measure (cf. (Manning and Schütze, 1999)).

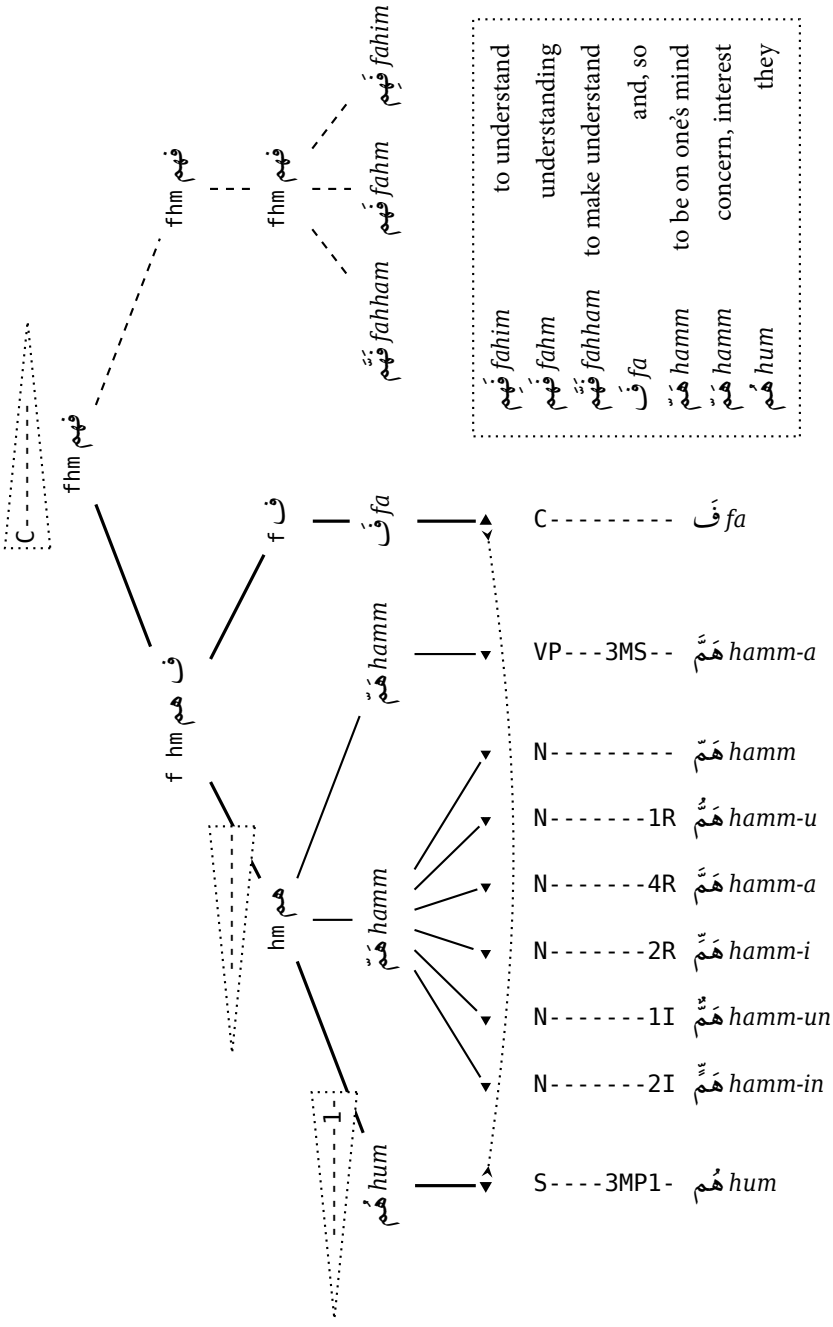


Figure 3. MorphoTrees of the orthographic string `fhm` فهم *including annotation with restrictions. The dashed lines indicate there is no solution suiting the inherited restrictions in the given subtree. The dotted line symbolizes there might be implicit morphosyntactic constraints between the adjacent tokens in the analyses.*

8. Lexicon versus Treebank

This chapter outlines the structure of linguistic description in the framework of Functional Generative Description and motivates our specific concerns about Arabic within the Prague Arabic Dependency Treebank.

8.1. Functional Description of Language

Prague Arabic Dependency Treebank (Hajič et al., 2004a, Hajič et al., 2004b) is a project of analyzing large amounts of linguistic data in Modern Written Arabic in terms of the formal representation of language that originates in the Functional Generative Description (Sgall, 1967, Sgall, Hajičová, and Panevová, 1986, Panevová, 1980, Hajičová and Sgall, 2003).

In this theory, the formal representation delivers the linguistic meaning of what is expressed by the surface realization, i.e. the natural language. The description is designed to enable generating the natural language out of the formal representations. By constructing the treebank, we provide a resource for computational learning of the correspondences between both languages, the natural and the formal.

Morphological annotations identify the textual forms of a discourse lexically and recognize the morphosyntactic categories that the forms assume. Processing on the analytical level describes the superficial syntactic relations present in the discourse, whereas the tectogrammatical level reveals the underlying structures and restores the linguistic meaning (cf. (Sgall, Panevová, and Hajičová, 2004)).

Linguistic data, i.e. mostly newswire texts in their written form, are gradually analyzed in this system of levels, and their linguistic meaning is thus reconstructed and made explicit.

8.2. Analytical Syntax

The tokens with their disambiguated grammatical information enter the annotation of analytical syntax (Žabokrtský and Smrž, 2003, Hajič et al., 2004b). This level is formalized into dependency trees the nodes of which are the tokens. Relations between nodes are classified with analytical syntactic functions. More precisely, it is the whole subtree of a dependent node that fulfills the particular syntactic function with respect to the governing node.

In Figure 4, we analyze a verbal sentence with coordination and a subordinate relative clause. Coordination is depicted with a diamond node and dashed 'dependency' edges between the coordination node and its member coordinants.

Both clauses and nominal expressions can assume the same analytical functions—the attributive clause in our example is Atr, just like in the case of nominal attributes. Pred denotes the main predicate, Sb is subject, Obj is object, Adv stands for adverbial. AuxP, AuxY and AuxK are auxiliary functions of specific kinds.

The coordination relation is different from the dependency relation, but we can depict it in the tree-like manner, too. The coordinative node becomes Coord, and the subtrees that are the members of the coordination are marked as such (cf. dashed edges). Dependents modifying

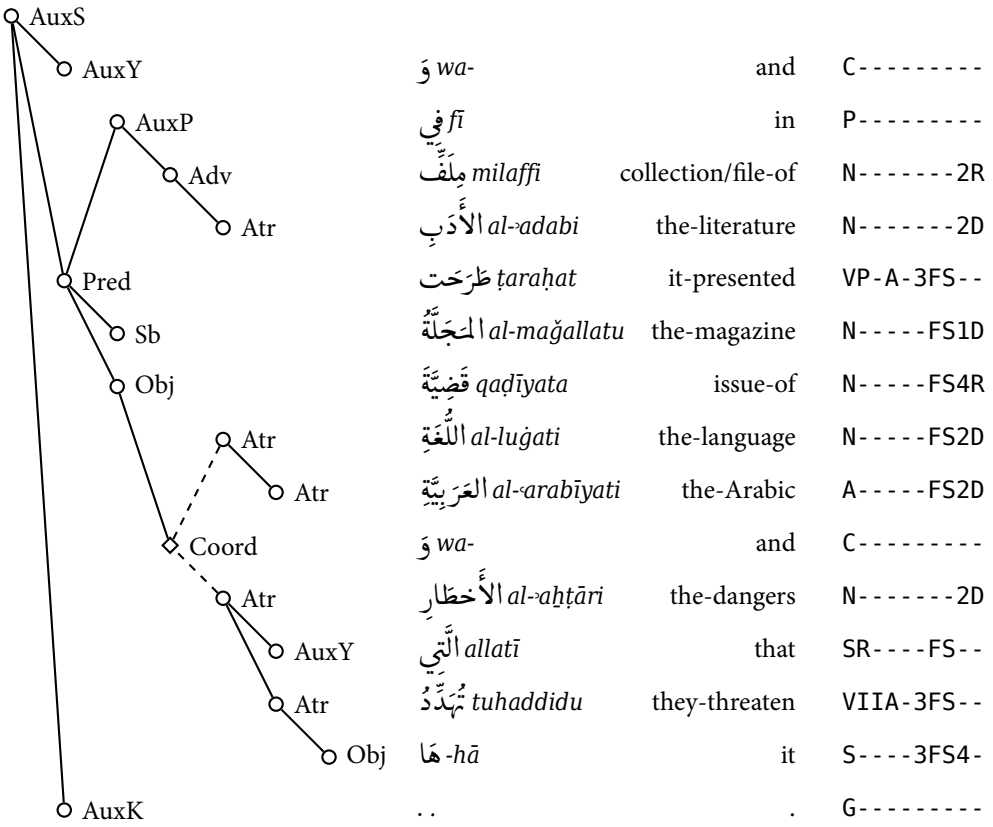


Figure 4. An analytical representation of the sentence In the section on literature, the magazine presented the issue of the Arabic language and the dangers that threaten it.

the coordination as a whole would attach directly to the Coord node, yet would not be marked as coordinants—therefrom, the need for distinguishing coordination and pure dependency in the trees.

The immediate-dominance relation that we capture in the annotation is independent of the linear ordering of words in an utterance, i.e. the linear-precedence relation (Debusmann, 2006). Thus, the expressiveness of the dependency grammar is stronger than that of phrase-structure context-free grammar. The dependency trees can become non-projective by featuring crossing dependencies, which reflects the possibility of relaxing word order while preserving the links of grammatical government.

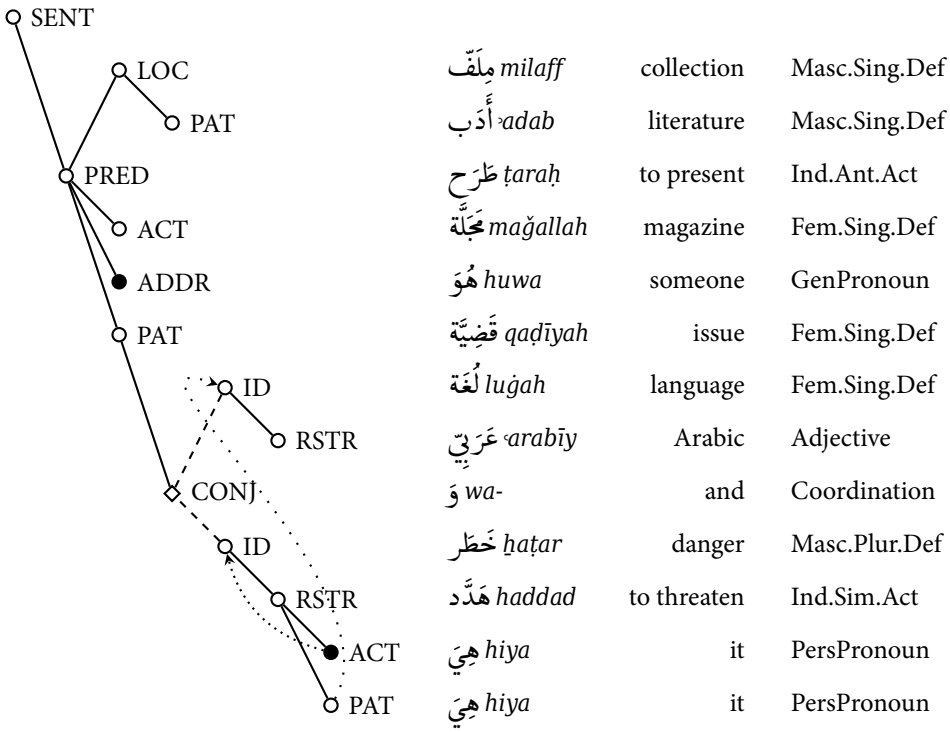


Figure 5. A tectogrammatical representation of the sentence In the section on literature, the magazine presented the issue of the Arabic language and the dangers that threaten it.

8.3. Tectogrammatics

The analytical syntax is yet a precursor to the deep syntactic annotation (Sgall, Panevová, and Hajičová, 2004, Mikulová and others, 2006) with the following characteristics:

deleted nodes only autosemantic lexemes and coordinative nodes are involved in tectogrammatics; synsemantic lexemes, such as prepositions or particles, are deleted from the trees and may be instead reflected in the values of deep grammatical categories, called *grammatemes*, that are associated with the relevant autosemantic nodes

inserted nodes autosemantic lexemes that do not appear explicitly in the surface syntax, yet that are required as obligatory by valency frames or by other criteria of tectogrammatical well-formedness, are inserted into the deep syntactic structures; the elided lexemes may be copies of other explicit nodes, or may be restored even as generic or unspecified

functors are the tectogrammatical functions describing deep dependency relations; the underlying theory distinguishes *arguments* (inner participants: ACTor, PATient, ADDRessee, ORIGin, EFFect) and *adjuncts* (free modifications, e.g.: LOCation, CAUSe, MANNer, TimeWHEN, ReSTRictive, APPurtenance) and specifies the type of coordination (e.g. CONJunctive, DISJunctive, ADVerSative, ConSeQuential)

grammatemes are the deep grammatical features that are necessary for proper generation of the surface form of an utterance, given the tectogrammatical tree as well, cf. (Hajič et al., 2004b)

coreference pronouns are matched with the lexical mentions they refer to; we distinguish *grammatical* coreference (the coreferent is determined by grammar) and *textual* coreference (otherwise); in Figure 5, the densely dotted arc indicates grammatical coreference, the loosely dotted curve denotes textual coreference

Compare the representations in Figures 4 and 5. Note the differences in the set of nodes actually represented, esp. the restored ADDRessee which is omitted in the surface form of the sentence, but is obligatory in the valency frame of the semantics of the PREDicate.

8.4. Dependency and Inherent vs. Inflectional Properties

Analytical syntax makes the agreement relations more obvious. We can often use those relations to infer information on inherent lexical properties as to gender, number, and humanness, as other words in the relation can, with their inflectional or inferred inherent properties, provide enough constraints.

So this problem is a nice example for constraint programming. Our experiments with the treebank so far have been implemented in Perl, and the inference algorithm was not optimal. Neither was the handling of constraints that (perhaps by an error in the annotation) contradict the other ones. Anyway, we did arrive at promising preliminary results.

These experiments have not been fully completed, though, and their revision is needed. In view of that, we consider formulating the problem in the Mozart/Oz constraint-based programming environment ((Van Roy and Haridi, 2004), chapters 9 and 12).

8.5. Tectogrammatics and Derivational Morphology

We can define default derivations of participles and deverbal nouns, the *maṣdar*s, or consider transformations of patterns between different derivational forms, like in the case of Czech where lexical-semantic shifts are also enforced in the valency theory (cf. (Žabokrtský, 2005)). If the default happens to be inappropriate, then a lexical entry can be extended to optionally include the lexicalized definition of the information that we might require.

The concrete transformations that should apply on the tectogrammatical level are a research in progress, performed by the whole PADT team.

The ability to do the transformations, however, is expected in near future as a direct extension of the ElixirFM system.

9. Encode Arabic

This chapter contains details about the implementations related to processing the ArabTEX notation and its extensions. The mentioned software is open-source and is available via `http://sourceforge.net/projects/encode-arabic/`.

Extending ArabTEX The `alocal` package implements some of the notational extensions of Encode Arabic to work in ArabTEX.

The `acolor` package adds colorful typesetting to ArabTEX. Thanks are due to Karel Mokrý who implemented the core of this functionality originally.

Independent Libraries The Perl implementation of Encode Arabic is documented at `http://search.cpan.org/dist/Encode-Arabic/`.

In the thesis, we present parts of the implementation of our Haskell library for processing the Arabic language in the ArabTEX transliteration (Lagally, 2004), a non-trivial and multi-purpose notation for encoding Arabic orthographies and phonetic transcriptions in parallel. Our approach relies on the Pure Functional Parsing library developed in (Ljunglöf, 2002), which we accommodate to our problem and partly extend. We promote modular design in systems for modeling or processing natural languages.

Conclusion

In the thesis, we developed the theory of Functional Arabic Morphology and designed ElixirFM as its high-level functional and interactive implementation written in Haskell.

Next to numerous theoretical points on the character of Arabic morphology and its relation to syntax, we proposed a model that represents the linguistic data in an abstract and extensible notation that encodes both *orthography* and *phonology*, and whose interpretation is customizable. We developed a domain-specific language in which the lexicon is stored and which allows easy manual editing as well as automatic verification of consistency. We believe that the modeling of both the *written* language and the *spoken* dialects can share the presented methodology.

ElixirFM and its lexicons are licensed under GNU GPL and are available on `http://sourceforge.net/projects/elixir-fm/`. The implementations of Encode Arabic, MorphoTrees, and the ArabTEX extensions are published likewise.

We intend to improve our work further and integrate ElixirFM closely with MorphoTrees as well as with both levels of syntactic representation in the Prague Arabic Dependency Treebank.

Acknowledgement The research for the thesis was supported by the Ministry of Education of the Czech Republic (MSM 0021620838), by the Grant Agency of Charles University in Prague (UK 373/2005), and by the Grant Agency of the Czech Academy of Sciences (1ET101120413).

Bibliography

Al-Sughaiyer, Imad A. and Ibrahim A. Al-Kharashi. 2004. Arabic Morphological Analysis Techniques: A Comprehensive Survey. *Journal of the American Society for Information Science and Technology*, 55(3):189–213.

Badawi, Elsaid, Mike G. Carter, and Adrian Gully. 2004. *Modern Written Arabic: A Comprehensive Grammar*. Routledge.

Baerman, Matthew, Dunstan Brown, and Greville G. Corbett. 2006. *The Syntax-Morphology Interface. A Study of Syncretism*. Cambridge Studies in Linguistics. Cambridge University Press.

Beesley, Kenneth R. 1998. Arabic Morphology Using Only Finite-State Operations. In *COLING-ACL'98 Proceedings of the Workshop on Computational Approaches to Semitic languages*, pages 50–57.

Beesley, Kenneth R. 2001. Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001. In *EACL 2001 Workshop Proceedings on Arabic Language Processing: Status and Prospects*, pages 1–8, Toulouse, France.

Beesley, Kenneth R. and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Studies in Computational Linguistics. CSLI Publications, Stanford, California.

Buckwalter, Tim. 2002. Buckwalter Arabic Morphological Analyzer Version 1.0. LDC catalog number LDC2002L49, ISBN 1-58563-257-0.

Cavalli-Sforza, Violetta, Abdelhadi Soudi, and Teruko Mitamura. 2000. Arabic Morphology Generation Using a Concatenative Strategy. In *Proceedings of NAACL 2000*, pages 86–93, Seattle.

Chalabi, Achraf. 2004. Sakhr Arabic Lexicon. In *NEMLAR International Conference on Arabic Language Resources and Tools*, pages 21–24. ELDA.

Crochemore, Maxime, Costas S. Iliopoulos, Yoan J. Pinzon, and James F. Reid. 2000. A Fast and Practical Bit-Vector Algorithm for the Longest Common Subsequence Problem. In *Proceedings of the 11th Australasian Workshop On Combinatorial Algorithms*, Hunter Valley, Australia.

Dada, Ali. 2007. Implementation of the Arabic Numerals and their Syntax in GF. In *ACL 2007 Proceedings of the Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, pages 9–16, Prague, Czech Republic, June. Association for Computational Linguistics.

Debusmann, Ralph. 2006. *Extensible Dependency Grammar: A Modular Grammar Formalism Based On Multigraph Description*. Ph.D. thesis, Saarland University.

Ditters, Everhard. 2001. A Formal Grammar for the Description of Sentence Structure in Modern Standard Arabic. In *EACL 2001 Workshop Proceedings on Arabic Language Processing: Status and Prospects*, pages 31–37, Toulouse, France.

El Dada, Ali and Aarne Ranta. 2006. Open Source Arabic Grammars in Grammatical Framework. In *Proceedings of the Arabic Language Processing Conference (JETALA)*, Rabat, Morocco, June 2006. IERA.

El-Sadany, Tarek A. and Mohamed A. Hashish. 1989. An Arabic morphological system. *IBM Systems Journal*, 28(4):600–612.

Fischer, Wolfdietrich. 2001. *A Grammar of Classical Arabic*. Yale Language Series. Yale University Press, third revised edition. Translated by Jonathan Rodgers.

Forsberg, Markus and Aarne Ranta. 2004. Functional Morphology. In *Proceedings of the Ninth ACM SIGPLAN International Conference on Functional Programming, ICFP 2004*, pages 213–223. ACM Press.

Habash, Nizar. 2004. Large Scale Lexeme Based Arabic Morphological Generation. In *JEP-TALN 2004, Session Traitement Automatique de l'Arabe*, Fes, Morocco, April 2004.

Habash, Nizar, Owen Rambow, and George Kiraz. 2005. Morphological Analysis and Generation for Arabic Dialects. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 17–24, Ann Arbor, Michigan. Association for Computational Linguistics.

Hajič, Jan, Otakar Smrž, Tim Buckwalter, and Hubert Jin. 2005. Feature-Based Tagger of Approximations of Functional Arabic Morphology. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*, pages 53–64, Barcelona, Spain.

Hajič, Jan, Otakar Smrž, Petr Zemánek, Petr Pajas, Jan Šnaidauf, Emanuel Beška, Jakub Kráčmar, and Kamila Hassanová. 2004a. Prague Arabic Dependency Treebank 1.0. LDC catalog number LDC2004T23, ISBN 1-58563-319-4.

Hajič, Jan, Otakar Smrž, Petr Zemánek, Jan Šnaidauf, and Emanuel Beška. 2004b. Prague Arabic Dependency Treebank: Development in Data and Tools. In *NEMLAR International Conference on Arabic Language Resources and Tools*, pages 110–117. ELDA.

Hajičová, Eva and Petr Sgall. 2003. Dependency Syntax in Functional Generative Description. In *Dependenz und Valenz – Dependency and Valency*, volume I. Walter de Gruyter, pages 570–592.

Holes, Clive. 2004. *Modern Arabic: Structures, Functions, and Varieties*. Georgetown Classics in Arabic Language and Linguistics. Georgetown University Press.

Huet, Gérard. 2002. The Zen Computational Linguistics Toolkit. ESSLLI Course Notes, FoLLI, the Association of Logic, Language and Information.

Kay, Martin. 1987. Nonconcatenative Finite-State Morphology. In *Proceedings of the Third Conference of the European Chapter of the ACL (EACL-87)*, pages 2–10, Copenhagen, Denmark. ACL.

Kiraz, George Anton. 2001. *Computational Nonlinear Morphology with Emphasis on Semitic Languages*. Studies in Natural Language Processing. Cambridge University Press.

Lagally, Klaus. 1992. ArabTEX: Typesetting Arabic with Vowels and Ligatures. In *EuroTEX 92*, page 20, Prague, Czechoslovakia, September 14–18.

Lagally, Klaus. 2004. ArabTEX: Typesetting Arabic and Hebrew, User Manual Version 4.00. Technical Report 2004/03, Fakultät Informatik, Universität Stuttgart, March 11.

Ljunglöf, Peter. 2002. *Pure Functional Parsing. An Advanced Tutorial*. Licenciate thesis, Göteborg University & Chalmers University of Technology.

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge.

McCarthy, John and Alan Prince. 1990. Foot and Word in Prosodic Morphology: The Arabic Broken Plural. *Natural Language and Linguistic Theory*, 8:209–283.

McCarthy, John J. 1981. A Prosodic Theory of Nonconcatenative Morphology. *Linguistic Inquiry*, 12:373–418.

Mikulová, Marie et al. 2006. A Manual for Tectogrammatical Layer Annotation of the Prague Dependency Treebank. Technical report, Charles University in Prague.

Nelken, Rani and Stuart M. Shieber. 2005. Arabic Diacritization Using Finite-State Transducers. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 79–86, Ann Arbor.

Othman, Eman, Khaled Shaalan, and Ahmed Rafea. 2003. A Chart Parser for Analyzing Modern Standard Arabic Sentence. In *Proceedings of the MT Summit IX Workshop on Machine Translation for Semitic Languages: Issues and Approaches*, pages 37–44.

Panevová, Jarmila. 1980. *Formy a funkce ve stavbě české věty [Forms and Functions in the Structure of the Czech Sentence]*. Academia.

Ramsay, Allan and Hanady Mansur. 2001. Arabic morphology: a categorial approach. In *EACL 2001 Workshop Proceedings on Arabic Language Processing: Status and Prospects*, pages 17–22, Toulouse, France.

Sgall, Petr. 1967. *Generativní popis jazyka a česká deklinace [Generative Description of Language and Czech Declension]*. Academia.

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

Sgall, Petr, Jarmila Panevová, and Eva Hajičová. 2004. Deep Syntactic Annotation: Tectogrammatical Representation and Beyond. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 32–38. Association for Computational Linguistics.

Smrž, Otakar. 2007. *Functional Arabic Morphology. Formal System and Implementation*. Ph.D. thesis, Charles University in Prague.

Smrž, Otakar and Petr Pajas. 2004. MorphoTrees of Arabic and Their Annotation in the TrEd Environment. In *NEMLAR International Conference on Arabic Language Resources and Tools*, pages 38–41. ELDA.

Stump, Gregory T. 2001. *Inflectional Morphology. A Theory of Paradigm Structure*. Cambridge Studies in Linguistics. Cambridge University Press.

Van Roy, Peter and Seif Haridi. 2004. *Concepts, Techniques, and Models of Computer Programming*. MIT Press, Cambridge, March.

Yaghi, Jim and Sane Yagi. 2004. Systematic Verb Stem Generation for Arabic. In *COLING 2004 Computational Approaches to Arabic Script-based Languages*, pages 23–30, Geneva, Switzerland.

Žabokrtský, Zdeněk. 2005. *Valency Lexicon of Czech Verbs*. Ph.D. thesis, Charles University in Prague.

Žabokrtský, Zdeněk and Otakar Smrž. 2003. Arabic Syntactic Trees: from Constituency to Dependency. In *EACL 2003 Conference Companion*, pages 183–186, Budapest, Hungary.

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 88 DECEMBER 2007 31–52

Verb Valency Frames Disambiguation

Dissertation Summary

Jiří Semecký

Abstract

This is a summary of the author's PhD dissertation defended on September 17, 2007 at the Faculty of Mathematics and Physics, Charles University in Prague. Semantic analysis has become a bottleneck of many natural language applications. Machine translation, automatic question answering, dialog management, and others rely on high quality semantic analysis.

Verbs are central elements of clauses with strong influence on the realization of whole sentences. Therefore the semantic analysis of verbs plays a key role in the analysis of whole sentences. We believe that solid disambiguation of verb senses can boost the performance of many real-life applications.

In this thesis, we investigate the potential of statistical disambiguation of verb senses. Each verb occurrence can be described by diverse types of information. We investigate which information is worth considering when determining the sense of verbs. Different types of classification methods are tested with regard to the topic. In particular, we compared the Naïve Bayes classifier, decision trees, rule-based method, maximum entropy, and support vector machines. The proposed methods are thoroughly evaluated on two different Czech corpora, VALEVAL and the Prague Dependency Treebank. Significant improvement over the baseline is observed.

1. Introduction

Natural language processing (NLP) research has already grown up from the early phases of its life. Many tasks concerning the early stages of the linguistic analysis of written text, including lemmatization, morphological tagging and surface parsing, might today be considered sufficiently resolved for the mainstream NLP languages. Even if their development will probably further continue to improve, their current results are near to approaching the upper limits and they are already good enough for many practical applications.

However, the complex linguistic applications, including machine translation, question answering, dialog systems, information retrieval, and others need a deeper semantic analysis of

Please cite this article as: Jiří Semecký, Verb Valency Frames Disambiguation: Dissertation Summary. The Prague Bulletin of Mathematical Linguistics No. 88, 2007, 31–52.

text which is becoming the center of interest for current NLP research. Such an analysis tries to understand and describe not only the structure of text but also its meaning. But not all parts of speech are equally important for deep analysis.

Verbs have special roles in the analysis of text. From the syntactic point of view they are the central elements of clauses with direct influence on the presence and realization of other constituents. From the semantic point of view they are the bearers of events and their proper analysis is fundamental for a correct analysis of the rest of the sentence.

Moreover, verbs are also interesting from the linguistic perspective because they have the richest syntactic structure and also the highest level of ambiguity compared to other parts of speech.

Let us take a highly ambiguous Czech verb *dát* as an example. If we want to translate the verb into English, the most obvious translation will be *to give* as in the sentence:

*Petr **dal** Janě knihu. = Peter **gave** Jane a book.*

If we use the verb in combination with a reflexive particle *si* it changes the meaning of the sentence, and the verb needs to be translated as *put*:

*Petr **si dal** klíče do kapsy. = Peter **put** his keys in his pocket.*

Even with the same syntactic structure, we can get a completely different meaning which, again, translates differently:

*Petr si dal Guinness do půllitru. = Peter **ordered** a pint of Guinness.*

Needless to say, that when used in an idiomatic expression, the verb has a completely different translation:

*Petr **si** na tom **dal** záležet. = Peter **made** a point of it.*

*Petr **dal** na jeho slova. = Peter **took** what he said **into** account.*

*Petr **se dal** konečně dohromady. = Peter finally **got** better.*

As has been illustrated, the same Czech verb may have different English equivalents, depending on the sense in which it is used. Therefore, the correct assignment of the sense seems to be essential for the translation of the sentence. For other applications dealing with the semantic content of the text, it is naturally important, too, to take these differences into account.

Our contribution concerns the process and methods of automatic selection of the proper sense of verbs in their given contexts, i.e. verb disambiguation[1] according to a certain definition of verb senses.

[1] *to disambiguate = "to remove uncertainty of meaning from"* (Oxford Dictionary)

Czech is one of the languages which are the center of study of the world-wide computational linguistic community. A significant reason for this is the fact that there is a large amount of high-quality linguistically annotated data. As there are only ten million Czech native speakers, other languages, mainly English, Chinese, French, Spanish, and Arabic definitely receive more attention because of the far larger number of target users. However, the Czech language surely has the highest ratio of linguistically annotated tokens per native speaker[2].

In our experiments we use two Czech corpora:

First, **VALEVAL**, a small but reliable corpus, containing a few thousand running verbs in contexts annotated by three annotators in parallel. The corpus was put together as a lexical sampling experiment for an existing valency lexicon, and contains sentences randomly selected from the Czech National Corpus. Only the selected verbs are annotated in the corpus. The sentences are not selected in any larger continuous blocks except for a small context attached to each annotated unit. Only the golden part of the corpus was taken into account in our experiments. This assured highly reliable labeling which had, however, low coverage and does not respect the actual verb distribution.

Second, the tectogrammatical part of the **Prague Dependency Treebank 2.0**, a large corpus, containing almost 70,000 verb tokens[3]. The tectogrammatical annotation layer describes many linguistic characteristics, including valency which was used as an approximation of verb senses as is explained below. Each sentence of the relevant portion of the Prague Dependency Treebank was annotated on the tectogrammatical layer by one annotator only, i.e. no parallel annotations were performed. Therefore, the quality of the valency annotation is not guaranteed to be as high as for the first corpus. On the other hand, the quantity highly exceeds VALEVAL and the distribution of verbs reflects the real distribution in Czech (newspaper) text.

Our disambiguation process can be simply described by a sequence of the following steps. First, we automatically analyzed linguistically the sentences containing the annotated verbs. Second, we created a vector of features for each annotated verb in the dataset, describing its context. We experimented with a large number of different features, a great attention was paid to the comparison of individual feature types. Third, the generated features were used in machine learning algorithms. Again, we experimented with several machine learning methods, including the Naïve Bayes classifier, decision trees, rule-based learning, support vector machines, and maximal entropy model. Finally, we evaluated the obtained results. In the evaluation section, we stated the results obtained by using all types of features separately, as well as using their different combinations. Also the difference in performance of individual classification methods are evaluated, as well as several other aspects.

[2] We state here this claim without precise proof, and assuming the exclusion of dead (or nearly dead) languages where the ration is (or approaches) infinity, even with a very limited corpus.

[3] The number refers only to the portion annotated on the tectogrammatical layer.

2. Word Senses

In this section, we show that what we are going to disambiguate in this work are actually not senses of verbs but their valency frames. We explain this approximation and show that under a specific assumption it does not really matter so much.

We have worked with two different lexicons, namely VALLEX, and PDT-VALLEX.

For building a statistical word sense disambiguation system, two types of data resources are needed – a lexicon defining word senses and a corpus annotated with the senses of this lexicon.

We have decided to modify the task slightly by approximating verb senses with verb valency frames. Valency is a property of verbs which correlates with the senses to a certain extent, it is formally well defined and there are lexical resources of sufficient size available describing and using verb valency. In the following paragraphs, we point out that in our choice of valency frame lexicons, the correlation between frames and senses is relatively high.

2.1. Valency

Valency (Panevová, 1980), (Panevová, 1974), (Panevová, 1994) is the ability of a lexical item to combine with another lexical items in syntactic structures. The valency is defined for four different parts of speech — verbs, substantives, adjectives and adverbs. There is no doubt that the valency of verbs is the most differentiated and therefore the most interesting for studying. In this work we are only concerned with verb valency, leaving the valency of other parts of speech aside.

Valency is described in terms of **valency frames** which defines the ability of the given lexical item to syntactically combine with other lexical items. From a technical point of view a valency frame is usually described by a central lexical item (predicate, frame evoking element, …) and a list of participants of the frame (arguments, frame elements, …) corresponding to individual lexical items linked to the central element described by their linguistic (usually morphological and syntactic) characteristics and semantic labels. Different configurations of participants imply different valency frames. The participants are further categorized in different ways, depending on the concrete valency theory (e.g. usually distinguishing the level of obligatoriness).

2.2. Approximation of senses

The valency lexicons built at the Institute of Formal and Applied Linguistics in Prague – VALLEX and PDT-VALLEX (introduced in Section 3.1) – are, however, different from the general definition in this point: the **clearly different senses of a verb with equal valency frames are distinguished in the lexicon**. The following examples demonstrate this statement:

VALLEX:
- **Frame** 1 : **ACT**$_1$ **PAT**$_4$
  absolvovat studium
  graduate from a place

- **Frame** 2 : **ACT**$_1$ **PAT**$_4$
  absolvovat operaci
  undergo an operation

PDT-VALLEX:
- **Frame** v-w1184f1 : **ACT**$_1$ **PAT**$_4$
  chová prasata na farmě.LOC
  He breeds pigs on the farm.

⋮

- **Frame** v-w1184f4 : **ACT**$_1$ **PAT**$_4$
  chová dítě v náručí.LOC
  He cuddles the child in his arms.

When the difference in the meaning was not clear, frames did not have to be differentiated which corresponds to the uncertainty in the sense distinction.

From this perspective, **verb sense** (without any precise definition) **is a function of frames** (in VALLEX and PDT-VALLEX). The frame distinction in these lexicons is in fact driven by the combination of the valency and sense characteristics. Therefore these frames can be used as a suitable approximation of senses.

For the automatic assignment of word senses we need a lexicon containing formal definitions of senses. As already suggested above, instead of using such lexicons we are using lexicons of valency frames which take senses distinction into account.

3. Data resources

In this section, we introduce the data which we used or referred to in the experiments discussed in the thesis – two valency lexicons together with two corresponding corpora. The lexicons define the senses of verbs and the corpora use those lexicons to annotate the verbs.

3.1. VALLEX and VALEVAL

3.1.1. VALLEX

VALLEX(Žabokrtský and Lopatková, 2004) is a manually created valency lexicon of Czech verbs, which is based on the theoretical framework of Functional Generative Description.

The construction of VALLEX started in 2001 and the work is still in progress. The VALLEX

version 1.0 [4] (autumn 2003) (Lopatková et al., 2003) which we used in our task and which was published in 2003, defines valency for over 1,400 Czech verbs and contains over 3,800 frames. In 2005, the VALLEX version 1.5 was published, containing roughly 2500 verbs with more than 6000 valency frames. At the time this thesis is submitted, the new version 2.0 of the VALLEX is about to be published.

The basic structure of the VALLEX lexicon is shown in Figure 3.1.1[5]. Elements of the chart are described in more detail below.

3.1.2. VALEVAL

The manually annotated corpus VALEVAL (Bojar, Semecký, and Benešová, 2005) was created in 2005 as a lexical sampling experiment for the VALLEX lexicon. It contains frame annotations for 109 base lemmas selected from VALLEX. The term **base lemma** is used for a lemma excluding its possible reflexive particle.

For all verbs in VALEVAL, their aspectual counterparts, including iterative forms, were added, too. For each base lemma, 100 sentences from the Czech National Corpus[6] (Kocek, Kopřivová, and Kučera, 2000) (a large corpus containing over 100 million of words) were randomly selected to be present in VALEVAL. This selection resulted in an average number of frames per base lemma of 6.77 (according to VALLEX definition).

[4] http://ckl.ms.mff.cuni.cz/zabokrtsky/vallex/1.0/

[5] The rough structure of PDT-VALLEX is the same as that of VALLEX.

[6] http://ucnk.ff.cuni.cz/english/index.html

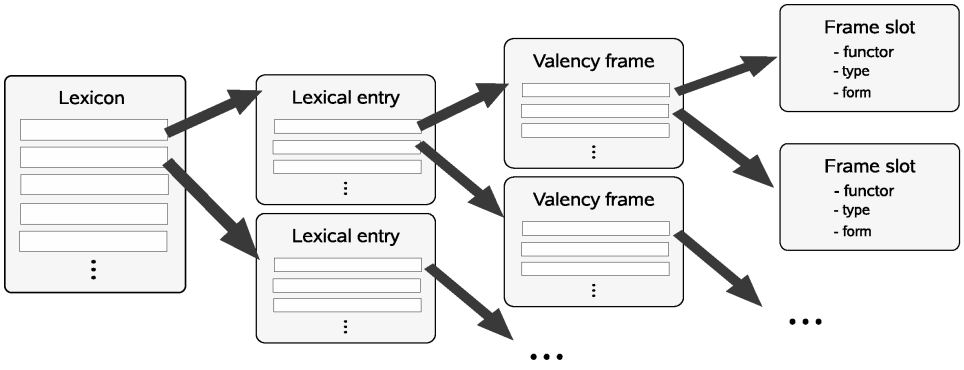


Figure 1. Structure of VALLEX and PDT-VALLEX lexicons.

3.2. Prague Dependency Treebank

The Prague Dependency Treebank (PDT) (Hajič, 2004) is a manually annotated corpus based on the theory of Functional Generative Description (FGD). Data of the PDT are part of the Czech National Corpus (Kocek, Kopřivová, and Kučera, 2000).

Data are annotated on three different layers (Hajičová, 2002), namely morphological, analytical, and tectogrammatical. This differs from the original definition of layers in the FGD.

The current version of the Prague Dependency Treebank is the version 2.0 published by the Linguistic Data Consortium in late 2006 under the number *LDC2006T01*.

Different layers contain different amounts of data. The data are organized so that each part annotated on a higher level is also annotated on all lower levels.

Moreover, the data in each section are divided into the training part, the development testing part (*dtest*), and the evaluation testing part (*etest*). The training part contains approximately 80% of the entire portion, the testing parts each contain approximately 10% of the data.

As frame annotation belongs to the tectogrammatical level, we were restricted to the tectogrammatically annotated portion of the data.

3.2.1. PDT-VALLEX

PDT-VALLEX (Hajič and Honetschläger, 2003), (Hajič et al., 2003) is a valency frames lexicon, created as a part of the PDT. It contains the definition of valency frames for four parts of speech – verbs, nouns, adjectives and adverbs. The PDT-VALLEX was created during the annotation and it contains all auto-semantic words occurring in the corpus. The lexicon was dynamically updated as the annotation went on, unlike VALLEX, described above.

4. Feature Design

To disambiguate a word or a phrase, we are looking at linguistic characteristics within its context. In our work, we look at the sentence in which the verb occurs.

The linguistic characteristics of a sentence are complex structures – trees, vectors, sets, …. On the contrary, machine learning methods can only deal with a simple description of samples, usually vectors.

The natural solution to deal with this contrast is to convert complex linguistic characteristics into simple vectors of features. As the vectors of features only describe linguistic information in a limited way, there will always be a loss of information in the feature creation process. Therefore the selection of a suitable set of features is essential for the success of the method.

4.1. Morphological features

These features are generated only from the morphological information, they are not a result of parsing.

Because syntactic parsing is computationally much more demanding than morphological tagging, those features are very simple and easy to obtain.

The morphological features are based on the Czech positional morphology (Hajič, 2000) used in the Prague Dependency Treebank. The morphological tags consist of 15 positions (characters), each stating the value of one morphological category.

In this work, we use all positions of the morphological tags, except positions 13, 14, and 15, which are not actively used.

For lemmas within a n-word window centered around the verb we used each position as a single feature.

Figure 2 shows an example of generation of morphological features for verb *odvolat – remove (from the office)*.

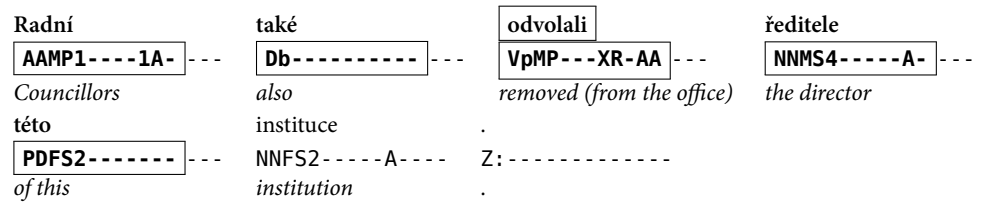


Figure 2. Generation of morphological features.

4.2. Syntax-based features

Syntax-based features, in contrast to the morphological features, are based on the result of the syntactic (analytical dependency) parser.

Syntax-based features also use morphological characteristics, but combine them with the shape of the dependency tree. As the term *syntactic features* might suggest using only syntactic information by analogy with the *morphological features* using only information about morphology, we prefer to use the term *syntax-based features*. Moreover, other types of features (idiomatic, WordNet-based, and animacy) also use the analytical syntax, however, they are in special categories because of their narrow scope.

For our experiments, we did not use a tectogrammatical parser, as we understand verb valency as a part of the tectogrammatical analysis. Therefore the tectogrammatical parsing and subsequent analysis (assignment of tectogrammatical functions) should be processed only after the valency is resolved.

We expected that syntax-based features would be very useful for the disambiguation of the valency frames as the valency frames describe the syntactic behavior of the verbs. Special care was paid to selecting the proper features. Nevertheless, since statistical parsing achieves much lower accuracy than morphological tagging, syntax-based features as opposed to morphological features can suffer much more from errors in analysis.

Based on the results of statistical syntactic parsers we extracted the following groups of features:

- Reflexive *se*
- Reflexive *si*
- Subordinate verb
- Superordinated verb
- Subordinating conjunctions
- Substantives in particular cases
- Adjectives in particular cases
- Prepositional with particular cases

A detailed description of each group follows.

4.3. Idiomatic features

Certain idiomatic expressions evoke a special (usually figurative) senses of verbs. To depict such senses, we introduced this type of features.

Each idiomatic construction (multi-word expression) described in the VALLEX lexicon was used as one boolean feature. This feature was set to *true* if this construction occurred in the raw text of the sentence containing the verb continuously. Features corresponding to non occurring idiomatic constructions were set to *false*.

In this way, we could have missed some idiomatic expressions which were in fact present in sentences but did not occur in a subsequent list of words. This could happen if the writer paraphrased the idiomatic expression. However, simply allowing the inflexion and the gaps in the multiword expression could heavily over-generate and introduce positive errors.

4.4. Animacy features

Animacy is a grammatical category of nouns and pronouns specifying whether the noun or pronoun refers to an animate object.

The introduction of the animacy features was based on an assumption that animacy can often suggest the meaning of the verb. This assumption follows from the fact that some senses of verbs can only describe a relation between (living) beings.

The main problem related to the animacy features is the difficulty of the determination of animacy. There is no simple way to determine animacy automatically, and we can only predict it for specific cases. The algorithm we used for partial animacy resolution differs for nouns and pronouns.

4.5. WordNet features

In some cases, dependency of a certain lemma or a certain type of lemma on the verb can imply a particular sense of the verb. From this perspective, it might be useful to capture the presence of each lemma among the nodes dependent on the verb. However, storing the pres-

ence for all possible lemmas would lead to a huge number of features, to a loss of generality, and possible over-fitting.

There are several possibilities of how to deal with this issue. One of them is, instead of capturing presence of each and every lemma, capturing only the "class" of the lemma. This class should generalize the meaning of each word, so words with a similar meaning should belong to the same class. This solution requires usage of some kind of ontology which maps the lemmas or meanings (disambiguated lemmas) to the classes.

WordNet (Fellbaum, 1998) seemed to be a good choise for this purpose. To define a system of coarse-grained classes of WordNet items (synsets[7]), we used the WordNet top ontology designed at the University of Amsterdam (Vossen et al., 1998). This ontology is described as a tree-based system of 64 WordNet synsets which represents the top of the WordNet hierarchy.

Using hyperonymy relation defined in WordNet we can easily determine all classes to which a given noun belongs, i.e. is related by the transitive relation of hyperonymy. This means that "the noun is type/kind of the class". Because of the transitivity of the hyperonymy relation, if a word belongs to a given class, it also belongs to all classes which are governing this class in the top-ontology.

4.5.1. Combination with Czech WordNet

For each lemma present in the synsets of the top ontology, we used the WordNet **Inter-Lingual-Index** to map the English WordNet to the Czech EuroWordNet (Pala and Smrž, 2004), extracting all Czech lemmas belonging to the top level classes. After this step we ended up with 1564 Czech lemmas associated to the WordNet top-level classes.

5. Evaluation

This section summarizes the empirical results of the experiments described in this work. We ran several machine learning algorithms on two corpora using various types of features. Because of size, we used cross-validation for the VALEVAL corpus. Moreover, two different ways of counting the overall results for the VALEVAL corpus are considered. In the first one, we computed the average of the results for individual lemmas weighted by the frequencies in the corpus, but in the second one, we weighted the results by the relative frequencies measured in the Czech National Corpus relative frequencies measured in the Czech National Corpus (CNC) (Kocek, Kopřivová, and Kučera, 2000). For the Prague Dependency Treebank, we presented results for two different evaluation data sets – the development test set, and the evaluation test set. We used the development test set throughout the development period and only performed the evaluation on the evaluation data set once, for the purpose of this thesis. After that, we did not modify the methods anymore.

[7] The term **synset** is used in the WordNet for a lexicon item capturing single meaning. One lemma can belong to more synsets (suggesting different meaning of the lemma), as well as one synset can consist of more lemmas.

	VALEVAL		PDT	
	$\oslash_{data}$	$\oslash_{CNC}$	**dtest**	**etest**
Average number of frames	4.45	5.31	2.39	2.27
Baseline	68.27	60.74	73.19	71.98

$\oslash_{data}$ denotes average weighted by the number of sentences in the dataset.
$\oslash_{CNC}$ denotes average weighted by the number of sentences in the Czech National Corpus.

Table 1. Difficulty of the frame disambiguation task

As the baseline of the disambiguation task we took **the relative frequency of the most frequent frame of each lemma in the training data**. For the VALEVAL corpus, we determined the baseline using 10-fold cross validation.

For the Prague Dependency Treebank, the baseline was measured on the testing data (the dtest, and the etest section, respectively) but the most frequent frame was determined from the training data.

We computed the overall baseline as the weighted average of the individual baselines. The overall baseline for the VALEVAL corpus was 68.27% when weighted by the number of sentences in our data set and 60.74% when weighted by the relative frequency in the Czech National Corpus. The overall baseline for PDT was 73.19% for the development testing set and 71.98% for the evaluation testing set. The baseline statistics are summarized in Table 1.

5.1. Results

This section presents the evaluation results of the valency frame disambiguation using each presented type of features separately, as well as different combinations of feature types, computed by different classifiers.

Table 2 shows the results weighted by the relative frequencies in the CNC. Table 3 present the results for the Prague Dependency Treebank for evaluation testing set.

The columns of the tables correspond to different classification methods: Naïve Bayes classifier (NBC), Christian Borgelt's implementation of the decision trees (DTREE), C5 decision trees (C5-DT), and C5 rule-based learning (C5-RB), Support Vector Machines (SVM), and Maximum Entropy (ME). The rows of the table correspond to different types of features, the first five rows state the results when using each type of features separately, the following rows state the results for different combinations of the type.

The best accuracy on VALEVAL – 77.56% – was achieved by the C5 rule-based algorithm using the full set of features.

5.2. Methods Comparison

Different methods achieved different results on different data. Generally, we can claim that the C5 decision trees, C5 rulesets, Support Vector Machines and the Maximum Entropy

Corpus:	VALEVAL					
Weighting:	Relative frequencies in the Czech National Corpus					
Type of features	**NBC**	**DTREE**	**C5-DT**	**C5-RB**	**SVM**	**ME**
Baseline	60.74					
Morphological (M)	61.62	59.81	67.50	67.83	58.48	66.36
Syntactic (S)	69.98	69.34	71.01	70.43	67.90	68.51
Animacy (A)	52.87	59.86	62.32	62.67	55.12	59.60
Idiomatic (I)	60.89	60.21	61.01	61.10	60.96	62.77
WordNet (W)	45.32	53.62	58.34	59.22	50.72	54.30
M + S	63.52	60.25	69.69	69.15	63.34	64.11
M + I	61.65	59.81	67.77	68.40	58.61	63.65
S + W	59.37	60.85	71.28	70.87	60.60	61.70
S + A	63.44	61.67	70.56	70.56	63.96	63.26
S + I	69.42	69.61	70.96	70.55	68.03	69.95
M + S + I	63.52	60.25	69.27	68.54	63.43	68.76
M + S + A	63.13	58.19	69.91	69.46	64.39	64.74
M + S + W	64.80	60.28	76.61	75.08	65.27	62.62
S + A + W	60.68	61.43	70.65	71.07	58.75	65.05
S + A + I	63.32	61.67	70.95	71.31	64.04	67.22
S + I + W	59.63	60.94	71.10	71.23	61.57	65.84
M + S + I + W	64.78	60.28	76.90	77.25	65.30	63.62
M + S + A + W	64.59	58.36	76.85	77.10	62.62	67.51
S + A + I + W	60.78	61.43	71.33	71.31	58.67	64.65
M + S + A + I + W	64.58	58.36	76.97	77.56	62.64	67.45

Results are obtained by weighting individual results with the relative frequencies in the Czech National Corpus.

Table 2. Accuracy [%] of the frame disambiguation task for VALEVAL corpus.

model achieved comparably good results throughout the experiments. As has already been mentioned, we did not expect the Naïve Bayes classifier to beat other state-of-art methods. The second implementation of the decision trees algorithm (DTREE) also did not achieve results comparable with C5.

The **C5** algorithm proved to be a reliable classification method. Compared to other methods, it performed well even if the number of training samples was low. When the number of samples was higher, the Maximum Entropy models tended to outperform C5.

C5 decision trees and rule-sets are comparably powerful, sometimes one scores slightly better, sometimes the other one does. The differences are usually not significant. Still, the rule-sets seemed to work slightly better in our tasks, which corresponds to the statement of the C5's authors. On the PDT evaluation test set, both C5 algorithms achieved the same result (78.06%).

The C5 method showed some gain even with very poor feature sets (animacy or idiomatic features alone), compared to other methods which usually scored below the baseline. As a matter of fact, the C5 methods never scored worse than the baseline, which does not hold for any other method examined.

Corpus:	PDT - etest					
Weighting:	Sample counts in the corpus.					
Type of features	**NBC**	**DTREE**	**C5-DT**	**C5-RB**	**SVM**	**ME**
Baseline	71.98					
Morphological (M)	73.03	73.72	73.66	73.62	72.55	74.59
Syntactic (S)	77.84	77.89	77.47	77.35	78.63	78.60
Animacy (A)	70.23	71.05	72.37	72.37	71.99	71.44
Idiomatic (I)	72.45	72.26	72.49	72.49	72.59	72.35
WordNet (W)	68.04	70.41	72.14	72.09	70.15	70.58
M + S	75.24	75.18	77.48	77.54	76.78	78.06
M + I	73.30	73.73	73.66	73.73	72.82	74.89
S + W	74.89	76.43	77.66	77.50	76.35	76.85
S + A	76.19	74.22	77.51	77.40	77.19	77.70
S + I	78.17	78.15	77.76	77.66	78.88	78.85
M + S + I	75.18	75.22	77.71	77.80	76.89	78.10
M + S + A	75.52	75.09	77.25	77.33	75.75	78.09
M + S + W	75.72	74.97	77.60	77.75	76.46	78.17
S + A + W	75.12	73.61	77.00	76.93	75.37	76.89
S + A + I	76.45	74.38	77.75	77.61	77.42	78.04
S + I + W	74.98	76.68	77.80	77.66	76.56	76.95
M + S + I + W	75.79	75.00	78.06	78.06	76.70	64.48
M + S + A + W	75.67	75.10	77.74	77.76	75.93	78.00
S + A + I + W	75.35	73.74	77.57	77.50	75.51	77.07
M + S + A + I + W	75.51	75.13	77.91	78.04	76.10	78.26

Table 3. Accuracy [%] of the frame disambiguation task for the evaluation test set of the Prague Dependency Treebank.

Support vector machines is a popular classifier which is in general performing well. However, it requires a fine tuning of the parameters.

In our experiments, the linear kernel always scored best. This can be explained by the fact that we largely used boolean features which could be easily separated by a superspace in the linear space. Using a more sophisticated kernel adds freedom in the methods which makes the classifier more difficult to train. If there were more real-number features, the situation would probably differ. However, linguistic characteristics are rarely described by real-number features.

The support vector machines achieved the absolutely best result on both, the development and the evaluation testing dataset of the Prague Dependency Treebank.

5.3. Features Comparison

This section gives comparison of individual types of features.

Tables 2 and 3 show that the syntax-based features (see Section 4.2) clearly performed best

in all datasets. They contain most of the information which is linguistically relevant to the valency.

The morphological features turned out to be the second best. The strong difference between syntax-based and morphological features shows how much the statistical parsing helps to analyze the meaning of the verbs. The remaining feature types achieved similar results, usually in the following order: idiomatic features, animacy features, WordNet features.

When we look at the combination of syntax-based features with another type of features, the best result was achieved with the idiomatic features, while the combination with morphological features usually performed worst. In our opinion, this is because the information stored in the morphological features is already included in the syntactic features and adding it does not bring any new information. On the other hand, the other types of features contain information of a different kind, hence they help the syntactic features when combined.

5.4. Differences in Words

The success of the disambiguation task is not flat across all the verbs, it differs from one verb to another, according to the characteristics of the given verb. Most of the verbs have a single dominant sense which is assigned to the majority of the running verbs. Typical examples are the verbs *být* (the most frequent Czech verb), *říci* or *začít*. There are, however, other verbs, whose different senses are widely spread and used in the language. Typical examples are the verbs *mít* (the second most frequent Czech verb), *dát*, or *vědět*.

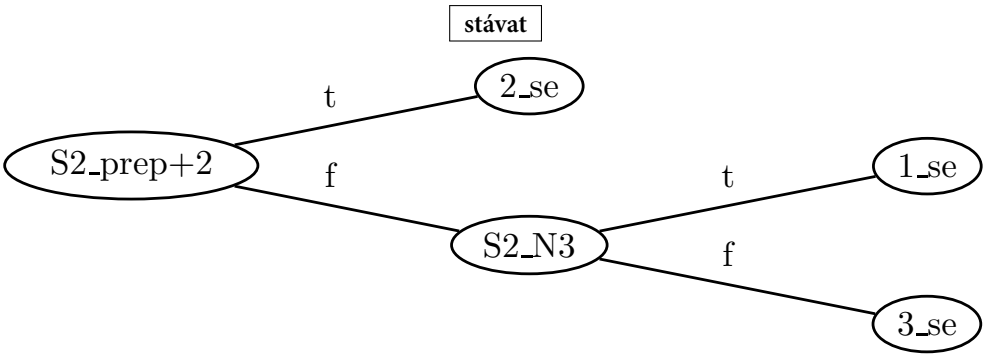
In the following sections, we present decision trees generated by the C5 algorithms. We have chosen decision trees because it is a white-box model, so they clearly show how the classifier works.

5.4.1. VALEVAL

The C5 decision trees scored worse than the baseline for eight verbs in the VALEVAL corpus. The following table lists the verbs with possible explanations of the failures:

zachytnout	(29 % loss)	low number (7) of training samples (4 frames)
spojit	(3 % loss)	high number (6) of frames
držet	(3 % loss)	high number (8) of frames
přidat	(2 % loss)	high number (7) of frames
ponechávat	(1 % loss)	
stávat	(1 % loss)	

Figure 3 shows the decision tree for the verb *stávat*, the decision trees for the other verbs from the previous list are not interesting.



S2_prep+2 …presence of a preposition in genitive dependent on the verb
S2_N3 …presence of a dative noun dependent on the verb

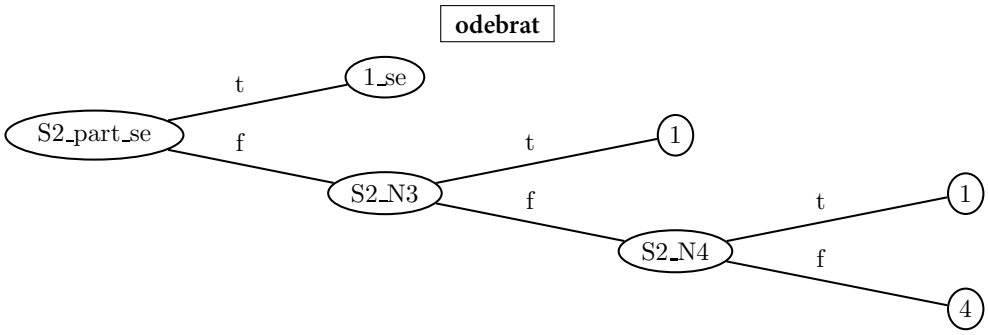
1_se	přiházet se; uskutečňovat se (Eng: *happen*) • často se mi stávalo, že jsem přišel pozdě → *lit.* it often happened to me that I came late
2_se	přeměňovat se (Eng: *become*) • pomalu se z něj stávala příšera → *lit.* slowly he became a monster
3_se	přeměňovat se v něco (Eng: *change into*) • z chlapce se stával mužem → *lit.* from a boy he changed into a man

Figure 3. Decision tree for the verb stávat from VALEVAL.

The verbs with the highest performance gain ($accuracy - baseline$) were the following:

odebrat	(48 % gain)
stát	(43 % gain)
určit	(35 % gain)
přihlížet	(33 % gain)
vyvíjet	(32 % gain)
udržovat	(31 % gain)
připadnout	(31 % gain)
orientovat	(31 % gain)
dát	(31 % gain)
umístit	(30 % gain)
vyvinout	(30 % gain)
přiznat	(30 % gain)

Figures 4 and 5 show the decision trees for the verb *odebrat* and *udržovat* respectively.



1_se	odejít; vydat se (Eng: *leave (for)*)
	• odebral se na schůzi
	→ *lit.* he left for a meeting
1	odejmout (Eng: *to take away*)
	• odebrali jí děti
	→ *lit.* they took away from her the children
4	odkoupit; převzít (Eng: *to buy, to take over*)
	• odebrali všechno objednané zboží
	→ *lit.* they bought all ordered goods

Figure 4. Decision tree for the verb odebrat from VALEVAL.

5.4.2. PDT

The C5 decision trees scored worse than the baseline for 64 verbs out of 1712. The verbs with the lowest performance were the following:

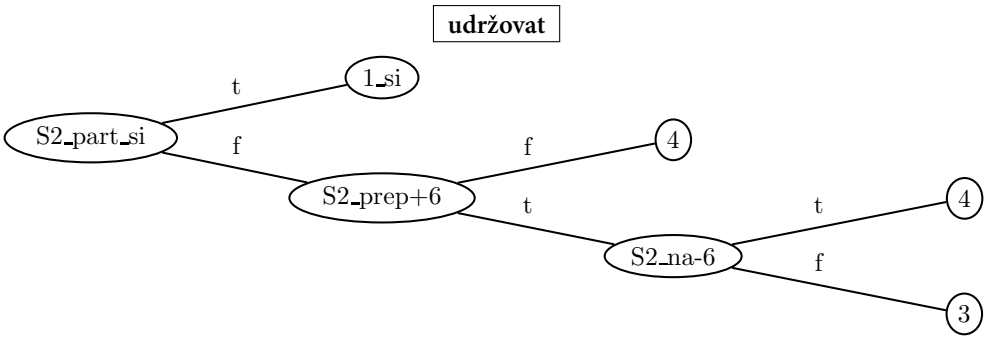
znát, *držet*, *učinit*, *přijímat*, *předpokládat*, *růst*, *fungovat*, *vyhrát*, *přinést*.

The most often reason for the fails were a low number of training data (unreliable classifier) or testing data (unreliable result), high number of frames compared to the size of training data (e.g. verb *držet* – 18 frames for 55 running verbs) and inability to distinguish two frames.

The verbs with the highest positive influence on the total performance ($accuracy-baseline$) were the following (in this order):

být, *mít*, *stát*, *dostat*, *rozhodnout*, *myslit*, *dát*.

Figures 6 and 7 show examples of decision trees for the verbs *rozhodnout* and *dělit*, respectively.



3	zachovávat v určitém stavu (Eng: *to keep in a certain state*) • udržoval byt v čistotě → *lit.* he kept the flat in good order
4	dodržet; uchránit; pečovat (Eng: *to maintain*) • udržoval kázeň / pořádek / kontakty / zahradu → *lit.* to maintain order
1_si	zachovávat (Eng: *to preserve, to keep*) • udržoval si nadhled / kondici → *lit.* he kept condition (up to the mark)

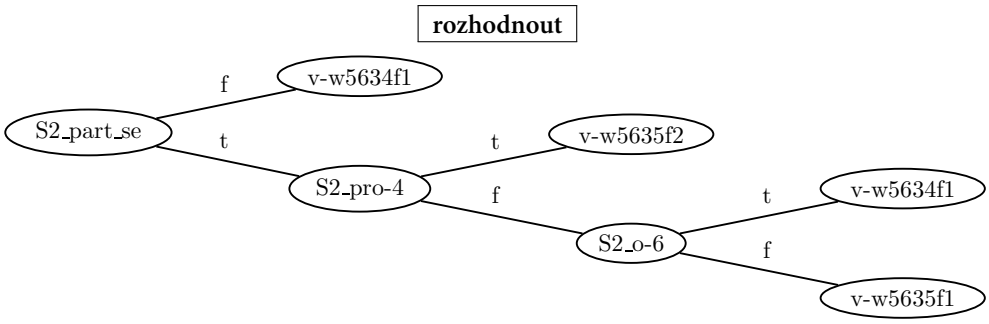
Figure 5. Decision tree for the verb udržovat from VALEVAL.

6. Conclusion

The disambiguation of verb senses in Czech has been extensively studied in this thesis. Different machine learning methods and different approaches to WSD and related tasks were introduced.

We investigated which type of information is important to consider when determining the sense of verbs. In fact, instead of senses we used the valency frames. Each verb occurrence was described by hundreds of features of five basic types. The types of the features were evaluated separately and compared to each other. The most important features turned out to be the ones using information about the surface syntax.

Experiments using different machine learning methods were performed, including the Naïve Bayes Classifier, decision trees, rule-based methods, Maximum Entropy model, and Support Vector Machines. The methods were validated on two qualitatively and quantitatively different corpora — the VALEVAL corpus and the Prague Dependency Treebank. For the smaller VALEVAL corpus, the C5 decision trees and rule-based methods turned out to be the most accurate. For the large Prague Dependency Treebank, the support vector machines and maximum entropy model performed better than other methods.



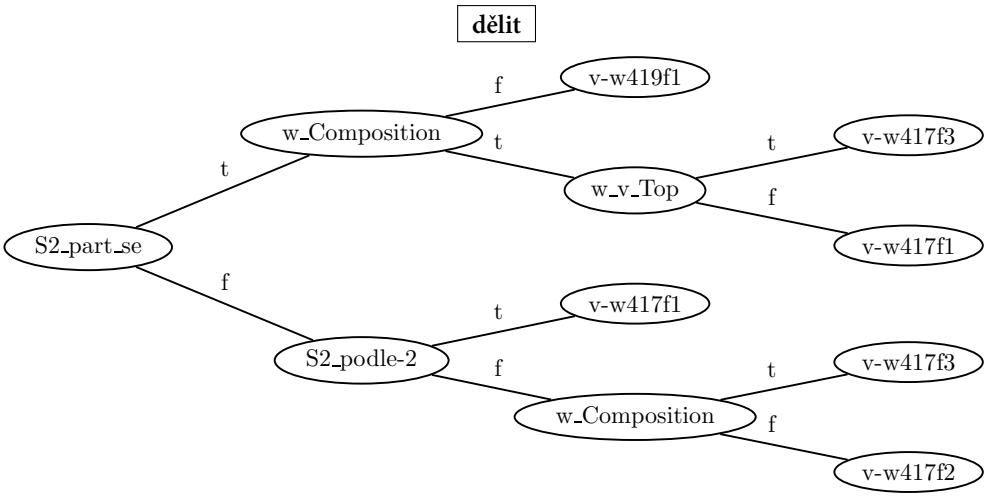
v-w5634f1	určit (Eng: *to decide*) • rychle rozhodl o jeho přijetí → *lit.* to decide on his admission • r. přijmout všechny • r., kam půjdeme
v-w5635f1 (Eng: *to decide*)	• rychle se rozhodl o dalším postupu → *lit.* to quickly decide where to go • r. se přijmout opatření • r. se, kam půjde • r. se rychle, jestli mu vydají....
v-w5635f2	volit, vybrat (Eng: *to choose*) • rozhodnout se pro Prahu mezi dvěma možnostmi → *lit.* he choose Prague as one of the two possibilities • r. se pro Karla

Figure 6. Decision tree for the verb rozhodnout from PDT.

On the VALEVAL corpus, we achieved improvement 12% absolute over the baseline. On the more challenging Prague Dependency Treebank, improvement 6.5% absolute over the baseline was measured on both the development and the evaluation testing set.

In the evaluation section we investigated the results from different perspectives giving alternative analysis and evaluations.

To summarize the thesis, different techniques of disambiguation of verb senses were proposed, implemented and thoroughly evaluated on two Czech corpora. The achieved improvement over baseline validated the correctness of the underlying ideas.



V-w417f1	členit, rozdělit, kouskovat (Eng: *to divide*) • dělit příjmení na části • d. republiku na dva státy • d. salám na poloviny • d. salám nožem v polovině • d. úkol na několik etap → *lit.* to divide the task into several phases
v-w417f2	odloučit Eng: to separate • minuta dělila kajakářku od medaile
v-w417f3	rozdělit, dát, podělit (Eng: *to distribute*) • dělit archívy mezi republiky • dělit dětem dárky → *lit.* to distribute presents among children • d. mezi děti dárky • d. aktivity na střediska, do středisek, střediskům • d. peníze do rozpočtu obcí
v-w419f1	rozdělit se (Eng: *to go share with a person*) • dělil se s příbuznými o majetek • ODS se dělí s ČSSD o politickou moc

Figure 7. Decision tree for the verb dělit from PDT.

Further perspectives Even though this work deals with the disambiguation task, extensively discussing many alternatives, there still remain several directions for the potential extension of the work.

In our opinion, more attention given to the tuning of parameters of non-linear SVM kernels might bring some improvement in performance.

The problem with low number of training samples can be partially avoided by merging aspectual counterparts which often share the valency behavior. However, this might not be applicable for all verbs, and it would require a further exploration. We would also need the mapping of aspectual pairs which is part of the VALLEX lexicon but is missing in the PDT-VALLEX.

The proposed methods might also be further adapted to other languages. However, for languages with limited morphology, e.g. English, a revision of features should be considered, as the current feature set is heavily based on information resulting from morphology.

Acknowledgement This research has been supported in part or in full by the following grants: Grant Agency of the Czech Republic GA405/06/0589 and Ministry of Education of the Czech Republic projects ME 838 and ME 752.

Bibliography

Bojar, O., J. Semecký, and V. Benešová. 2005. VALEVAL: Testing VALLEX Consistency and Experimenting with Word-Frame Disambiguation. *Prague Bulletin of Mathematical Linguistics*, 83:5–17.

Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May.

Hajič, Jan. 2000. Morphological Tagging: Data vs. Dictionaries. In *Proceedings of the 6th Applied Natural Language Processing and the 1st NAACL Conference*, pages 94–101, Seattle, Washington.

Hajič, Jan. 2004. Complex Corpus Annotation: The Prague Dependency Treebank. Bratislava, Slovakia. Jazykovedný ústav Ľ. Štúra, SAV.

Hajič, Jan and Václav Honetschläger. 2003. Annotation Lexicons: Using the Valency Lexicon for Tectogrammatical Annotation. *Prague Bulletin of Mathematical Linguistics*, (79–80):61–86.

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, *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, pages 57–68, Vaxjo, Sweden. Vaxjo University Press.

Hajičová, Eva. 2002. Theoretical description of language as a basis of corpus annotation: The case of Prague Dependency Treebank. *Prague Linguistic Circle Papers*, 4:111–127.

Kocek, Jan, Marie Kopřivová, and Karel Kučera, editors. 2000. *Czech National Corpus - introduction and user handbook (in Czech)*. FF UK - ÚČNK, Prague.

Lopatková, Markéta, Zdeněk Žabokrtský, Karolína Skwarska, and Václava Benešová. 2003. Vallex 1.0 valency lexicon of czech verbs. Technical report, ÚFAL MFF UK.

Pala, Karel and Pavel Smrž. 2004. Building Czech Wordnet. *Romanian Journal of Information Science and Technology*, 7(1–2):pp. 79–88.

Panevová, Jarmila. 1974. On verbal frames in Functional generative description I. *Prague Bulletin of Mathematical Linguistics*, (22):3–40.

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

Panevová, Jarmila. 1994. Valency frames and the meaning of the sentence. *The Prague School of Structural and Functional Linguistics*, pages 223–243.

Vossen, Piek, Laura Bloksma, Horacio Rodriguez, Salvador Climent, Nicoletta Calzolari, Adriana Roventini, Francesca Bertagna, Antonietta Alonge, and Wim Peters. 1998. The eurowordnet base concepts and top ontology. Technical report, Centre National de la Recherche Scientifique, Paris, France, France.

Žabokrtský, Zdeněk and Markéta Lopatková. 2004. Valency Frames of Czech Verbs in VALLEX 1.0. In Adam Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 70—77, Boston. Association for Computational Linguistics.

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 88 DECEMBER 2007 53–72

Information Structure from the Point of View of the Relation of Function and Form

Eva Hajičová

Abstract

The function-form viewpont (means and ends, and the regard to the communicative function) is applied to the analysis of the information structure of the sentence, distinguishing between the semantically relevant topic-focus articulation and its means of expression (morphological, syntactic, prosodical).

1. Introduction

There are two attributes by which the Prague Linguistic School is generally characterized: 'structural' and 'functional'. While 'structural' is a common denominator of several linguistic trends that originated in the first decades of the 20th century following Ferdinard de Saussure's pioneering linguistic approach, the term 'function' was used by de Saussure only quite occasionally. It is supposed to be a distinctive feature of Prague scholars that at the same time as they recognized the necessity to describe and explain the collection of language phenomena as a structured whole rather than as mechanical agglomeration, they emphasized that this structured whole – language – should be understood as a functioning means of communication.

As has been observed already by the founding members of the School (e.g. Jakobson 1963, p. 482, says that "we could hardly find a unifying pattern for the Prague group which would distinguish it as a whole from other scholars …") and by many Praguian linguists afterwards (e.g. Vachek 1966; Novák and Sgall 1962, Sgall 1987, Leška 1999, Daneš 1987; 2006), the Praguian formulations of the guiding principles often differ from author to author or from one writing to another. The following quotation characterizes the situation quite well: "… the Prague group has never formed anything like a dogmatically closed body; while it has been united in the basic acceptance of the structuralist and functionalist standpoint, in matters of implementation of the common principles there has always been a great variety of opinion." (Vachek 1966, p. 8).

Please cite this article as: Eva Hajičová, Information Structure from the Point of View of the Relation of Function and Form. The Prague Bulletin of Mathematical Linguistics No. 88, 2007, 53–72.

However, Jakobson (1963, p. 482) points out that "at the same time, there is a typical drift which ties the work of all these explorers and strictly distinguishes them both from the older tradition and from some different doctrines which found their outspoken expression likewise in the '30's. … this common drift … (aims) toward a means-ends model of language. These efforts proceed from a universally recognized view of language as a tool of communication." This is what Oldřich Leška, one of the outstanding "second-generation" Prague School representatives, reflected in the title of his paper as *unity in diversity* (Leška 1999)..

In our present contribution we focus our attention on the necessity of the application of the function – form viewpoint ('means' and 'ends', and the regard to the communicative function) in the domain of one of the most important contributions of Prague School scholars to linguistic theory, namely the study of the information structure of the sentence.

2. Form and function in the mirror of authentic (historical) quotations

Let us first look at the use of the term *functional* and some related terms (relevant for our focus of attention) in two original sources, namely in the collective theses presented to the First International Congress of Slavicists (published in Vol. 1 of TCLP, 5–29) and in Vachek's Dictionary of the Prague School of Linguistics (originally published in 1960; its English translation appeared in 2003).

The following places in the text of the *Thèses* are characteristic for the use of the term function and its derivatives (the numbers at the beginnings of the lines refer to the respective chapters of the *Thèses*):

1.a) Conception de la langue comme système **fonctionnel**

… la langue est un système de moyens d'expression appropriés à **un but**

2.a) … Nécesssité de distinguer le son comme fait physique objectif, comme representation et **comme élément du système fonctionnel**

… les images acoustico-motrices subjectives … remplissement, dans se système, une **fonction différenciatrice de significations**

The entries in Vachek's *Dictionary of the Prague School of Linguistics* (1960, transl. 2003) mostly refer to what the author considered to be the most characteristic or typical uses of the given terms rather than bringing definitions of the head words (collocations); it is no wonder then that some of the entries reflect a certain vagueness of the use and differences between the authors quoted. We do not reproduce here the whole entries but just the relevant passsages. Our comments (mostly just abbreviated Vachek's commentary from the Dictionary) are in square brackets.

Function: Skalička 1948b, 139: … the term function is used where the meaning (the function of a word, a sentence) or the structure of semantic units (the function of a phoneme) is concerned [as opposed to Hjelmslev, with whom …"the notion of function is close to the notion of function in mathematics"]

Functional onomatology

"two important parts of linguistic investigation, that of the ways and means of calling selected elements of reality by names, and that of the ways and means organizing these names,

as applied to an actual situation into sentences … we may call these respective sections of linguistics functional onomatology and functional syntax. [Mathesius 36a, 97–98].

Functional sentence perspective

[is not specified in general, just the means of FSP are mentioned as if the very term FSP were 'given']

Form and function in language

'it cannot be denied that form and function are not simply two sides of one thing, but they often intersect. This is … also the essence of homonymy and homosemy, and in my opinion an important impulse for language changes. Though language is a system, the system of language is perhaps never completely balanced. For this reason in analysing language, systems which are too logical and thus too simplifying will fail to some extent." [Mathesius 36b, 50]

Analytical comparison and the functional viewpoint

"If we are to apply analytical comparison with profit, the only way of approach to different languages as strictly comparable systems is the functional point of view, since general needs of expression and communication, common to all mankind, are the only denominators to which means of expression and communication, varying from language to language, can reasonably be brought." [Mathesius 36a, 95]

3. The hierarchy of levels and relations between their units

3.1. Introduction

The need for a systematic and integrated description of the relation of functions and forms has led to conceive the core of language system as consisting of levels the units of which have their functions in that they represent units of the adjacent higher levels, up to the non-linguistic layer of cognitive content. Under this understanding, the relation of means and function is interpreted as "functions as" (in the upwards direction) and "is constructed of" (in the downward direction).

From the methodological standpoint, Mathesius (influenced apparently by Marty) adopted the speaker's point of view and emphasized the necessity to proceed from function to form; i.e. from needs of communication to means of expression (see e.g. Mathesius 1929, p.119 "… functional approach consists in the convergence of linguistics to the standpoint of the speaker"; according to Daneš (1987), in his respect to the communicative needs, Mathesius himself was influenced by sociology). For Mathesius, form is subordinated to function. As duly noted by Novák and Sgall (1964), several questions may arise: are the needs quite common? what are the basic units of such needs? etc.

However, it is possible to take an opposite point of view and to proceed from form to function, which is the method applied in Jakobson's structural morphology. Leška (1995, p.10) notes that such a new arrangement opens the way to a stratification model of language, introduced by Skalička (1935) and fully developed by Trnka (see esp. Trnka1964).

3.2. Relations between units of levels

With the system of levels, two hierarchies have to be distinguished:

the relation between the (units of the) adjacent levels in the hierarchy; Hockett (1961) speaks about the "R" (representation) relation;

the relation between units of a given level: complex units are composed of more elementary units (morph of phonemes, morpheme of semes, word of morphemes, sentence of word forms; Hockett (1961) speaks about the "C" (composition) relation.

As pointed out by Sgall (1987, p. 171), three different approaches how to account for these two hierarchies can be found in the writings of Prague scholars; Sgall's (1967a) original model of functional generative description works with levels based on the hierarchy R and within each level the hierarchy C obtains. For our discussion, we will restrict ourselves to the discussion of the R relation.

A far-reaching significance for the understanding of the relations between units of adjacent levels is the notion of *asymmetrical dualism* introduced by S. Karcevskij (1929). The main idea consists in the recognition that a form and its meaning (or rather function; Karcevskij uses the French term *signification*) do not cover the same field in all their points: the same sign has several functions and the same function can be expressed by several signs. There is always a certain tension between *signifiant* and *signifié* and the asymmetrical dualism of the structure of the sign makes it possible for language to develop.

Another distinction relevant for the understanding of the relations between levels (esp. for the specification of the functions of a given form) is that of ambiguity and vagueness as discussed e.g. by Zwicky and Sadock (1975): it is possible to ask the speakers if two morphemes, or constructions differ in their functions or if they are synonymous. Similarly, two different meanings of a single morph can be distinguished from a single vague meaning. In the former case, rather than in the latter, the speaker is always able to tell which of the two different lexical or grammatical functions s/he had in mind (although not knowing the precise linguistic wording).

4. The communicative role of language and the position of TFA in the function – form hierarchy

4.1. Some historical milestones

The focal point of Mathesius' interest was "functional onomatology" (means employed by language for the purpose of naming) and "functional syntax". In the latter domain, Mathesius understood sentence as comprising a patterning primarily conditioned by the interactively based role the sentence plays in the context, in discourse. His innovative and consistent regard to this role has led to his introduction of the notions of *theme* and *rheme* into syntactic studies, which is one of the fundamental issues discussed in modern linguistic theories up to present.

The writings on what is more generally (and recently) covered by the term *information structure* date back centuries ago; the issue is treated under different terms and this is not always

possible to find a one-to-one mapping between them; also, they receive a slightly different interpretation. However, they share the underlying idea: a description of the structure reflecting the functioning of language in communication, which is different from the subject-verb-object structure (described in any formalism). One of the oldest and most stimulating, not only for its time, is Weil's (1844) comparison of the means expressing information structure in languages of different types. Of great interest is his proposal to distinguish two types of 'progressions' of sentences in a discourse, in relation to which part of a given sentence serves as a starting point for the subsequent one. Sentences may follow each other in a parallel mode, i.e. they share their starting points (*marche parallèle*), or in a sequential mode, i.e. the starting point of a given sentence follows up the second (final) part of the preceding sentence (*progression*). In more modern terms, one can say that in the parallel mode, the sentences share their themes (topics), in the sequential mode the theme (topic) of one sentence relates to the rheme (focus) of the preceding sentence. (It should be noted that more than one hundred years later, a similar, though a more subtle approach was developed by Daneš 1970 in his paper on thematic progressions).

It is not our intention here to present a historical survey; let us only mention that though the first hints for a systematic treatment of these issues within structural linguistics were given by Vilém Mathesius and later continued (on the initiative of Josef Vachek) by Jan Firbas, one should not forget that the topic was, so to say, hanging in the air, receiving attention esp. in German linguistics (for a more detailed discussion, see Sgall et al., 1973, 1980 and 1986).

With the entrance of formal linguistics on the scene, it is not surprising that the first suggestions for the inclusion of TFA into an integrated formal description of language came from Prague; Sgall's Functional Generative Description (Sgall 1967a) working with a tectogrammatical (underlying, deep) level of sentence structure has incorporated the TFA opposition into the description of this level (Sgall 1967b).

An important terminological (but not only terminological) side-step is in place at this point. As Svoboda duly notes (Svoboda 2003) Mathesius' Czech term *aktuální členění větné* is not directly translatable into English; Firbas – on the advice of Josef Vachek (Firbas 1992, p. xii) and apparently inspired by Mathesius' use of the German term *Satzperspektive* in his fundamental paper from 1929 – changed it into *functional sentence perspective* (FSP). However, this is not the only name under which this domain of research entered linguistics: German researchers often speak about *Thema-Rhema Gliederung*, M.A.K. Halliday, one of the leading European linguists who has been influenced by the Praguian theory, speaks about information subsystem (Halliday 1967) or information structure (reflecting the given-new strategy) distinguishing it from thematic structure (Halliday 1970); another pair of terms used are topic and comment, etc. These terminological differences often indicate some notional distinctions, as is the case of the Praguian theory of *Topic-Focus Articulation* (TFA) we subscribe to. TFA is not a mere "translation" or "rephrasing" of the term FSP; a different term was used basically to indicate certain differences in the starting points: Firstly, theme was originally defined by Firbas as the item that carries the lowest degree of communicative dynamism; if understood in this way, the existence of sentences without a theme (so-called topicless sentences in linguistic literature, or hot-news) would be excluded (every sentence *has* an item with a lowest degree of commu-

nicative dynamism); to avoid such a misunderstanding, we used the term topic rather than theme. (Firbas 1992, however, modifies his definition of theme by adding that in the absence of theme, the lowest degree of CD is carried by the first element of non-theme – referring to Sgall's objection against his original definition of theme made at a FSP conference in Sofia in 1976). Second, even though we accept the postulate that every item in the sentence carries a certain degree of CD, our analysis of negation gives an indisputable support for understanding TFA as based on the 'aboutness' relation, i.e. not just on the degree of CD but on the opposition of contextual boundness (see Sect. 3 below) and also on (as a derived notion, though) the notion of a bipartition (the focus of a sentence conveys some information about its topic). Third, certain notions have been found formulated more precisely in the TFA theory than in Firbas' insightful writings. As Sgall (2003, esp. pp. 281ff) writes, this concerns differences in the nature of the four factors of linear arrangement, prosody, semantics and contexts (the first two belonging to the means of expression of information structure and the other two to its functional layers), as well as CD and contextual boundness. And last but not least, as will be discussed below, in our understanding, TFA is a structure belonging to the underlying, deep structure of sentences (tectogrammatical, in our terms).

It should be noted that the examples serving as arguments during the split of generative transformational grammar into interpretative and generative semantics reflected the difference in TFA (actually, on both sides of the dispute, though not recognized as such; see e.g. Chomsky 1971 and Lakoff 1971a, to name just the main figures). A "breakthrough" on that side of Atlantic was Mats Rooth's doctoral dissertation on association with focus (Rooth 1985), in which the author (referring i.a. to Jackendoff 1972) quite convincingly argues for the "semantic effect of focus" in the sentence offering the explanation of this effect in terms of a domain of quantification (p. 197); his starting arguments were restricted to the presence in the sentence of the so-called focusing particles such as *only, even,* but he extended his proposal also to the so-called adverbs of quantification (*often, always*) and cases such as cleft constructions in English.

The interest was aroused, and after Barbara Partee's (who was one of Mats Rooth's supervisors) involvement in the discussion of the semantic consequences of different TFA structures (see e.g. Partee 1991) the TFA issues took up an important position in the discussions of formal semanticists (for a Czech contribution to that discussion see Peregrin 1994; 1996), but not only within that domain (quite noticeable is the interest in the TFA issues in German linguistics).

One of the crucial contributions of the above mentioned discussions was the due respect to the reflection of the differences in TFA in the prosodic shape of the sentences (which view, actually, has been present in the Praguian studies of TFA). Let us mention here only Jackendoff's (1972) introduction of the difference in A and B prosodic contour and Rooth's (1985) consistent regard to the placement of the intonation pitch in his example sentences.

4.2. The position of TFA in the function – form hierarchy

To give an answer to the question posed in the title of this section, let us start with some examples (maybe notoriously known). The capitals denote the intonation centre, the names in brackets indicate the source of the examples.

(1)(a) Everybody in this room knows at least two LANGUAGES.
(b) At least two languages are known by everybody in this ROOM. (Chomsky 1957;1965)
(2)(a) Many men read few BOOKS.
(b) Few books are read by many MEN. (Lakoff 1971a)
(3)(a) Londoners are mostly at BRIGHTON.
(b) At Brighton, there are mostly LONDONERS. (Sgall 1967b)
(4)(a) I only introduced BILL to Sue.
(b) I only introduced Bill to SUE. (Rooth 1985)
(5)(a) I work on my dissertation on SUNDAYS.
(b) On Sundays, I work on my DISSERTATION.
(6)(a) English is spoken in the SHETLANDS.
(b) In the Shetlands, ENGLISH is spoken. (Sgall et al. 1986)
(7)(a) Dogs must be CARRIED.
(b) DOGS must be carried. (Halliday 1967)
(c) Carry DOGS. (a warning in London underground, around 2000)
(d) CARRY dogs.

It is not difficult to understand that the pairs of sentences under each number differ not only in their outer shapes or in their contextual appropriateness, but also in their meanings, even in their truth conditions. This difference may be attributed to the presence of quantifiers and their order (with an explicit quantification in (1) and (2) and a more or less explicit in (3) and (4)), but from (5) on, such an explanation is not possible. Also, an exclusive reference to the surface order of the sentence elements would not be correct, as illustrated by (4) and (7).

A more adequate explanation is that based on the relation of *aboutness*: the speaker communicates something (the Focus of the sentence) about something (the Topic of the sentence), i.e. F(T), the Focus holds about the Topic. In case of negative sentences, the Focus does not hold about the Topic: F(T).

A supportive argument for the semantic relevance of TFA can be traced in the discussions on the kinds of entailments starting with the fundamental contributions of Strawson. Strawson (1952, esp. p. 173ff.) distinguishes a formal logical relation of entailment and a formal logical relation of presupposition; this distinction – with certain simplifications - can be illustrated by (8) and (9):

(8) All Johns' children are asleep.
(9) John has children.

If John's children were not asleep, the sentence (8) would be false; however, if John did not have children, the sentence as well as its negation would not be false but meaningless. Thus (9) is a presupposition of (8) and as such it is not touched by the negation of (8).

Returning to the relation of aboutness, we can say that (8) is about John's children, and for (8) to be meaningful, there must be an entity John's children the speaker can refer to.[1]

[1] This need not mean that the entity the sentence is 'about' should exist in the real world, but it should be referentially available; cf. the discussion of the notion of referential vs. existential presuppositions in Hajičová 1976, 55–58, reflected also in Sgall et al. 1986).

The close connection between the notion of presupposition and TFA can be documented by a more detailed inspection of the notion of presupposition, exemplified here by sentences (10) and (11).

(10) The King of France is (not) bald.

(11) The exhibition was (not) visited by the King of France.

It follows from the above mentioned discussions on presuppositions that Strawson's (1964) ex. (10) is about the King of France and the King's existence (referential availability) is presupposed, it is entailed also by its negative counterpart; otherwise (10) would have no truth value, it would be meaningless. On the other hand, there is no such presupposition for (11): the affirmative sentence is true if the King of France was among the visitors of the exhibition, while its negative counterpart is true if the King of France was not among the visitors. The truth/falsity of (11) does not depend on the referential availability of the entity "King of France". This specific kind of entailment was introduced in Hajičová (1972) and was called allegation: an allegation is an assertion A entailed by an assertion carried by a sentence S, with which the negative counterpart of S entails neither A nor its negation (see also Hajičová 1984; 1993, and the discussion by Partee 1996). Concerning the use of a definite noun group in English one can say that it often triggers a presupposition if it occurs in Topic (see sentence (10)), but only an allegation if it belongs to Focus (see sentence (11)).

These considerations have led us to the attempt at a more systematic analysis of the relations between affirmative and negative sentences (Hajičová 1972, 1984, 1993). The scope of negation can be specified, in the prototypical case, as constituted by the Focus, so that the meaning of a negative declarative sentence can be interpreted as its Focus (F) not holding of it, i.e. F(T). In this way it is possible to understand the semantic difference present in (10) and (11).

In a secondary case, the assertion holds about a negative Topic: F(T), see (12) on the reading when answering the question "Why didn't he come?".

(12) He did not come because he was out of money.

Here again, the scope of negation is dependent on TFA: it is restricted to the Topic part of the sentence. The assertion entailed (on this reading) by the *because*-clause in Focus is not touched by negation.[2]

4.3. TFA as an integral part of the underlying layer of linguistic description

The analysis summarized in Sect. 4.2. points out very clearly that TFA undoubtedly is a semantically relevant aspect of the sentence and as such should be represented at a level of an integrated language description capturing the meaning of the sentence (whatever interpretation we assign to the notion of 'meaning'). For the formal description of language we subscribe to, namely the Functional Generative Description, this is the underlying, *tectogrammatical* layer; the tectogrammatical representations of sentences (TRs) are specified as dependency tree structures, with the verb (of the main clause) as the root of the tree. While the labels of

[2]On another possible reading of (12), e.g. if the sentence is followed by *but because he was on his leave of absence,* his being out of money is neither entailed nor negated, i.e. the entailment belongs to the allegations of the sentence, i.e. he might have come for some other reason. The scope of negation concerns Focus, schematically: F(T).

the nodes of the tree are counterparts to the autosemantic words of the sentence, counterparts of function words as well as of grammatical morphemes are just indices of the nodes and the edges of the tree: the morphological values of number, tense, modalities, and so on, are specified by indices of the labels of the nodes. For each node of the TR it is specified whether it is contextually bound or non-bound.[3] The edges of the tree are labeled by underlying syntactic relations (such as Actor/Bearer, Addressee, Patient, Origin, Effect, several Local and Temporal relations, etc.). The appurtenance of an item to the Topic or Focus of the sentence is then derived on the basis of the features cb or nb assigned to individual nodes of the tree (see Sgall 1979).

An underlying structure specified in this way can be understood as the 'highest' level of the language description viewed from the point of view of the hierarchy from function to form. The inclusion of TFA into this level can serve well as a starting point for connecting this layer with an interpretation in terms of intensional semantics in the one direction and with a description of the morphemic and phonemic means expressing TFA (Sgall 2003, p. 280; see also Fig. 1 in Sect. 6 below).

The semantico-pragmatic interpretation of sentences (for which the TRs represent suitable input) may then include an application of Tripartite Structures (Operator - Restrictor - Nuclear Scope), as outlined by B. H. Partee in Hajičová et al. (1998). Let us briefly recall some of the characteristic sentences discussed there (with their relevant TRs) and specify (in a maximally simplified notation) which parts of their individual readings belong to the Operator (O), Restrictor (R) and Nuclear Scope (N) of the corresponding tripartite structures. We assume that in the interpretation of a declarative sentence, O corresponds to negation or to its positive counterpart (the assertive modality) or to some other operators such as focusing particles, R corresponds to Topic (T), and N to Focus (F).

(13) John sits by the TELEVISION.

(13') O ASSERT, R John, N sits by the TELEVISION.

(13") O ASSERT, R John sits, N by the TELEVISION.

From the point of view of TFA, (13) - leaving aside its possible interpretation as a topicless sentence (hot news) - may be analyzed in two ways: either it conveys information about John (i.e. John being its Topic and the rest its Focus), or it conveys information about John's sitting; in the latter case, the dividing line between Topic and Focus will be drawn after the verb. The ASSERT operator (introduced by Jacobs 1984) indicates the assertive modality of the sentence, and the two possible divisions into Topic and Focus are reflected by (13') and (13").

In (14), the particle *only* occupies its prototypical position in the underlying structure, so

[3] A contextually bound (*cb*) node represents an item presented by the speaker as referring to an entity assumed to be !easily accessible by the hearer(s), i.e. more or less predictable, readily available to the hearers in their memory, while a contextually non-bound (*nb*) represents an item presented as not directly available in the given context, cognitively 'new'. While the characteristics 'given' and 'new' refer only to the cognitive background of the distinction of contextual boundness, the distinction itself is an opposition understood as a grammatically patterned feature, rather than in the literal sense of the term. This point is illustrated e.g. by *(Tom entered together with his friends.) My mother recognized only HIM, but no one from his COMPANY.* Both Tom and his friends are 'given' by the preceding context (indicated here by the preceding sentence in the brackets), but in the given sentence they are structured as non-bound (which is reflected in the surface shape of the sentence by the position of the intonation center).

that the focus of the particle is identical with the Focus of the sentence on either reading, i.e. with the verb included in Focus in (14'), and in Topic in (14").

(14) John only sits by the TELEVISION.

(14') O only, R John, N sits by the TELEVISION.

(14") O only, R John sits, N by the TELEVISION.

A contextually bound (*cb*) node represents an item presented by the speaker as referring to an entity assumed to be !easily accessible by the hearer(s), i.e. more or less predictable, readily available to the hearers in their memory, while a contextually non-bound (*nb*) represents an item presented as not directly available in the given context, cognitively 'new'. While the characteristics 'given' and 'new' refer only to the cognitive background of the distinction of contextual boundness, the distinction itself is an opposition understood as a grammatically patterned feature, rather than in the literal sense of the term. This point is illustrated e.g. by *(Tom entered together with his friends.) My mother recognized only HIM, but no one from his COMPANY.* Both Tom and his friends are 'given' by the preceding context (indicated here by the preceding sentence in the brackets), but in the given sentence they are structured as non-bound (which is reflected in the surface shape of the sentence by the position of the intonation center).

Let us just note that in the cases in which Topic or Focus is complex, as illustrated by (15), it is the opposition of contextual boundness that is responsible for the difference: while contextually bound items then belong to the local (partial) R, the non-bound ones belong to the corresponding N.

5. Means of expression of TFA

5.1. Introduction

From the methodological point of view, Mathesius' emphasis on the virtual identity of the facts to be expressed by all languages of the world directs the analyst's attention to the diversity of ways by which these identical facts are referred to in various languages. As Vachek (1966, p.7) notes, this is a specific characteristic of the Prague structuralist conception delimiting it from other structurally oriented linguistic currents (Danish glossematics, American descriptivism).

5.2. The order of words

The most frequently and extensively discussed means of expression of the information structure is the word order. In some approaches, the differences in the information structure are even identified with the differences in the order of words in the surface shape of the sentence; as indicated by our set of examples in (1) through (7) this is not correct; the word order is only one of the means (forms) of the expression of the underlying difference of meaning. This is not only due to the fact that not in all languages the word order is flexible enough to express this distinction. The order of words in the surface shape of the sentence might be the same and yet the sentences acquire different information structure, see (7) above or (15), offered by the late Prof. Ivan Poldauf (pers. comm.):

(15) John and Mary saw an EXPLOSION.
(15') An explosion was seen by JOHN and MARY.
(15") An EXPLOSION was seen by John and Mary.

While either (15) or (15") might be used both if the two people saw the same explosion or each of them saw a different one, the (only, or at least preferred) interpretation of (15') is that the two people saw the same explosion (meaning: there was an explosion John and Mary saw) even though the order of elements in the surface shape of (15') and (15") is the same.

5.3. Sentence prosody

Examples such as (7) and (16) illustrate that sentence prosody, especially the placement of the intonation centre, is as an important way of expression of the TFA differences as word order is. In this respect, the pioneering analyses of M.A.K.Halliday have to be mentioned (dating as back as to Halliday 1967, see his example (7)); it was probably him who first 'exported' the issues relevant for information structure to the other side of the Atlantic. This might be attested by Chomsky's (1965; this example was used for the first time in Chomsky 1957) first reference to 'topic' as a possible source of the semantic distinction between the active sentence (16) and its passive counterpart (16'); the intonation center is assumed to fall on the last word of the sentence.

(16) Everybody in this room knows at least two languages.
(16') At least two languages are known by everybody in this room

Also, it should be acknowledged that in his paper on presupposition and focus as related to his notions of deep and surface structure, Chomsky (1971) consistently took into consideration the position of intonation center (giving it a special graphic notation by capitals). This respect to the prosodic expression is most perspicuously reflected in the above mentioned doctoral dissertation on 'association with focus' by Rooth (1985).

The issues related to the notion of 'association with focus' and its assumed acoustic realization by a pitch accent are connected with such expressions as English 'only', 'also', even'. As indicated by the name of the category of these particles (rhematizers by Firbas, or focusing or focus sensitive particles or focalizers by Rooth, Partee and others), the question can been raised whether these particles always stipulate association with a focused element in their scope, or whether there are contexts in which they can occur without such an association. The dialogue (17) (quoted from Hajičová, Partee and Sgall 1998, p. 153) supports the view that an association of these particles with the Focus of the sentence is not necessarily the case.

(17) A. Everyone already knew that Mary only eats vegetables.
B. If even Paul knew that Mary only eats vegetables, then he should have suggested a different restaurant.

In (b), there are two 'focalizers': one of them, the particle *only*, is associated with the material repeated from the first sentence (A) of the dialogue, the second is the particle *even*. Such a complex situation is referred to in linguistic literature as "second-occurrence focus", SO (for a most recent discussion, see Beaver et al. 2007). It has been empirically testified by Bartels (1997) that the realization of second-occurrence focus (on several acoustic dimensions) is dif-

ferent from the "regular' focus; in a follow-up production experiment reported in Beaver et al. (2007), it was confirmed that not only the SO focus is marked differently from the 'regular' focus but that it is also differs acoustically from the non-focused expressions. In Hajičová, Partee and Sgall (1998), the authors therefore differentiate focus of the focusing particle (i.e. its scope) from the Focus of the sentence (i.e. the part of the sentence the sentence is 'about'). In terms of the above mentioned tripartite structures, the analysis of a complex sentence with two focusing particles is as indicated in (18). If the operator is included in Topic, its own focus (which differs from the sentence Focus in such marked cases) does not cross the boundary between the Topic and the Focus of the sentences.

(18) (What did even PAUL realize?) Even Paul realized that
Jim only admired MARY.

(18') O ASSERT, R (O even, R realized, N Paul), N (O only, R Jim admired, N Mary)

It is, of course, not only the position of the intonation center that should be taken into account in the analysis of TFA. The studies on contrastive topic (see e.g. Hajičová and Sgall. 2004, Veselá, Peterek and Hajičová 2003) covering also instances of the above-mentioned 'second-occurrence focus' convincingly support the view that one should consider the whole intonation contour of the sentence (its F0 characteristics) when deciding on the status of the given elements of the sentence in its TFA. For a very inspiring general discussion of the relation between syntax and prosody see Selkirk (1984; 1995).

It should be noted in the connection of the discussion of the prosodic means of TFA, that it is not always the case that the most dynamic element of Focus is to be prosodically marked: Firbas (1992, p. 176) quotes the English sentence (19) as an example of an 'automatic placement' of the intonation center at the end of the sentence even if it is the subject which is 'rhematic' rather than the end of the sentence.

(19) A boy came into the room.

It is worth mentioning that due to the fact that the grammatically fixed word order of English does not allow to linearly order the elements of a sentence so as to reflect the information structure of the sentence (its CD), even the written form of English has a means to indicate the position of the intonation center in the sentence, namely the use of italics.This has been observed already by Alena Skaličková in the 1970's; her observation reoccurred, surprisingly enough, in a paper by Saldanha (2007), analyzing the use of italics to mark focus in English translations of Spanish and Portugese original texts.

5.4. Syntactic constructions

The best known example of a syntactic construction used as the means of rendering the information structure of an English sentence are the so-called cleft constructions. It is a commonly accepted assumption that the *it*-clefts (in contrast to the pseudo-clefts, sometimes referred to as *wh*-clefts) make it possible to 'prepose' the rhematic element and thus to give it some kind of prominence; the rest of the sentence is then understood as being in a kind of 'shadow', backgrounded. The 'preposing' of the focused element is prototypically accompanied by placing the intonation center on this element. A typical example is (20); as its translation to Czech

in (20') illustrates, there is no need to use a specific construction in Czech (unless in a special emphatic situation), a simple reordering of the elements of the sentence is enough.

(20) It was JOHN who talked to few girls about many problems.

(20') S málo děvčaty mluvil o mnoha problémech HONZA.

Lit. With few girls talked about many problems John-Nominative

Though the above interpretation of the cleft constructions is the one prevailing in linguistic literature on English, it is not the only possible one. As recalled by Dušková (1993), Quirk et al. (1985, p.1379) offer the interpretation of 'divided focus'; the authors assume that the decision which of the two items of 'focus' is dominant ('new') depend on the context. Dušková (1993) compares their example (21) with (21') and suggests that in (21') Frost as the rheme of the it-clause gets more prominence and thus can be regarded as dominant, while in (21) the dominant item is the that-clause.

(21) They hoped that Herbert Frost would be elected and Frost indeed it was that topped the poll.

(21') They hoped that Herbert Frost would be elected and it was indeed Frost that topped the poll.

Cleft constructions may also serve as an additional support for the view that not only the division of the sentence into its Topic and Focus, but also the degrees of communicative dynamism as such play their role in the semantic interpretation of the sentence.

(22) It was JOHN who talked about many problems to few girls.

(22') O mnoha problémech mluvil s málo děvčaty HONZA.

Lit. About many problems talked with few girls John-Nominative

The interpretation (at least the preferred one) of (20) suggests that there was a group of few girls with which John talked about many problems, not necessarily the same set of many problems. For (22), the (preferred) interpretation suggests that there was a (single) set of many problems about which talked with few girls (not necessarily a single group of girls).

5.5. Morphemic means

To make the repertoire complete, information structure may be also rendered by morphemic means. There belong the notorious example of the Japanese particles wa and ga discussed in linguistic literature since Kuno's (1972; 1973) pioneering analysis of the function of these particles in the information structure of Japanese (most recently, the thematic function of 'wa' was discussed e.g. by Fukuda 2003).

There are many other examples of languages where morphemics serves as (one of the means of expression) of information structure quoted in linguistic literature up to now, let me only give two of them mentioned by Novák (1974, p. 177) referring also to Dahl (1959). Information structure is expressed obligatorily and by using morphological means in Yukaghir, a Paleo-Asiatic language (Krejnovič 1958). There are three series of forms for each transitive verb there (distinguished from one another by the presence or absence of personal inflection, by morphological exponents, and by the presence or absence of certain prefixes) which are used whether the rheme-component coincides with the subject of the verb, or its object, or the verb

itself, respectively. In addition, a suffix is attached to the subject or object under conditions that pertain to the distribution of the rheme. In Tagalog, an Indonesian language, the theme of the sentence is distinguished by means of certain particles (articles) and word order; the syntactic roles of the given participants are indicated by an appropriate from of the verb (Bowen 1965).

6. Conclusions

In the present contribution we argue that (i) topic-focus articulation as a semantically relevant language phenomenon is an integral part of the description of the sentence at the underlying level of language description (Sect. 4.), (b) that as such, TFA belongs to 'langue', to the language system rather than to parole understood as the domain of communication and discourse, as sometimes claimed. From the point of view of the function - form relation as postulated by the Prague School scholars (shortly recapitulated in Sect. 1 of the present contribution) it is then not precise to characterize TFA (or FSP, for that matter) as an interplay of four factors, namely context, semantics, linearity, intonation (as continuously characterized by Firbas and his followers).

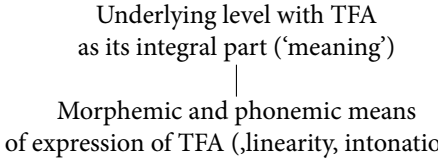


Figure 1.

While linearity and intonation (together with syntactic and morphemic means) belong to the side of 'means' or 'forms' in the hierarchy (see Fig. 1), the other two 'factors', namely the 'semantic' one (including the presentation scale: setting – presentation – phenomenon presented and the quality scale: setting – quality bearer – quality – specification(s)) and the contextual factor are of a different nature. They, of course, may help the linguist to determine what is the TFA of the sentence s/he examines (or whether the sentence is ambiguous); for the participants of the discourse the TFA of a sentence is relevant both for the suitability of the sentence for this or that context (from the point of view of the speaker) and for its semantico-pragmatic interpretation (from the viewpoint of the addressee (see Sgall 2003, p. 281).

Note: Parts 4.1 and 4.2 of the present contribution are modified and substantially enlarged versions of Sect.2 and 3.1, respectively, of Hajičová (2007).

Acknowledgements

The research reported on in the present contribution has been supported by the grant MSM 0021620838.

References

Bartels Christine (1997), Acoustic correlates of 'second occurrence' focus: Towards and experimental investigation. In Kamp and Partee (1997), 11–30.

Beaver David, Clark Brady Zack, Flemming Edward, Jaeger T. Florian and Wolters Maria (2007), When semantics meets phonetics: Acoustical studies of second-occurrence focus. *Language* 83, 245–276.

Bowen D.J., ed. (1965), *Beginning Tagalog.* Berkeley and Los Angeles.

Chomsky Noam (1957), *Syntactic Structures.* The Hague: Mouton.

Chomsky Noam (1965), *Aspects of the Theory of Syntax.* Cambridge, Mass.: The MIT Press.

Chomsky Noam (1971), Deep Structure, Surface Structure and Semantic Interpretation, in Steinberg and Jakobovits (1971),193–216.

Dahl Ő. (1969), *Topic and comment: a study in Russian and general transformational grammar.* Slavica Gothoburgensia 4, Gőteborg.

Daneš František (1970), Zur linguistischen Analyse der Textstruktur. *Folia linguistica* 4, 72–78.

Daneš František (1987), On Prague School Functionalism in Linguistics. In: René Dirven and Vilém Fried, eds.: *Functionalism in Linguistics*, Amsterdam/Philadephia: John Benjamins Publishing Company, 3–38.

Daneš František (2006), Pražská škola: názorová univerzália a specifika. Delivered at the Seminar on the occasion of the 80 years of Prague Linguistic Circle The Role of PLK in the Development and in the Perspectives of Czech Linguistics Prague, September 27, 2006; to be printed in a special issue of *Slovo a slovesnost*.

Dušková Libuše (1993), On some syntactic and FSP aspects of the cleft construction in English. In*: AUC Philologica* 1, Prague Studies in English 20, 71–87. Reprinted in Dušková (1999), p. 318–332.

Dušková Libuše (1999), *Studies in the English Language.* Part II. Section IV. Hypersyntax. Prague: Karolinum.

Firbas Jan (1964), On defining the theme in functional sentence analysis. *Travaux Linguistiques de Prague* 1, Prague, 267–280.

Firbas Jan (1992*), Functional Sentence Perspective in Written and Spoken Communication.* Cambridge: Cambridge University Press.

Firbas Jan (1999), Dogs must be carried on the escalator. *Brno Studies in English* 25, 7–18.

Fukuda Fazuo (2003), A consideration of the thematiser 'wa' in Japanese. In Hladký ed. (2003), 147–160.

Hajičová Eva (1984), Presupposition and Allegation Revisited. *Journal of Pragmatics* 8, 155–167; amplified as "On presupposition and allegation" in: Sgall, P. (ed.): *Contributions to*

Functional Syntax, Semantics and Language Comprehension. Amsterdam: Benjamins - Prague: Academia. 1984, 99–122.

Hajičová Eva (1976), *Negace a presupozice ve významové stavbě věty*. Prague: Academia.

Hajičová Eva (1993), *Issues of Sentence Structure and Discourse Patterns.* Prague: Charles University.

Hajičová Eva (2007), The Position of TFA (Information Structure) in a Dependency Based Description of Language, invited paper presented at the Meaning-Text Theory Conference.2007, Klagenfurt, in press in Proceedings.

Hajičová Eva, Partee Barbora and Petr Sgall (1998), *Topic/Focus Articulation, Tripartite Structures and Semantic Content.* Dordrecht: Reidel.

Hajičová Eva and Petr Sgall (2004), Degrees of Contrast and the Topic-Focus Articulation. In: A. Steube (ed.): *Information Structure – Theoretical and Empirical Aspects.* Berlin - New York: Walter de Gruyter, 1–13.

Halliday M.A.K. (1967), *Intonation and Grammar in British English.* The Hague: Mouton.

Halliday M.A.K. (1970), Language structure and language function. In: John Lyons, ed.: *New Horizons in Linguistics,* Penguin Books. 140–165.

Hladký Josef , ed. (2003*), Language and Function: To the Memory of Jan Firbas.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Hockett Charles F. (1961), Linguistic elements and their relations. *Language* 37, 29–53.

Jackendoff, Ray (1972), *Semantic Interpretation in Generative Grammar*, Cambridge, Mass,: MIT Press.

Jacobs J. (1984), Funktionale Satzperspektive und Illokutionssemantik. *Linguistische Berichte* 91:25–58.

Jakobson Roman (1963), Efforts toward a Means-Ends Model of Language in Interwar Continental Linguistics, In: *Trends in European and American Linguistics 1930–1960*, II, Utrecht 1963. Reprinted in Vachek 1964, 481–485.

Kamp Hans and Barbara Partee, eds. (1997), Context-*dependence in the analysis of linguistic meanin*g: Proceedings of the workshops in Prague and Bad Teinach. Stuttgart: Institut fuer Maschinelle Sprachverarbeitung, University of Stuttgart.

Karcevskij Serge (1929), Du dualism asymétrique du signe linguistique, *TCLP* 1, 88–93.

Krejnovič, E. A. (1958), *Jukagirskij jazyk.* Moscow – Leningrad.

Kuno Susumu (1972), Functional sentence perspective. *Linguistic Inquiry* 3:296–320.

Kuno Susumu (1973), *The structure of the Japanese language.* Cambridge, Mass.

Lakoff Geroge (1971a), On Generative Semantics. In: Steinberg and Jakobovits (1971), 232–296.

Lakoff George (1971b), Presupposition and Relative Well-Formedness. In: Steinberg and Jakobovits (1971), 329–340.

Leška Oldřich (1995), Prague School teachings of the classical period and beyond. In: *Prague Linguistic Circle Papers* 1, ed. by Eva Hajičová, Miroslav Červenka, Oldřich Leška and Petr Sgall, Amsterdam/Philadelphia: John Benjamins Publishing Company, 3–22.

Leška Oldřich (1999), Prague School Linguistics: Unity in diversity. In: *Prague Linguistic Circle Papers* 3, ed. by Eva Hajičová, Tomáš Hoskovec, Oldřich Leška, Petr Sgall and Zdena

Skoumalová, Amsterdam/Philadelphia: John Benjamins Publishing Company. 3–14.

Mathesius Vilém (1929a), Zur Satzperspektive im modernen Englisch. *Archiv für das Studium der neueren Sprachen und Literaturen* 155:202–210.

Mathesius Vilém (1929b), Funkční lingvistika. In: *Sborník přednášek pronesených na Prvém sjezdu československých profesorů filosofie, filologie a historie v Praze* 3.-7. dubna 1929. Printed also in Z klasického období pražské školy 1925–1945, ed. by J. Vachek, Prague, Academia 1972, 27–39. Also in Vilém Mathesius, Jazyk, kultura a slovesnost, Odeon, Prague,1982, 29–38. Translated into German as: Die funktionale Linguistik. In: Stilistik und Soziolinguistik, ed. by Eduard Beneš and Josef Vachek, Berlin, List Verlag 1971, 1–18.

Mathesius Vilém (1936a), On some problems of the systematic analysis of grammar. *TCLP* 6, 95–107.

Mathesius Vilém (1936b), Pokus o teorii struturální mluvnice [An attempt at a theory of structural grammar]. *Slovo a slovesnost* 2, 47–54.

Novák Pavel (1974), Remarks on devices of functional sentence perspective. In *Papers on Functional Sentence Perspective,* Prague: Academia, 175–178.

Novák Pavel and Petr Sgall (1968) On the Prague functional approach, *Travaux linguistique de Prague* 3, Academia, Prague, 291–297

Partee Barbara H. (1991), Topic, Focus and Quantification. In: S. Moore & A. Wyner, eds., *Proceedings from SALT I*, Ithaka, N.Y.: Cornell University, 257–280.

Partee Barbara H. (1996), Allegation and Local Accommodation. In: Partee & Sgall (eds.), 1996, 65–86.

Partee Barbara H. & Petr Sgall, eds. (1996), *Discourse and Meaning: Papers in Honor of Eva Hajičová*. Amsterdam/Philadelphia: Benjamins.

Peregrin Jaroslav (1994), Topic-Focus Articulation as Generatlized Quantification. In: Bosch & van der Sandt (eds.), 1994, 379–388.

Peregrin, Jaroslav (1996), Topic and Focus in a Formal Framework. In: Partee & Sgall (eds.), 1996, 235–254.

Quirk Randolph, Greenbaum Sydney, Leech Geoffrey and Jan Svartvik (1985), *A Comprehensive Grammar of the English Language.* London and New York: Longman.

Rooth,Mats (1985), *Association with Focus.* PhD Thesis, Univ. of Massachusetts, Amherst.

Saldanha Gaby (2007), The use of italics as stylistuic devices marking information focus in Englihs transl;ations. Presented atCorpus Linguistic 2007, Birmingham July 2007; abstract oprinted in *Abstracts Corpus Lingustics* 2007, p.55.

Selkirk Elisabeth (1984), *Phonology and Syntax: The Relation between Sound and Structure.* Cambridge, Mass.: MIT Press.

Selkirk Elisabeth (1995), Sentence Prosody: Intonation, Stress and Phrasing. In: J. A. Goldsmith (ed.): *Handbook of Phonological Theory.* London: Blackwell, 550–569.

Sgall Petr (1967a), *Generativní popis jazyka a česká deklinace* [Generative description of language and Czech declension], Prague: Academia.

Sgall Petr (1967b), Functional Sentence Perspective in a generative description of language. *Prague Studies in Mathematical Linguistics* 2, Prague, 203–225. Reprinted (shortened) in Sgall (2006), 275–301.

Sgall Petr (1979), Towards a Definition of Focus and Topic. *Prague Bulletin of Mathematical Linguistics* 31, 3–25; 32, 1980, 24–32; printed in *Prague Studies in Mathematical Linguistics* 78, 1981, 173–198.

Sgall Petr (1987), Prague Functionalism and Topic vs. Focus. In: René Dirven and Vilém Fried, eds.: *Functionalism in Linguistics.* Amsterdam/Philadephia: John Benjamins Publishing Company , 170–189.

Sgall Petr (2003), From functional sentence perspective to topic-focus articulation. In Hladký ed. (2003), 279–287.

Sgall Petr (2006) *Language in Its Multifarious Aspects.* Edited by Eva Hajičová and Jarmila Panevová. Prague: Karolinum.

Sgall Petr, Hajičová Eva and Eva Benešová (1973), *Topic, Focus and Generative Semantics*. Kronberg/Taunus: Scriptor.

Sgall Petr, Hajičová Eva and Eva Buráňová (1980), *Aktuální členění věty v češtině* [Topic-focus articulation of the sentence in Czech]. Prague: Academia.

Sgall Petr, Hajičová Eva and Jarmila Panevová (1986), *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*, ed. by J. L. Mey. Dordrecht:Reidel - Prague:Academia.

Skalička Vladimír (1935)., *Zur ungarischen Grammatik.* Prague: Facultas philosophica Universitatis Carolinae. Reprinted in Skalička (1979:59–125).

Skalička Vladimír (1948), Kodaňský strukturalismus a Pražská škola [Copenhagen structuralism and Prague School]. *Slovo a slovesnost* 10, 135–142.

Steinberg, D. D. and L. A. Jakobovits, eds. (1971), *Semantics – An Interdisciplinary Reader*. Cambridge University Press, Cambridge, UK.

Strawson Peter (1952), *Introduction to Logical Theory*, London: Methuen.

Strawson Peter (1964), Identifying Reference and Truth Values. *Theoria* 30, 96–118. reprinted in Steinberg and Jakobovits (eds.), 1971, 86–99.

Svoboda Aleš (2003), Jan Firbas – An outstanding personality of European linguistics. In: In Hladký ed. (2003), 1–7.

Thèses présentées au Premiér Congrès des philologues slaves. Travaux de Cercle Linguistique de Prague I, 5–29:

Trnka Bohumil (1964), On the linguistic sign and the multilevel organization of language, *TLP L'école de Prague d'aujourd'hui* 1, 33–40. reprinted in Bohumil Trnka, Selected Papers in Structural Linguistics, ed. by Vilém Fried, Mouton Publishers. Berlin – New York – Amsterdam. 86–93.

Vachek Josef (1960), *Dictionnaire de linguistique de l'Ecole de Prague*, Spectrum -Editeurs, Utrecht/Anvers. Translated as (and quoted from) Dictionary of the Prague School of Linguistics, edited by L. Dušková, translated from the original sources by Aleš Klér, Pavlína Šaldová, Markéta Malá, Jan Čermák and Libuše Dušková, Amsterdam-Philadelphia: John Benjamins Publishing Company, 2003.

Vachek Josef, ed. (1964*), A Prague School Reader in Linguistics.* Bloomington: Indiana University Press.

Vachek Josef (1966), *The Linguistic School of Prague*, Bloomington and London: Indiana University Press.

Veselá Kateřina, Peterek Nino and Eva Hajičová (2003), Topic-Focus Articulation in PDT: Prosodic Characteristics of Contrastive Topic. *The Prague Bulletin of Mathematical Linguistics* 79–80, 2003, p. 5–22.

Weil Henri (1844), *De l'ordre des mots dans les langues anciennes comparées aux langues modernes*, Paris. Translated as *The Order of Words in the Ancient Languages Compared with That of the Modern Languages*, Boston 1887, reedited, Amsterdam: Benjamins 1978.

Zwicky A. M. and J. M. Sadock (1975), Ambiguity tests and how to fail them. In *Syntax and Semantics 4,* ed. by J.P.Kimball, New York, 1–35.

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 88 DECEMBER 2007 73–90

How Can Typological Distances between Latin and Some Indo-European Language Taxa Improve Its Classification?

Yuri Tambovtsev

Abstract

The article deals with an Antique language—Latin. A new method of phonostatistics is proposed here. It is based on the structure of the frequency of occurrence of consonants in the speech sound chain. It is a good clue for defining the typological closeness of languages. It allows a linguist to find the typological distances between Latin and the other languages of different genetic groups of the Indo-European language family. This method can put any language in a language taxon, i.e. a sub-group, a group or a family. The minimum distance may be a good clue for placing Latin in this or that language taxon. The method of calculating Euclidean distances is used. It adds new information for classifying languages.

Key words: consonants, phonological, distance, typology, frequency of occurrence, speech sound chain, statistics, Euclidean distances, closeness, language taxon, taxa of languages, classification.

The aim of the article is to analyse an Antique language—Latin in order to put it in this or that language taxon. The new method of phonostatistics developed by the author is proposed here (**Tambovtsev, 1977; 2002-c; 2002-d; 2003-b; 2004-a; 2004-b**). It allows a linguist to find the typological distances between the languages under study (**Tambovtsev, 1994-b; 2001-d; 2002-a**). The obtained distances indicate to which language taxon a language belongs. In fact, the received language distances show similarity between the languages in question, the less the distance—the more similar the languages (**Tambovtsev, 2001-e; 2002-b; 2004-a**).

Now Latin is classified into the Italic group of the Indo-European language family (**Crystal, 1992: 199; JaDM, 1982: 19**). However, not so long ago Latin was placed into one group with the Romance languages (**Chikobava, 1953: 207–208**). May be, it is more logical, when the parent language is in the same group with its offsprings. It would be very strange if we put Old Slavonic in some separate group, but not in the Slavonic group. Our method shows the typological distances which may lit light on the closeness of Latin to the Romance languages since it is not

Please cite this article as: Yuri Tambovtsev, How Can Typological Distances between Latin and Some Indo-European Language Taxa Improve Its Classification?. The Prague Bulletin of Mathematical Linguistics No. 88, 2007, 73–90.

possible to find enough long and reliable texts in the true Italic languages: Faliscan, Oscan, Umbrian and Venetic which are dead by now. Therefore, Latin may have been placed in this language group for the lack of information. Though the number of texts in the Italic languages is limited and they are short, there are some linguists who claim that Latin belongs to the group of Italic languages. Rex E. Wallace goes even further than that. He claims without much evidence that Latin enters the Latino-Faliscan group of the Italic branch of the Indo-European language family (**Wallace, 2001: 412**). One must pay attention to the fact that he opens a new group and a new branch. More logically it is to call his new group the Latino-Faliscan subgroup. While his new branch is nothing else but the commonly accepted Italic group within the Indo-European language family. Though the information on the other Italic languages is scarce and unreliable, Rex E. Wallace insists that Oscan, Umbrian, South Picene, Vestinian, Marrucinian, Paelignian, Marsian, Volscian, Aequian and Hernican are more distant from Latin that Faliscan (**Wallace, 2001: 412**). However, it is quite possible that all the Italic languages mentioned above are just the sub-dialects and dialects of Latin. Though usually Latin is a term for the Classical Latin language, which was used only by the educated classes of Rome. Rex E. Wallace correctly points out that there were numerous different sub-dialects and dialects of Latin. He is also right to state that there were different variants of Latin for different social levels, e.g. Vulgar Latin as the speech of the common folk (**Wallace, 2001: 412**).

It is possible to agree that meanwhile it is advisable to place Latin into the Italic group of the Indo-European language family until more solid and reliable information is received. At the same time one cannot agree to the fact that this group is called a language family. A fair representative of the linguists who believe that there could be a family inside a family is David Crystal (**Crystal, 1992: 199**). Unfortunately, he is not the only one who makes a logical mistake like this. April McMahon and Robert McMahon also speak about the Germanic family, which is embraced into the Indo-European language family (**McMahon et al., 2005: 3-4**). However, if one takes into consideration all the reasoning of their book, one may realise that the abundance of data leads them to the conclusion that Indo-European family looks like a sort of a super-family, called here **a language unity**, i.e. the next level of classification. Usually, the languages as the objects at this higher level are not so similar as at the lower levels. If a classification is correct, i.e. natural, then the languages at the lower levels are more similar (**Tambovtsev, 2004-a: 201–210; Tambovtsev, 2004-b: 147–151).**

It is high time to reconsider all the established language families and other language taxa. If it is done so, then it may be discovered that Italic and Romance groups must be merged together into one group called Romano-Italic with two subgroups: Romance and Italic. There are some arguments, which allow us to do it. One of the arguments may be the distance between Latin and the Romance languages (**Tambovtsev, 2001-a**). If Latin is closer to the languages of the Romance group of languages, then it surely belongs to them, rather than to any other set of languages. Our results show the shortest mean distance of Latin to the languages of the Romance group, than to the other languages (c.f. Tab 1–13).

It is good to see that the logical mistake of classification described above is not made by other classifiers. Thus, Kenneth Katzner calls Italic a subgroup of the Indo-European language family (**Katzner, 1986:2**). However, strictly speaking he also makes a sort of a logical mistake,

since his subgroup does not enter a group, but a family. Thus, he omits one classification step. A logical classification of languages must incorporate subgroups into a group, groups into a family, families into a unity, unities into a phylum, phyla into a union, unions into a language community (**Tambovtsev, 2004: 145**).

It is high time to establish a universal and strict logical hierarchy of language taxa. All the linguists in the world should keep to one and the same order of language taxa (**Tambovtsev, 2003-a: 3**). The ordered series of the taxa of the world languages should include old and dead languages like Latin, Old Greek, Old Russian, Old Turkic, etc (**Tambovtsev, 2001-b; Tambovtsev, 2001-b; Tambovtsev, 2001-c**). While reconsidering and building new language taxa linguists should take into account the special rules. First of all it is the idea that they must separate all world languages into sets in such a way that the distances between languages in a language taxon must be less than the distances of these languages to the other world languages (**Tambovtsev, 2003-a**). The structure of a taxon is more dense (tight), that is compact, if the languages selected for it are more similar (**Tambovtsev, 2002-b**). In our studies it is usually the total of the distances between the ideal language in this or that set of language, which is expressed by the mean of a set (**Tambovtsev, 2001-e**). In a compact set the distances between the mean and the other values are minimal. First we developed this idea of compact and sparse sets of languages on the data of the frequency of occurrence of phonemes in the speech chain (**Tambovtsev, 1977**). Then, we went on applying the idea of the measure of compactness on the basis of the consonantal coefficient, which is the ratio of the frequency of occurrence (**Tambovtsev, 1986**)

We have nothing against placing Latin into the group of Italic languages of the Indo-European language family. Nevertheless, it is necessary to point out that in physics, chemistry, biology and other natural sciences old classifications are often reconsidered (**Kuhn, 1977; Rozova, 1986**). We must also point out that basing on the same known Indo-European isoglosses, Tomas V. Gamkrelidze and Vjacheslav Vs. Ivanov do not construct the group of Romance languages and the Italic group of the Indo-European language family. Instead, they define only one group of languages, i.e. the Italic group. Presumably, their Italic group embraces both Italic and Romance languages, since they do not provide a separate Romance group (**Gamkrelidze et al.,1984: 415**). It is fruitful that they also include not only the phonetical but the lexical and grammatical isoglosses, which allows them to obtain a more complete and reliable scheme. We have analysed this scheme in detail elsewhere and came to conclusion that their scheme is different from the usual traditional one in this aspect (**Tambovtsev, 1989, 134–137**).

Comparing the distances between Latin and Old Greek or Modern Greek one must bear in mind that Old Greek and Modern Greek are considered genetically isolated languages (**Crystal, 1992: 11; JaDM, 1982: 23**). There are some other languages, which have not been placed into any language family: Basque, Japanese, Korean, Ainu, Nivhi, Yukaghir and Ket (Yug). However, for the latter, a new language family—Yenissey has been invented. So, now Ket with all its dialects is the only memeber of the Yenissey family. Nevertheless, it is not a solution of the problem. If we follow this way, then we must also establish separate language families for Ainu, Basque, Japanese and the other isolated languages.

The new data, which we received for Latin may allow it to enter this or that group of lan-

guages. It is the first attempt to establish the phonostatistical measures for the typological closeness of Latin with the language groups, to which it may be supposed to enter. Usually, genetically close languages are also typologically close. However, the typologically close languages may be or may not be genetically close. Nevertheless, in the majority of cases typologically close languages are genetically close. We can find the phonostatistical closeness, which can give a good clue for the genetic relatedness. It was found for some Finno-Ugric, Turkic, Mongolic, Tungus-Manchurian and Paleo-Asiatic languages (**Tambovtsev, 2001-d; 2001-e; 2002-a; 2002-b; 2002-c; 2002-d; 2003-a; 2003-b; 2004**). Therefore, it is a good reason to believe this method should also work for Latin or any other language.

Why should one use quantitative methods in studying languages? A great philosopher and scientist Immanuil Kant (1724–1804) in his well-known works explaining the structure of the world stated that everything in this world possesses quantity and quality. Quantitative data characterise an object sometimes better, especially when the objects are very similar. Languages are similar in their qualitative characteristics. This is why, one should rely on the quantitative characteristics more. Actually, quantity may go over into quality when it is great enough (**FS, 1980: 144**). In this case, English is a fair example. Must it be considered a Germanic or a Romance language? Many words of its stock are of Romance origin as the result of the Norman Conquest in 1066. It is believed that quantitative characteristics work better in the cases when qualitative characteristics fail to distinguish two linguistic objects.

Long ago, in 1935, George Kingsley Zipf stated that it was necessary to introduce the so-called "Dynamic Philology" to achieve fruitful results in studying the structure and entity of Language (**Zipf. 1935:XII**). As George A. Miller correctly put in the introduction to Zipf's book, one who wishes to study a rose should count its petals, not just enjoy it. G. K. Zipf believed that it is necessary to study the massive statistical regularity of every linguistic unit or phenomenon (**Zipf, 1935:V–VI**).

Quantitative research needs the use of mathematical statistics. One can't help agreeing with Christopher Butler, who requires a quantitative treatment in any linguistic research because it is difficult otherwise to understand and evaluate how relevant are the linguistic results (**Butler, 1998: 255–264**).

Establishing genetic language families linguists compare every language with some other language or a group of languages. Jiri Kramsky is correct to remark that one can establish a typology of languages basing on the quantitative data received after comparing languages. The quantitative data gives a clearer vision of the differences and similarities between languages. The quantitative load of particular language phenomena is different in different languages. Kramsky is quite right to observe that in linguistics there is a very close relation between quality and quantity, even if the conditions of the transition of quantity into quality are not established so safely as they are in natural sciences. Nevertheless J. Kramsky assumes that in linguistics qualitative changes are asserted with the help of quantitative factors (**Kramsky, 1972: 15**).

Our method measures distances between languages on the phonological level. It gives a vivid picture of the typological similarity of the sound pictures of the languages under investigation. It allows us to find out the archetype of this or that language family. The mean values of the frequency of the consonantal groups

The use of quantitative data ensures that the languages are similar if the frequency of occurrence of certain linguistic units are similar. It takes into account both cases when the units are used very frequently or very seldomly. However, in classical linguistics, where the frequency is not taken into consideration, it is more often than not that the usual elements are compared with the rare elements. J. Kramsky is correct to point out that the language units which are in the centre of some language system should not be compared to those of the periphery (**Kramsky, 1972: 15**). The quantitative analysis shows us the units, which are in the centre of a language system and those which are at the periphery of it. Therefore, the typology of languages based on the quantitative data may add much to the established language families (**Tambovtsev, 2001-a; 2001-b; 2001-c; 2003**).

Latin, as any other human language, has a specific structure of the speech sound chain. It can be distinguished by its structure from any other language. Every language has a unique structure of distributions of speech sounds in its phonemic chain. The distribution of Latin vowels will not be considered till the second stage of the investigation. The frequency of occurrence will be considered if and only if the frequency of occurrence of different groups of consonants will not differentiate Latin from the other world languages. Let's point out that consonants bear the semantic load in the word, not vowels. Therefore, it is more possible to understand the meaning of the message by consonants, rather by vowels. Some linguists use consonants to consider statistical models in language taxonomy.

Let us consider the way one of statistical methods, namely, Chi-square is applied to place English and German in one group. On the basis of the frequency of fricative consonants [s] and [f] Alan Ross proved, and then April and Robert McMahon proved again that English and German are related, i.e. the use of these fricative consonants is not random (**McMahon et al., 2005: 59–61**). Actually, an outstanding American mathematician of Hungarian origin G. Polya used the same way of reasoning to establish the similarity of Hungarian to English, Swedish, Danish, Dutch, German, French, Spanish, Italian and Polish. He came to the conclusion on the sample of ten numerals that Hungarian is quite different from these languages (**Polya, 1975: 315–319**)

However, if we fail to recognise and distinguish two languages, then we resort to the structure of occurrence of vowels in the speech sound chain. While comparing languages, it is necessary to keep to the principle of commensurability. Having it in mind, it is not possible to compare languages on the basis of the frequency of occurrence of separate phonemes, because the sets of phonemes in languages are usually different. The articulartory features may serve as the basic features in phono-typological reasoning.

Before the computer measures the phonological distances, one has to choose the phonological features, which are necessary and sufficient. One has to select the system of the informative features. In pattern recognition such features are called basic (**Zagoruiko, 1972: 54–75**). Therefore, we have chosen all the features basic for the articulation of any speech sound. At the first stage we shall deal with consonants.

First of all, it is the classification of consonants according to the work of the active organ of speech or place of articulation (4 features). Secondly, it is the classification from the point of view of the manner of articulation or the type of the obstruction (3 features). Thirdly, it

is the classification according to the work of the vocal cords (1 feature). In this way, 8 basic features are obtained: 1) labial; 2) forelingual or front; 3) mediolingual or palatal; 4) guttural or back or velar; 5) sonorant; 6) occlusive non-sonorant; 7) fricative non-sonorant; and 8) voiced non-sonorant consonants. One should take the values of the frequency of occurrence of these 8 features in the speech chain of Latin and compare them to those of the other languages. On the basis of the "chi-square" test and Euclidean distance, we have developed our own method of measuring the phono-typological distances between languages (**Tambovtsev, 1994-a; 1994-b; 2004**). It takes into account the frequency of occurrence of the 8 consonantal groups mentioned above and builds up the overwhelming mosaic of the language sound picture.

It is very important to find some typological characteristics in order to endeavour to place it in some defined language family. Some linguists consider it impossible to put Latin in any of the known language families because it was unsufficiently studied before. Actually, it is considered here that it is possible to put Latin in a language family if its phonostatistical characteristics are studied better. Therefore, we undertook the study of the frequency of Latin phonemes on the vast sample of Latin texts. Fortunately, unlike the other Italic languages mentioned above, Latin has an abundance of reliable texts.

We fed into the computer the following Latin texts: 1) Latin proverbs and sayings from the book by V. N. Kuprejanova and N. M. Umnova and small texts by different Latin authors from the book by Ja.M. Borovskij and Bildyrev (Kuprejanova et al., 1975; Borovskij et al., 1949). 2) Aeneid by Vergilius.

After Aleksandr A. Derjugin, Larisa M. Lukjanova, Ja.M. Borovskij and A. V. Boldyrev, we define the following Latin phonemes:

Vowels: [i, u, e, o, a, i:, u:, e:, o:, a:, ae, oe, au, eu]

Consonants: [p, b, v, f, m, t, d, ts, s, z, n, l, r, j, k, g, h]

The classification of the Latin consonants by the work of the active of speech (i.e. place of articulation):

Labial: [p, b, v, f, m]

Forelingual (front): [t, d, ts, s, z, n, l, r]

Mediolingual (palatal): [j]

Guttural (velar or back): [k, g, h]

The classification by the manner of articulation (the character of the obstruction):

Sonorant: [m, n, l, r, j]

Occlusive non-sonorant: [p, b, t, d, ts, k, g]

Fricative non-sonorant: [v, f, s, z, h]

The classification by the work of the vocal cords:

Voiced non-sonorant consonants: [b, v, d, z, g]

After computing the Latin text by V. N. Kuprejanova, N. M. Umnova, Ja.M. Borovskij and A.V. Boldyrev, we received the following frequencies of the phonemic occurrence in the sound chain:

	Frequency	% to all ph.	% to cons.
Labial:	4561	13.82	24.12
Forelingual (front)	12248	37.12	64.77
Palatal (mediolingual)	140	0.42	0.73
Guttural (back)	1964	5.95	10.38
Sonorant	7463	22.62	39.47
Occlusive non-sonorant	7297	22.11	38.58
Fricative non-sonorant	4153	12.58	21.95
Voiced non-sonorant	2702	8.19	14.29

The total of consonants: 18913 phonemes — 57.31 %

The total of vowels: 14087 — 42.69 %

The value of the consonantal coefficient (i.e. the ration of consonants to vowels): 1.34

Sample volume of the Latin proverbs: 33000 phonemes.

Zip's data has the following frequency of the phonemic occurrence in the sound chain (Zipf et al., 1939):

	Frequency	% to all ph.	% to cons.
Labial:	560	11.20	20.86
Forelingual (front)	1705	34.10	63.50
Palatal (mediolingual)	25	0.50	0.93
Guttural (back)	395	7.90	14.71
Sonorant	1076	21.52	40.07
Occlusive non-sonorant	1149	22.98	42.79
Fricative non-sonorant	460	9.20	17.13
Voiced non-sonorant	260	5.20	9.68

The total of consonants: 2685 phonemes — 53.70 %

The total of vowels: 2315 — 46.30 %

The value of the consonantal coefficient (i.e. the ratio of consonants to vowels): 1.16

Sample volume of the Zipf's Latin text: 5000 phonemes.

The author has also computed the epic poem "Aeneidos" by Vergilius. It is long and consists of 12 chapters describing the legends dedicated to Rome. Publius Vergilius Maro received a good education in philosophy, poetry and rhetoric. He wrote his poem for some 11 years. It is considered to be a good sample of classical Latin. "Aeneid" has the following frequency of the phonemic occurrence in the sound chain:

	Frequency	% to all ph.	% to cons.
Labial:	43514	12.15	21.19
Forelingual (front)	135892	37.95	66.20
Palatal (mediolingual)	1504	0.41	0.72
Guttural (back)	24411	6.82	11.89
Sonorant	80515	22.48	39.21
Occlusive non-sonorant	82351	23.00	40.12
Fricative non-sonorant	42455	11.85	20.67
Voiced non-sonorant	25218	7.04	12.28

The total of consonants: 205321 phonemes — 57.33 %
The total of vowels: 152800 — 42.67 %
The value of the consonantal coefficient (i.e. the ratio of consonants to vowels — 1.34
Sample volume of the Latin text of Aeneid: 358121 phonemes.

The united data computed by the author consists of Latin proverbs and "Aeneid". It has the following frequency of the phonemic occurrence in the sound chain:

	Frequency	% to all ph.	% to cons.
Labial:	48075	12.29	20.97
Forelingual (front)	148140	37.88	64.63
Palatal (mediolingual)	1644	0.42	0.73
Guttural (back)	26375	6.74	11.76
Sonorant	87978	22.49	39.23
Occlusive non-sonorant	89648	22.92	39.98
Fricative non-sonorant	46608	11.92	20.79
Voiced non-sonorant	27920	7.14	12.45

The total of consonants: 224234 phonemes — 57.33 %
The total of vowels: 166887 — 42.67 %
The value of the consonantal coefficient (i.e. the ratio of consonants to vowels — 1.34
Sample volume of the Latin text of Aeneid: 358121 phonemes.

It is recommended to use in linguistics some exact measure to place the languages more objectively. In pattern recognition such exact measures of distances between two objects are used. Nikolai G. Zagoruiko recommends to use the Euclidean distances when the value of the features are equal (**Zagoruiko, 1999: 198–199**). We consider all our features to be equal since we cannot claim that the frequency of occurrence of labials is more important than the frequency of occurrence of sonorants, or the frequency of occurrence of palatals is more important than the frequency of occurrence of the fricatives and so on.

We measure here the distances by the well-known formula of measuring the distance between points in the Euclidean space:

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 + \text{etc.}}$$

where
D - distance
x_1 - the frequency of occurrence of labials in Latin
x_2 - the frequency of occurrence of labials in the second language
y_1 - the frequency of occurrence of front consonants in Latin
y_2 - the frequency of occurrence of front consonants in the second language
z_1 - the frequency of occurrence of palatals in Latin
z_2 - the frequency of palatals in the second language, etc.

The details of calculating Euclidean distances may be found elsewhere (**Tambovtsev, 2003-c: 122**). This method is good because it can use any number of features in any number of languages. Therefore, a linguist can take as many linguistic features as he wants. The number

of languages is not limited either. So, this method calculated the distance between Basque and Latin (10.54). Though the least distances were between Basque and Kazah (5.310) or Tofalar (5.96) and the other Turkic languages (**Tambovtsev, 2003-c: 125**).

It is necessary to introduce some system of references when dealing with the distances between Latin and the other languages. Such point may be the distance between two texts in some language. We calculated the distances between two texts in the Markiz language, one of the Austronesian languages. It is 0.505. Now let us take any other language as a point for the system of references. It can be any language, which is far away from Latin and the contacts with which is not probable. Such a language may be Ainu. The native speakers of Ainu live in Japan. So the influence of Latin on Ainu is not possible. For calculating the distances between Ainu and the other languages we used the same method. The language closest to Ainu is one of the Austronesian languages—Tagalog with the distance of 9.310. The closest language to Latin by this method is Moldavian (4.275), then comes Italian (5.242) and then Romanian (6.913). We can see that Latin is much closer to Moldavian, than Ainu to its closest language. In fact, it is by two times closer. We can see the other distances between Latin and Romance languages in Tab. 1.

The least distance between Latin and Moldavian means that they are the closest languages among the chosen Romance and other languages (c.f. Tab. 1–13). It is not surprising since Moldavian and Romanian are spoken by the descendants of Roman soldiers and settlers, who occupied the Roman province of Dacia (Carlton, 2001: 598). In my mind, Italian, Moldavian and Romanian preserved the articulation base of Latin and thus the frequency of occurrence of sounds in Latin and in these languages is more similar, than in the others. Actually, the smallest distance between Latin and Moldavian may speak for many more remnants in Moldavian rather than in Italian. It is always so that at the periphery there are more obsolete features than in the centre. These distances may also point out that the articulation base of these three languages is rather similar.

As a matter of fact, articulation base is the main factor in ruling the frequency of occurrence of speech sounds in any language. We can see it on the examples of other languages, e.g. Ainu. Let us remember the words of N. A. Nevskij that Ainu is close to Paleo-Asiatic languages (**Tambovtsev, 2001-b**). Indeed, one of the Paleo-Asiatic languages, i.e. the Chookchi language with the distance 10.954 is rather close. The next closest language is also a Paleo-Asiatic language—Koriak with the distance 12.781. Korean is a bit closer — 12.636. Japanese is more far away — 15.269. As we can see from the tables below the other languages are also rather far away. So, the closest Tungus-Manchurian language is Ul'ch with the distance 13.464.

However, the most close to Ainu turned the American Indian languages of the North and South America. So, Quechua has the distance of 5.451 and Inga 7.388. They both belong to the Quechua family of American Indian languages. Quechua and Inga Indians live in South America.

Let us take some other languages as reference points. Japanese is a good choice since it is an isolated language. Having compared Japanese to some languages, we received the following phono-typological distances: Japanese–Ujgur (6.77); Japanese–Nanaj (8.12); Japanese–Jakut (8.26); Japanese–See Dajak (8.86); Japanese–Kazah (9.02); Japanese–Turkish (9.05); Japanese–Ket

(9.52); Japanese–Baraba Tatar (9.76); Japanese–Uzbek (10.63); Japanese–Hausa (10.98); Japanese–Georgean (11.05); Japanese–Kazan Tatar (11.07) and so on. One can see, that Ujgur, Jakut, Kazah, Turkish, Baraba Tatar, Uzbek and Kazan Tatar are Turkic languages. Nanaj is a Tungus-Manchurian language. Therefore, one can notice that Japanese is closer to the so-called Altaic languages which include Turkic, Mongolian and Tungus-Manchurian languages. Many world languages were compared to Japanese. We can't show all the distances here for the lack of space. However, the maximum distances were found for Japanese–German (22,24); Japanese–English (19.83); Japanese–Rumanian (15,08) and Japanese–Swedish (17.03). As a conclusion, we can also state that speech sound picture of Japanese is rather far away from the languages, which are geographically close: Chinese, Nivh, Itelmen or Indonesian. It was a surprise to us. Our data state that the speech sound pattern of Japanese resembles that of Ujgur–one of the Turkic languages spoken in the Middle Asia. The Ujgur people are often linked to the Old Turkic tribes, who used to live in the stepps of Southern Russia before the Tatar-Mongols captured them in the 9th century A.D. We must point out that it is not a coincidence since the other native Altaic people have a very similar data of closeness to Japanese. Turkic and Tungus-Manchurian tribes may have had a sort of common origin with Japanese. It may verify the Altaic hypothesis of Japanese origin. It is especially vivid, when the Austro-Oceanic and other languages do not show such a great closeness.

Considering the mean distance between Latin and the other languages and sets of languages, one may notice a clear preference. The mean distance between Latin and the Romance languages is the least 6.706 (c.f. Tab.1). The Baltic languages (Latvian and Lithuanian) are also rather close (8.504) to Latin (c.f. Tab.5). Latin is closer in general to the Eastern Slavonic languages (Russian, Old Russian, Ukrainian and Belorussian), than to the other two Slavonic subgroups. The mean distance is less (9.259) than that of Latin to Southern Slavonic (9.810) or Western Slavonic (13.008). So, it speaks again for similarity between Eastern and Southern Slavonic subgroups (c.f. Tab 1–4).

The Iranian group is closer (10.673), than Germanic (11.160) or Indic (12.400) groups.

It is possible to see that Old Greek (8.482) and Modern Greek (8.653) are not so close to Latin. However, Armenian is a bit further (8.838). Albanian is not close enough either (9.325).

Nevertheless, the Indo-European languages are closer to Latin than the Samoyedic family (15.400) or the Ob-Ugrian subgroup of the Ugric group of the Finno-Ugric family(16.333). The Northern dialect of Mansi (19.017) or the Konda dialect of Mansi (18.261) may be the champions (c.f. Tab. 14).

In conclusion, it is possible to state a great typological closeness between Latin and some languages of the Romance group of the Indo-European family. We are far from stating that genetically Latin is closer to the languages of the Romance group than to the languages of the Italic group. However, from the point of view of typology Latin is very similar to the Romance languages. Having this typological clue, linguists may have a closer look at Latin from the genetic point of view. May be, it is advisable to reconsider both Italic and Romance groups and unite them into one group Romano-Italic with two sub-groups: Romance and Italic.

References

Borovskij et al., 1949 - Borovskij Ja. M. and Boldyrev A. V. Latinskij jazyk [Latin]. - Moskva: ILIJA, 1949.

BSE - Bol'shaja Sovetskaja Entsyklopedija [Great Soviet Encyclopaedia]. 30 Volumes. Moskva: Sovetskaja Entsiklopedija Publishing House, 1978.

Butler, 1998 - Butler, Christopher. Statistics. - In: Projects in Linguistics. A Practical Guide to Researching Language. Alison Wray, Kate Trott and Aileen Bloomer with Shirley Reay and Chris Butler. - London-New York: Arnold-Hodder, 1998.–303 p.

Carlton, 2001 - Carlton, Charles M. Romanian. - In: Facts About the World's Languages: An Encyclopaedia of the World's Major Languages, Past and Present (Edited by Jane Garry and Carl Rubino). - New York and Dublin: The H. W. Wilson Company- A New England Publishing Associates Book, 2001, p. 598–601.

Chikobava, 1953 - Chikobava A, S. Vvedenie v jazykoznanie [Introduction into Linguistics]. - Moskva: Prosveshchenie, 1953.–243 pages.

Crystal, 1992 - Crystal, David. An Ecyclopedic Dictionary of Language and Languages. - Oxford: Blackwell, 1992–428 p.

FS, 1980 - Filosofskij slovar' [Phylosophy Dictionary] (Editor I. T. Frolova). - Moskva: Politizdat, 1980–445 pages.

Derjugin et al., 1986 - Derjugin, Aleksandr A., Lukjanova Larisa M. Latinskij jzyk [Latin]. - Moskva: Vysshaja shola, 1986–295 p.

Gamkrelidze et al., 1984 - Tomas V. Gamkrelidze and Vjach. Vs. Ivanov. Indo-European and the Indo-Europeans. Volume 1 and 2. [in Russian] - Tbilisi: The Tbilisi University Press, 1984–1392 p.

JaDM, 1982 - Jazyki i dialekty mira. [Languages and Dialects of the World]. - Moskva: Nauka, 1982.

Katzner, 1986 - Katzner, Kenneth. The Languages of the World. - London: Routledge and Kegan Paul Publishers, 1986–376 pages.

Kuhn, 1977 - Kuhn T. S. The Structure of Scientific Revolutions [in Russian]. - Moskva: Progress Publishing House, 1977.–300 pages.

McMahon et al, 2005 - McMahon, April and McMahon, Robert. Language Classification by Numbers. - Oxford: Oxford University Press, 2005.–263 pages.

Polya, 1975 - Polya, George. Mathematics and Plausible Reasoning. - Moskva: Nauka, 1975. - 463 pages. [Russian translation of his book published in 1954 by Princeton University Press].

Rozova, 1986 - Rozova S. S. Klassifikatsionnaja problema v sovremennoj nauke [Classification Problem in Modern Science]. - Novosibirsk: Nauka, 1986.–223 pages.

Tambovtsev, 1977 - Tambovtsev, Yuri. Nekotorye harakteristiki raspredelenija fonem mansijskogo jazyka [Some Characteristics of the Distribution of the Mansi Language]. - In: Soviet Finno-Ugric Studies, Volume XIII, # 3, 1977 (Tallin), p. 195–198.

Tambovtsev, 1986 - Tambovtsev, Yuri. Konsonantnyj koeffitsient v jazykah raznyh semej [Consonant Coefficient in the Languages of Different Families]. - Novosibirsk: Novosibirsk

University, 1986. - 17 pages.

Tambovtsev, 1989 - Tambovtsev, Yuri. Review on the book by Tomas V. Gamkrelidze and Vjach. Vs. Ivanov. Indo-European and the Indo-Europeans. Tbilisi: The Tbilisi University Press, 1984. Volume 1 and 2. - In: Gengo Kenkyu, # 96, 1989 (Tokyo), p. 119–143.

Tambovtsev, 1994-a - Tambovtsev, Yuri. Dinamika funktsionorovanija fonem v zvukovyh tsepochkah jazykov razlichnogo stroja. - Novosibirsk : NGU, 1994-a. - 133 pages.

Tambovtsev, 1994-b - Tambovtsev, Yuri. Tipologija uporjadochennosti zvukovyh tsepej v jazyke. - Novosibirsk: NGU, 1994-b. - 199 pages.

Tambovtsev, 2001-a - Tambovtsev, Yuri. Kompendium osnovnyh statisticheskih harakteristik funktsionirovanija soglasnyh fonem v zvukovoj tsepochke anglijskogo, nemetskogo, frantsuskogo i drugih indoevropejskih jazykov. [Compendium of the basic statistical characteristics of functioning of consonants in the speech chain of English, German, French and other Indo-European languages]. - Novosibirsk: Novosibirskij klassicheskij institut, 2001-a. - 129 pages.

Tambovtsev, 2001-b - Tambovtsev, Yuri. Funktsionorovanie soglasnyh fonem v zvukovoj tsepochke uralo-altajskih jazykov. [Functioning of consonants in the sound chain of Ural-Altaic languages]. Novosibirsk: Novosibirskij klassicheskij institut, 2001-b. - 132 pages.

Tambovtsev, 2001-c - Tambovtsev, Yuri, Nekotorye teoreticheskie polozhenija tipologii upor'adochennosti fonem v zvukovoj tsepochke jazyka i kompendium statisticheskih harakteristik osnovnyh grupp soglasnyh fonem. [Some theoretical fundamentals of the typology of oderliness of phonemes in the sound chain of language and the compendium of statistical characteristics of the basic groups of consonants], - Novosibirsk: Novosibirskij klassicheskij institut, 2001-c. - 130 pages.

Tambovtsev, 2001-d - Tambovtsev, Yuri. The phonological distances between Mongolian and Turkic languages based on typological consonantal features. - In: Mongolian studies. Journal of the Mongolia Society (USA), Vol.24, 2001-d, p.41–84.

Tambovtsev, 2001-e - Tambovtsev, Yuri. Perednejazychnye soglasnye kak pokazatel' odnoj iz lingvisticheskih universalij jazykov mira [Front consonants as an indicator for one of the universals of world languages]. - In: Sibirskij Lingvisticheskij Seminar, # 2 (2), Novosibirsk, 2001-e, p.18–27.

Tambovtsev, 2002-a - Tambovtsev, Yuri. Comparative typological study of language distances based on the consonants in sound chains of various languages. - In: The 5th National Colloquium for Computational Linguistics in the UK. Proceedings of the Conference. (Edited by John Elliot). 8-9 January, 2002. University of Leeds, UK. - Leeds: University of Leeds, 2002-a, p.77–80.

Tambovtsev, 2002-b - Tambovtsev, Yuri. Is Kumandin a Turkic language? - In: Dilbilim Arashtirmalari (Istanbul), 2002-b, p. 63–104.

Tambovtsev, 2002-c - Tambovtsev, Yuri. Korean and Japanese as Members of the Altaic Language Family. - In: Abstracts. Permanent International Altaistic Conference 45th Meeting, Budapest June 23-28, 2002. Budapest: Research Group for Altaic Studies. Hungarian Academy of Sciences, 2002-c, p. 13–14.

Tambovtsev, 2002-d - Tambovtsev, Yuri. Structure of the frequency of occurrence of con-

sonants in the speech sound chain as an indicator of the phono-typological closeness of languages. - In: ALL - ACH 2002. New Directions in Humanities Computing. The 14th Joint International Conference, University of Tuebingen, 24–28 July, 2002. Conference Abstracts. - Tuebingen: Universitaet Tuebingen, 2002-d, p. 138–139.

Tambovtsev, 2003-a - Tambovtsev, Yuri. Tipologija Funktsionorovanija fonem v zvukovoj tsepochke indoevropejskih, paleoaziatskih, uralo-altajskih i drugih jazykov mira: kompaktnost' podgrupp, grupp, semej i drugih jazykovyh taksonov. [Typology of functioning of phonemes in sound chain of Indo-European, Paleo-Asiatic, Ural-Altaic and other world languages: compactness of subgroups, groups, families and other language taxins]. - Novosibirsk: Sibirskij nezavisimyj institut, 2003-a. - 143 pages. [in Russian].

Tambovtsev, 2003-b - Tambovtsev, Yuri. The phonological similarity between Turkic languages based on some phonological features of consonants. - In: Linguistic and Oriental Studies from Poznan', Vol. 5, Poznan', (Poland), 2003-b, p. 85–118.

Tambovtsev, 2003-c - Tambovtsev, Yuri. Phonological similarity between Basque and other world languages based on the frequency of occurrence of certain typological consonantal features. - In: The Mathematical Bulletin of Mathematical Linguistics # 79–80, 2003 (Praha), p. 121–126.

Tambovtsev, 2004-a - Tambovtsev, Yuri. Uralic Language Taxon: Natural or Artificial? (Typological Compactness of Uralic Languages and other Language Taxons: Branches, Subgroups, Groups, Families and Superfamilies). - In: Fenno-Ugristica, # 26, 2004 (Tartu), p. 200–246.

Tambovtsev, 2004-b - Tambovtsev, Yuri. Phonostatistical Distance Totals as an Indicator of the Compactness of Language Taxons: a Typologo-Metrical Approach. - In: Lingua Posnaniensis. Vol. XLVI, Poznan' (Poland), p. 145–172.

Wallace, 2001 - Wallace, Rex E. Latin. - In: Facts About the World's Languages: An Encyclopaedia of the World's Major Languages, Past and Present (Edited by Jane Garry and Carl Rubino). - New York and Dublin: The H. W. Wilson Company- A New England Publishing Associates Book, 2001, p.201–416.

Zagoruiko, 1972 - Zagoruiko, Nikolai G. Metody raspoznavanija i ih primenenie. [The Methods of Pattern Recognition and Their Application] (in Russian). - Moskva: Sovetskoe Radio, 1972.–206 pages.

Zagoruiko, 1999 - Prikladnye metody analiza dannyh i znanij. [Applied Methods of Data and Knowledge Analysis](in Russian). - Novosibirsk: Institute of Mathematics SORAN, 1999. - 269 pages.

Zipf, 1935 - Zipf, George Kingsley. The Psycho-Biology of Language. An Introduction to Psycho-Biology of Language. - Cambridge: Massachusetts Institute of Technology Prtess, 1935. - 336 pages.

EUCLIDEAN DISTANCES between Latin and other world languages, united in different genetic families and other language taxa

Tab. 1

Phonostatistical EUCLIDEAN DISTANCES between Latin and Romance language group of the Indo-European language family. The mean of the distances — 6.706.

Language	Distance
	Latin
0. Latin	0.000
1. Moldavian	4.275
2. Italian	5.242
3. Rumanian	6.913
4. Spanish	7.353
5. Portuguese	9.747

Tab. 2

Phonostatistical EUCLIDEAN DISTANCES between Latin and the Eastern Subgroup of the Slavonic language group of the Indo-European language family.

The mean of the distances – 9.259.

Language	Distance
	Latin
0. Latin	0.000
1. Russian	4.275
2. Old Russian	9.048
3. Belorussian	10.124
4. Ukrainian	10.169

Tab. 3

Phonostatistical EUCLIDEAN DISTANCES between Latin and the Southern Subgroup of the Slavonic language group of the Indo-European language family.

The mean of the distances – 9.810.

Language	Distance
	Latin
0. Latin	0.000
1. Macedonian	7.502
2. Slovenian	8.582
3. Serbian	9.579
4. Bulgarian	13.577

Tab. 4

Phonostatistical EUCLIDEAN DISTANCES between Latin and the Western Subgroup of the Slavonic language group of the Indo-European language family.

The mean of the distances – 13.008.

Language	Distance
	Latin
0. Latin	0.000
1. Slovak	11.653
2. Czech	11.743
3. Luzhits-Sorbian	11.789
4. Polish	16.848

Tab.5

Phonostatistical EUCLIDEAN DISTANCES between Latin and Baltic language group of the Indo-European language family.

The mean of the distances – 8.504.

Language	Distance
	Latin
0. Latin	0.000
1. Latvian	7.344
2. Lithuanian	9.664

Tab. 6

Phonostatistical EUCLIDEAN DISTANCES between Latin and Indic language group of the Indo-European language family.

The mean of the distances – 9.231.

Language	Distance
	Latin
0. Latin	0.000
1. Gypsy	6.939
2. Sanskrit	8.074
3. Marathi	8.097
4. Bengali	10.268
5. Hindi	12.779

Tab. 7

Phonostatistical EUCLIDEAN DISTANCES between Latin and Iranian language group of the Indo-European language family.

The mean of the distances – 10.673.

Language	Distance
	Latin
0. Latin	0.000
1. Persian (Iranian)	7.877
2. Osetian	9.804
3. Tadjik	14.338

Tab. 8

Phonostatistical EUCLIDEAN DISTANCES between Latin and Celtic language group of the Indo-European language family.

Language	Distance
	Latin
0. Latin	0.000
1. Irish	13.057

Tab. 9

Phonostatistical EUCLIDEAN DISTANCES between Latin and Germanic language group of the Indo-European language family.

The mean of the distances – 11.160.

Language	Distance
	Latin
0. Latin	0.000
1. Dutch	8.075
2. Norwegian	8.793
3. Old English	10.002
4. English	11.763
5. Gothic	12.258
6. German	16.067

Tab. 10

Phonostatistical EUCLIDEAN DISTANCES between Latin and Isolated languages of the Indo-European language family.

Language	Distance
	Latin
0. Latin	0.000
1. Old Greek	8.482
2. Modern Greek	8.653
3. Armenian	8.838
4. Albanian	9.325

Tab. 11

Phonostatistical EUCLIDEAN DISTANCES between Latin and Esperanto—an artificial language.

Language	Distance
	Latin
0. Latin	0.000
1. Esperanto	7.330

Tab. 12

Phonostatistical EUCLIDEAN DISTANCES between Latin and the Ob-Ugric Subgroup of the Ugric language group of the Finno-Ugric language family.

The mean of the distances – 16.333.

Language	Distance
	Latin
0. Latin	0.000
1. Eastern Hanty	11.823
2. Kazym Hanty	16.231
3. Konda Mansi	18.261
4. Northern Mansi	19.017

Tab. 13

Phonostatistical EUCLIDEAN DISTANCES between Latin and the Samoedic language family.

The mean of the distances – 15.400.

Language	Distance
	Latin
0. Latin	0.000
1. Nenets	14.375
2. Nganasan	15.572
3. Selkup	16.252

Tab. 14

The Ordered Series of the Mean Phonostatistical EUCLIDEAN DISTANCES between Latin and Some Subgroups and Groups of the Indo-European family. The mean of the distances inside every language taxon.

Language	Mean Distance
	Latin
0. Latin	0.000
1. Romance	6.706
2. Baltic	8.504
3. Eastern Slavonic	9.259
4. Southern Slavonic	9.810
5. Iranian	10.673
6. Germanic	11.160
7. Indic	12.400
8. Western Slavonic	13.008

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 88 DECEMBER 2007 91–92

NOTE

Our Lucky Moments with Frederick Jelinek

Barbora Vidová Hladká

This contribution is going to be a congratulation to Frederick Jelinek's birthday jubilee. Before I reach the very congratulation I would like to remind a lucky moment that had a strong influence on the life of a certain Institute of Charles University in Prague after 1989. And it is by no chance that the honored person witnessed the above mentioned moment and its consequences. From my personal point of view, I have become one of the "victims" of this lucky moment so I really appreciate the opportunity to wish well to Fred via the Prague Bulletin circulating the institutions over the world.

The crucial events in November 1989 in Czech Republic brought freedom to a lot of people. Freedom to scientists in the group of computational linguistics at the Faculty of Mathematics and Physics, Charles University changed (among other things) their subdepartment into an independent department of the faculty in 1990, namely the Institute of Formal and Applied Linguistics (ÚFAL) headed by Eva Hajičová. Freedom to Fred Jelinek made it possible for him (among other things) to give a two term course on spoken and written language analysis at the Czech Technical University in Prague in 1991-1992. At that time, Fred was a senior manager of the IBM T.J. Watson Research Center, Yorktown Heights, NY and he was heading a group carrying out research on continuous speech recognition, machine translation and text parsing and understanding. While being in Prague, he was looking for a Czech scientist to offer him/her a position in his IBM group. The first one who he asked refused. The second one did not. But what is more important, the second one was the present director of ÚFAL Jan Hajič. As far as I know the search for a candidate was running via the question Do you know someone who would be interested to spend some time at IBM? So Jan was among those addressed and he did accept Fred's offer. The experience with the statistically based machine translation that Jan acquired at IBM became crucial for the next progress of ÚFAL.

In 1993, Fred moved to Baltimore and became the director of the Center for Language and Speech Processing (CLSP) at Johns Hopkins University. His very nice idea of the summer workshops came into life for the first time two years later in 1995 and lasts till now when CLSP invites proposals already for the 14th workshop.

Since 1993, when Jan returned back to Prague, much work had been done in a field of corpus linguistics and corpus-based approaches at ÚFAL - the Prague Dependency Treebank v. 0.5 was released in 1998 and the statistically-oriented experiments on tagging and parsing were performed even before then. Thus the topic of parsing happened to be one of three projects solved during the 4th Summer workshop in 1998 and Fred had a lot to say! Going through the complete workshop participant listings during the whole time of existence of this wonderful event, I can summarize that the members of ÚFAL participated in five out of thirteen workshop series - please, recall a lucky moment described at the very beginning! The summer workshops do not present the only possibilities open by Fred to ÚFAL's people - the graduate students are invited for the stays in CLSP. The benefits and motivation gained over these stays are undeniable and exceptional.

Needless to stress that I have touched upon only a few of key moments for ÚFAL that Fred initiated. A more comprehensive birthday congratulation to Fred was presented by Eva Hajičová and Jan Hajič at the Text, Speech and Dialogue Conference 2007 in Pilsen. If you had no chance to hear it, do not despair. You can read it, see (Hajič and Hajičová, 2007).

To conclude I am happy to know that the word "speech" knows to elude smile while taking photo as well as the word "cheese". *Happy birthday and good luck, professor Jelinek! Thanks for being helpful, original and exceptional . . .*

Bibliography

Hajič, Jan and Eva Hajičová. 2007. Some of Our Best Friends Are Statisticians. In *Proceedings of the 10th International Conference on Text, Speech and Dialogue*, pages 2–10. Springer-Verlag Berlin Heidelberg.

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 88 DECEMBER 2007 93–94

NOTE

ACL 2007—the 45th Annual Meeting of the Association for Computational Linguistics, Prague, June 23–30, 2007

Eva Hajičová

The 45th ACL Annual Meeting has been held in Prague on June 23–30, 2007 organized by the Institute of Formal and Applied Linguistics at Charles University in Prague. It was the largest ACL meeting ever held by the ACL during the time of its existence. There were more than 1000 participants from abroad, coming from 48 countries, and about 70 researchers and students from the Czech Republic, mostly from Charles University in Prague. The meeting was held under the auspices of the former Rector of Charles University prof. ing. Ivan Wilhelm, the present-day Rector of the University prof. RNDr. Václav Hampl (who welcomed the participants on behalf of the University at the opening session) and the Mayor of the City of Prague MUDr. Pavel Bém. The General Chair of the Conference was John Carroll; the programme chairs were Annie Zeanen and Antal van den Bosch, the tutorial chair Joakim Nivre and the workshop chair Simone Teufel. The Local Organizing Committee was headed by Eva Hajičová with Jan Hajič as Local Coordinator and Anna Kotěšovcová as Local Arrangements Chair.

The three-day main conference consisted of four parallel sessions and one student session. There were 588 submissions for the main conference out of which 131 have been accepted (acceptance rate 22,30 %). The invited speakers were Tom Mitchell (on the relations between language, meaning, and brain), and Barney Pell (from Powerset) on intelligent text retrieval.

15 workshops were organized before and after the conference and there were also two adjoined conferences, the International Workshop on Parsing Technologies (IWPT) and the joint conference on Empirical Methods in Natural Language Processing and on Computational Natural Language Learning (EMNLP-CoNLL). The EMNLP-CoNLL joint conference this year was also exceptionally large, there were more than 340 participants and 398 submissions from which 66 were accepted as otral presentations and 43 as posters, making the acceptance rate 27 %. This conference also included some short reports on the results of a shared task concerning dependency based analysis applied to annotated corpora of several languages.

There were five tutorials before the main conference with a range of topics: special statistical methods for NLP, data mining from Internet, dialogue systems, methods of logical inferencing from texts and methods of evaluation and advancing the quality of corpus annotation.

The Best Paper Award went to Yuk Wah Wong a Raymond J. Money who delivered the paper "Learning Synchronous Grammars for Semantic Parsing with Lambda Calculus."

The recipient of the 2007 Life Achievement Award was a very influential theoretical and computational linguist, Lauri Karttunen from Stanford University, USA.

Among the sponsors of the 2007 ACL meeting there were Google, Microsoft, IBM, Xerox, TextKernel, BBN, Morphologic, NewsTin, Powerset, the Czech Association for Information Science and some others.

This was the third time when Prague hosted an international meeting on computational linguistics: after a rather small but for that time rather influential Colloquium on Algebraic Linguistics in 1964 there was the COLING Conference in Prague in 1982, with almost 400 participants, at the occasion of which the foundation of the European Chapter of ACL was announced by Donald Walker, the then ACL Secretary, accompanied by the establishment of the ACL International Fund which made it possible for considerably economically handicapped researchers from Central and Eastern European countries to be ACL members, receiving the journal and being supported in their participation at the ACL meetings. This was one of the greatest support we have got and by organizing the 2007 ACL meeting we also wanted to express our gratefulness.

Eva Hajičová
Local Organisation Chair

Jan Hajič
Sponsorship Chair

Pavel Straňák
Publicity Chair

link to webpage:
`http://ufal.mff.cuni.cz/acl2007`

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 88 DECEMBER 2007 95–98

REVIEWS

Connectives as Discourse Landmarks

Agnès Celle, Ruth Huart (eds.)

Amsterdam, Philadelphia: John Benjamins Publishing Company, 2007, 212 pp.
ISBN 978-90-272-5404-7

Reviewed by Šárka Zikánová

In the volume under review, papers by eleven authors are included which were presented at the international conference Connectives as Discourse Landmarks (University of Paris-Diderot, May 2005). The main point of interest of present studies are syntactic, semantic and pragmatic functions of several discourse connectives in English.

The term 'connective' is being used in a broad sense, without specific theoretical restrictions, herewith opening space for different treatments of discourse. In the book, this term covers not only traditional connective items like conjunctions *(and, but)* and relative pronouns *(which),* but also discourse adverbials *(rather, still, yet),* phrasal constructions *(after all)* as well as whole sentential frames *(the fact is that; it's not that; A because B so A')* and means of contact *(well, you know).* As this set of connectives shows, the studies deal with two large aspects of discourse research: with questions of syntax, semantics and lexicology, understanding connectives as items expressing the relations between sentences (abstract objects, events), and with pragmatics where connectives are understood as units linking the speaker and the hearer.

The editors' introduction describes briefly the historical context of present-day discourse studies, making short references to Richard G. Warner, Deborah Schiffrin and Jan-Ola Östman. Further, it explains the development of the discourse terminology and touches upon some open questions of the discourse research, namely the level of grammaticalization of connectives, the issue, whether for the meaning of connectives, their core lexical sense is more important or rather pragmatic sense variance in different contexts and, finally, the relation between the form (conjunctive, subjunctive) and the meaning of connectives. After general remarks on discourse, the main points of the studies included are shortly summarized.

In the "Part I. Connectives and modality", Raphael Salkie ("Connectives, modals and prototypes: A study of *rather*") focuses on common features of different senses of *rather* (connective,

degree modifier, part of modal *would rather*) and proposes a prototype approach to connectives and modality to catch its shared basic pragmatic function of narrowing down the possible interpretations of an utterance. Karin Aijmer ("The interface between discourse and grammar: *The fact is that*") explains the internal structure of 'shell noun phrases' such as *the fact / thing / trouble is (that)* and their development from matrix clause to a compound pragmatic marker. These pragmatic markers have several several variants, some of them are – from the syntactic point of view – ungrammatical *(fact is);* as the author claims, they can serve as an argument for the statement that 'shell noun phrases' are collocational frameworks rather than full matrix clauses. What can be found as confusing is the position of this article within the part of book concerning modality.

Mark de Vos ("*And* as an aspectual connective in the event structure of pseudo-coordinative constructions") in the "Part II. From syntax to pragmatics" deals with so called pseudo-coordinations of verbs including a verb such as *go / sit,* connective *and* and a lexical verb or including reduplicative coordination of the lexical verb *(Caesar went and read the parchment! Caesar sat and read the parchment. Caesar read and read in his tent all night.)* Describing carefully the meaning of these structures with regard to aktionsart and testing their syntactic properties in comparison with other coordinative constructions, the author points out that connective *and* can serve as means expressing the event structure on semantic and syntactic level. In Rudy Loock's article (*Are you a good which or a bad which?* The relative pronoun as a plane connective), specific utterances of *which* are analyzed, which fulfil no anaphoric function. In the surveyed atypical appositive relative clauses, either a resumptive pronoun appears at the position of a standard gap and one position seems to be expressed twice *(which – it),* or no gap (antecedent for the relative pronoun) is available. Thus, *which* in such constructions develops into a pure connective item. Diana M. Lewis ("From temporal to contrastive and causal: The emergence of connective *after all*") considers the historical evolution of the connective sense of the phrase *after all* arguing that it originates neither from a metaphorical use of an originally temporal *after,* nor from any ad hoc innovation of its justificative or counter-expectative sense, but rather from the metonymic expression of compressed information.

In the "Part III. Discourse strategies", Barbara Le Lan ("Orchestrating conversation: The multifunctionality of *well* and *you know* in the joint construction of a verbal interaction") emphasizes the pragmatic meaning of the term 'connective', describing the interpersonal role of these two items in the conversation, i.e. reference to the (supposed) point of view of the other speaker, as well as semantic components of cognitive control ('being familiar with something') and subjectivity. Frédérique Passot ("*A because B so A*': Circularity and discourse progression in conversational English") focuses on a quasi-repetitive conversational sequence of three sentences *A because B so A*' arguing that the structure of the sequence is not circular, but rather a dynamic spiral with a progression of information exchange and with the permanently updated confirmation of the shared knowledge between the speaker and the hearer. In Ruth Huart's article *("Not that… versus It's not that…"),* the different features of the two complex connectives are described, concerning especially the relation to presuppositions, the scope of negation, collocability with adverbs and the syntactic structure.

Martine Sekali ("*He's a cop but he isn't a bastard:* An enunciative approach to some pragmatic effects of the coordinator *but*") in "Part IV. In search of operations" suggests intralinguistic analysis for pragmatic aspects of different utterances of the connective *but* based on the Theory of Enunciative Operations. Working with the same theoretical frame, Graham Ranger ("Continuity and discontinuity in discourse: Notes on *yet* and *still*") analyzes how single senses of aspect, degree and argumentation with connectives *yet* and *still* are linked. François Nemo ("Reconsidering the discourse marking hypothesis. *Even, even though, even if,* etc. as morpheme/construction pairs") points out that the specific meaning of connectives in single utterances is influenced from two sides, by the on-going context and by the encoded meaning of morphemes and proposes a methodology for analysis of the meaning of discourse connectives.

As a whole, the book offers not only detailed descriptions of meanings and usage of single English connectives, but can be especially useful from the methodological point of view – as 'a textbook' of discourse studies, giving the reader variety of ways how to deal with discourse phenomena.

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 88 DECEMBER 2007 99–100

BOOK NOTICES

Víceslovné předložky v současné češtině
Studie z korpusové lingvistiky, sv. 3.
(Complex Prepositions in the Present-Day Czech)
(Studies from the Corpus Linguistics, vol. 3)

Renata Blatná

Praha: Nakladatelství Lidové noviny / Ústav Českého národního korpusu, 2006, 352 pp.
ISBN 80-7106-865-9

Notice by Jaroslava Hlaváčová

The third volume of the series Studies from the Corpus Linguistics brings an extensive analysis of the controversial concept of complex prepositions. The author tries to define complex prepositions using 7 conditions, the most important being the syntactic function. Then, she classifies the complex prepositions from both paradigmatic as well as syntagmatic points of view – form of their components, frequency (in the Czech National Corpus), valency, function, semantics. She also studies variants of complex prepositions (e.g., possibility of singular as well as plural form of incorporated nouns) and their representation in various genres, and also in spoken Czech. The book describes several hundreds of Czech complex prepositions. The list, together with all the reasonings, could serve as a basis for a profound discussion whether such a broad conception of complex prepositions is reasonable and legitimate.

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 88 DECEMBER 2007

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 88 DECEMBER 2007 103–106

INSTRUCTIONS FOR AUTHORS:

A Guide to Preparing Images of Trees with TrEd for Publishing

Petr Pajas

This short guide describes the process of preparing vector images of trees in the tree editor TrEd[1]. Vector image formats (such as PDF or EPS), unlike bitmap formats (JPG, PNG, BMP…), are suitable for publishing in printed journals or bulletins; in fact, PDF is a strongly recommended format for all graphics published in PBML. The greatest benefit of using a vector format over a bitmap format is that a vector image looks the same at arbitrary scale; it looks ok on the computer screen and even better when printed on paper by a high resolution printer. On the other hand, bitmap images that look ok on the screen often look ugly and rasterized when printed on paper (this applies especially to schemas with geometrical shapes, lines, and text, not that much to photographs). The purpose of this guide is therefore also to encourage you to use vector formats wherever possible.

Here we assume the document containing the images is prepared using XeLaTeX, PDFLaTeX, or LaTeX. Some popular office suites, such as OpenOffice.org 2.0, can handle images in PDF and EPS formats as well, so the images prepared according to these guidelines can be included in office documents as well.

The process requires the following steps:

1. Start TrEd. Use File→Open and select the file containing the desired tree, for example a file from the Prague Dependency Treebank 2.0. Find the desired tree in the file, e.g. using arrow buttons on the toolbar or by choosing the tree form the list displayed using View→List of sentences.

2. Customize the way the tree is displayed. Note that the tree on the resulting image will look just as it is displayed by TrEd on screen. The way TrEd displays the tree can be adjusted in many ways using so called stylesheets, which can be customized using View→Edit stylesheet. The syntax of stylesheets in TrEd is beyond the scope of this document; you can find the necessary information under the Help button in the left corner of the stylesheet editor or in the TrEd User's Manual[2].

[1] http://ufal.mff.cuni.cz/ pajas/tred

[2] http://ufal.mff.cuni.cz/ pajas/tred/ar01-toc.html

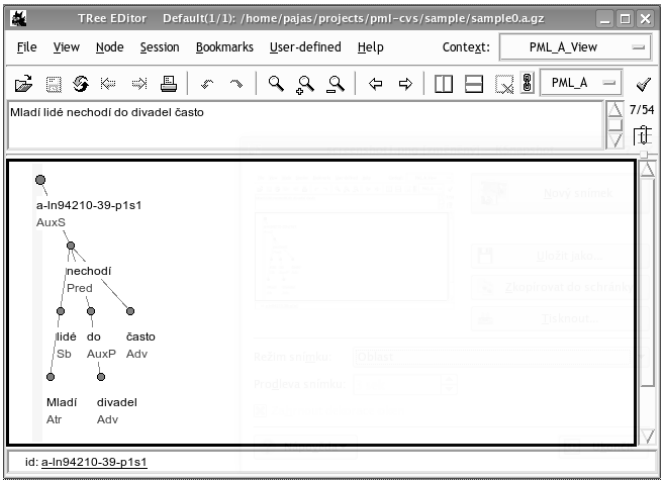


Figure 1. Tree Editor TrEd

3. If everything looks as it should in the resulting image, open the Print dialog using File→Print.

4. From the list Media, select the topmost item, BBox. Fill 0 as X margin and Y margin.

5. Select Print to file. If using XeLaTeX or PDFLaTeX, which is recommended for PBML, select Create PDF. TrEd will generate a list of TrueType fonts (TTF) available on your system; from this list, choose a font you would like to be used for node labels in the generated PDF. A good choice is a sans serif font, e.g. a PBML recommended and compatible font DejaVu Sans Book. (If you use LaTeX with DVI output, you may rather select Create EPS.)

6. To prepare a gray-scaled image (recommended for PBML), leave the option Use colors unchecked. If you check this option, the resulting image will have the same colors as displayed on the screen.

7. If you check Print filename and tree number, then the file name and the tree number will appear in a text line below the tree in the resulting image. Similarly, checking Print sentence causes the sentence of the tree (e.g. the text normally displayed above the tree in TrEd's main window) to appear below the tree in the resulting image. In most cases, you will want to leave both these options unchecked and provide a proper caption for the image using the `\caption{...}` command in the corresponding figure in your LaTeX document, where you can for example copy the sentence from TrEd's main window.

8. Adjust the resulting image file name in the field File name.

9. If you wish to create images of several trees from the current file, check One tree per file and use the formating pattern '%n' somewhere in the file name. This pattern will be replaced by the number of each tree. You may use '%03n' instead to force zero-padded 3-digit numbers.

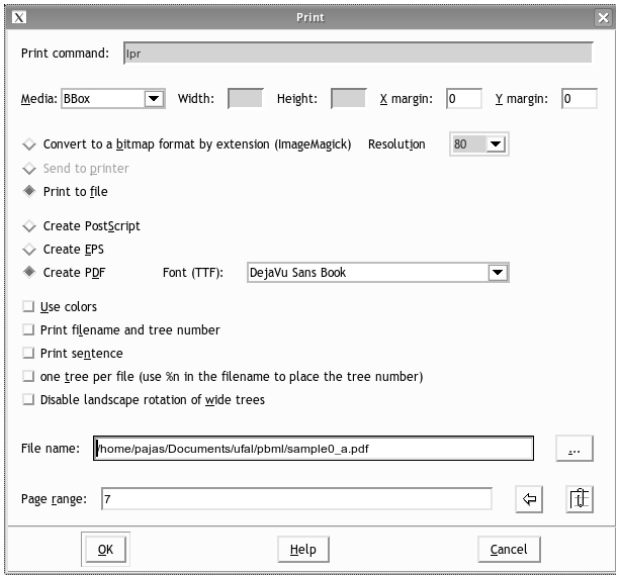


Figure 2. The Print dialog in TrEd

Now, list the tree numbers of the trees in the Page range field; for example, enter 3-11,30,42- in order to create image files for the trees three to eleven, thirty and for all trees starting from 42 to the end of the file. The button with an icon of a paper clip can be used to select the trees from the list of sentences.

10. Finally, press OK. This will generate the image file(s). You can now open the generated PDF images with Adobe Reader (or GhostView in case of EPS output) to see that the result looks as expected. Copy the resulting images to the folder containing your LaTeX document. Make sure to include the following line in the preamble of your LaTeX document.

```
\usepackage{graphicx}
```

In your document, use for example the following lines to include the image file sample0_a.pdf as a figure:

```
\begin{figure}[h]
\begin{center}
\includegraphics[scale=0.7]{sample0_a.pdf}
\end{center}
\caption{Sample image of an analytical tree of the Czech sentence
\emph{Mladí lidé nechodí do divadel často} from PDT 2.0.}
\label{fig:sample0-a}
\end{figure}
```

The parameter `scale=0.7` in the square brackets following `\includegraphics` scales the inserted PDF image by factor 0.7. You can use any other parameters accepted by the LaTeX `\includegraphics` command, e.g.
`\includegraphics[width=12.5cm]{sample0_a.pdf}`
or you can omit the sqaure brackets entirely, inserting the image as it is. The choice of parameters for `\includegraphics` and their values is a matter of your choice[3].

The result will look like Figure 3.

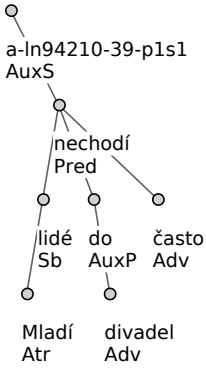


Figure 3. Sample image of an analytical tree of the Czech sentence Mladí lidé nechodí do divadel často from PDT 2.0.

[3]The list of all parameters can be found in the documentation of the `graphicx` package at http://www.ctan.org/tex-archive/macros/latex/required/graphics/grfguide.pdf

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 88 DECEMBER 2007

LIST OF AUTHORS

Eva Hajičová
Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
hajicova@ufal.mff.cuni.cz

Barbora Vidová Hladká
Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
hladka@ufal.mff.cuni.cz

Jaroslava Hlaváčová
Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
hlavacova@ufal.mff.cuni.cz

Petr Pajas
Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
pajas@ufal.mff.cuni.cz

Jiří Semecký
Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
semecky@ufal.mff.cuni.cz

Otakar Smrž
Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
smrz@ufal.mff.cuni.cz

Yuri Tambovtsev
Department of English and Linguistics
Novosibirsk Pedagogical University
P.O. Box 104
630123, Novosibirsk - 123, Russia
yutamb@mail.ru

Šárka Zikánová
Institute of Czech Language and Theory of Communication
Charles University
náměstí Jana Palacha 2
116 38 Praha 1, Czech Republic
sarka.zikanova@ff.cuni.cz