Formal Representation of Language Structures

Jan Hajič[1] Eva Hajičová[1] Alexandr Rosen[2]
[1]*Institute of Formal and Applied Linguistics*
Faculty of Mathematics and Physics
[2]*Institute of Theoretical and Computational Linguistics*
Faculty of Philosophy
Charles University
Prague, Czech Republic

Abstract

Building treebanks is a prerequisite for various experiments and research tasks in the area of NLP. Under a recently awarded grant,[1] we are developing (i) a formal definition of a (dependency based) tree, and (ii) a mid-size treebank based on this definition. The annotated corpus is designed to have three layers: morphosyntactic (linear) tagging, syntactic dependency annotation, and the tectogrammatical annotation. The project is being carried out jointly at the authors' Institutes.

1 The Current State and Motivation

Recent decades have seen a shift towards **expressing linguistic knowledge** in ways which allow **its verification and processing by formal means**. Tools originating in mathematics, logic and computer science have been applied to human language to model its structure and functioning. Various aspects of different languages are being described within formally defined frameworks proposed by a number of interacting linguistic theories.

The proposals deal with various levels of linguistic description, starting from the level of sounds (phonetics) up to the level of meaning. Partial grammars and lexicons now exist for many languages within various formal frameworks and collections of linguistic analyses of text and speech are accumulated to be employed both in theoretical research and applications. Besides approaches relying on symbolic means and 'rationalist' efforts which result in language models consisting of grammar rules and lexical entries, alternative methods employ statistics computed from input text or its analysis to produce a stochastic model.[2]

However, a common and crucial issue cutting across all types of enterprise in this domain is the need to adopt or design an **adequate formal representation of language structures** in order to accommodate relevant linguistic knowledge in its relation to the actual language data. There is a number of tasks which typically require soundly defined formal representation of language structures:

1. analysis (parsing) of input text or speech into a representation, tagging of text or speech collections;
2. synthesis (generation) of output text or speech from a representation;
3. mapping of one representation onto another -- transfer (typically in machine translation systems).

These are the elementary tasks which are parts of many natural language processing applications, some of which are listed below:

- machine translation systems;
- natural language interface to knowledge bases, question answering systems;
- automatic abstracting and knowledge acquisition systems;
- automatic acquisition of linguistic data and its integration into a language model.

---

[1] Grant of the Grant Agency of the Czech Republic No. 405/96/0198, which has now become an integral part of a newly awarded long-term grant of the same agency No. 405/96/K214

[2] When a linguistic description is implemented on computers, the usual goal is to parse sentences and produce representations of their analyses, thereby verifying the framework, the linguistic theory and the description itself. Another way to obtain (morphological and syntactic) analysis of sentences is by employing statistical methods on large samples of (already analyzed) texts in order to process a new text afterwards, performing some degree of linguistic analysis on the basis of the data acquired in the `learning' phase. Both these kinds of efforts converge and their increasing potential is reflected in the growing amount of text and speech data analyzed to a different degree for various purposes.

Formal representations of language structures which have been proposed by different linguistic theories and/or used in natural language processing applications reflect their context in many respects and suitable candidates for an intended more **general use** are difficult to find. This is due to various aspects of their design, such as (i) specific theoretical commitment, (ii) limited expressive power in partial coverage of language phenomena and restriction to certain levels of linguistic analysis, (iii) difficulties in expressing relationships between different levels of analysis, (iv) hard-wired reliance on some characteristics of a certain language or language group and the resulting difficulty in adapting the framework to a typologically different language, and, finally, (v) application-specificity. Thus, it is difficult to express a full-fledged syntactic analysis of a **'free word-order' language** by means of word-class labels and constituent brackets used for tagging (mostly English) texts.
Although it is not likely that a single framework could become a universally accepted vehicle of linguistic knowledge, we believe that a higher degree of generality and flexibility can be achieved for the benefit of both theoretical studies and application-oriented projects.

2 Characteristics of a Satisfactory Solution

From the conceptual point of view, an adequate design of formal representation should be able to express linguistic facts related to the following levels of description:

1. level of phonetics, phonology, graphemics: specification of phonemes, stress and prosodic patterns, etc.;
2. level of morphology: morphemes, morphological categories;
3. level of syntax: syntactic categories, syntactic structure (trees);
4. level of (linguistic) meaning: disambiguation of lexical meaning, specification of underlying structure and function, communicative dynamism and topic -- focus articulation, anaphora resolution.

There are several **important features** that should be reflected in the design to make it really useful:

- It should be possible to describe a language structure in **all its aspects** simultaneously, i.e., to be able to relate facts from all levels of linguistic analysis in a straightforward fashion. At the same time, the design should permit access to specific aspects of the description without other aspects intervening. Thus, a user interested only in syntactic structure should be able to filter out any other information.
- If a certain aspect of linguistic description can be structured and viewed differently depending on theoretical commitments, the design should provide an option to derive the desired way of presenting the linguistic facts from a common representation. Thus, **both phrase-structure and dependency trees** could be derived from the description.
- The design should be capable of accommodating **typologically different languages** without substantial modifications, especially, it should provide space for stating the relation between word-order variations and higher levels and for the interplay between morphology and syntax in the case of complex expressions.
- A related requirement concerns the possibility to express **links between parallel structures** and their analyses in different languages. This feature is important if parallel bi- or multilingual data are to be analyzed and studied as contrastive language structures.
- The design should provide space for as little or as much linguistic facts concerning a language structure as is possible or practical to collect or express. This feature would permit to integrate text or speech samples with their analyses in a **stepwise fashion**, possibly starting with a bare text/speech string and leaving some levels unspecified.
- It should be possible to represent at least some linguistic facts in an **underspecified** form. Wherever possible, an option to use a quantitative measure should accompany such cases. Disjunctions restricted to local domains, underspecified descriptions and weights could be the means to achieve this requirement.
- The formal representation should be **convertible** to another format, as required by an application or desired by another specification covering compatible conceptual issues.
- The design should be **flexible** in the sense that it should contain as few inherent restrictions to its extensions and modifications as possible.

3 Background, Methods and Problems

Without attempting to preview the results, the following points can be made to sketch the starting point situation, the outlines of the goal, and the path towards its achievement:

1. The project will be able **to profit from theoretical results and practical experience** gained in the field of formal description of natural language at our sites.

   The fruitful results concerning word-order variations and their relation to meaning, as well as the richness of syntactic studies based on a dependency-oriented model, both widely acknowledged and faithful to the high standards of the Prague School linguistic tradition, provide a wealth of stimulating material.

   At both sites, a number of application-oriented research projects have been at least in some respects tackling the problems of an adequate representation of language structures. The projects include machine translation, natural language interface to knowledge bases, automatic abstracting, automatic knowledge acquisition from texts and grammar checking.
2. The smallest piece of information (typically, a linguistic category) is expressed as **an attribute and its value** (i.e., a 'feature'). A collection (conjunction) of such pairs is used to describe a linguistic object (typically an aspect of linguistic description of a word or a collocation), allowing for partial information (underspecification) and entering into more complex structures, where some attribute values are not atoms but structures. Through the recursive nature of such a representation, linguistic structures of arbitrary complexity can be described. Two or more attributes can share a single value, which is a possible way to implement relations between linguistic facts at different levels of description.

   As structures of this type have become a kind of standard in modern linguistic research, the issues of compatibility with other approaches will be substantially simplified on many levels.
3. The design will be tested by its application on language data in at least two typologically different languages. A sample of **bilingual parallel text data** will be provided to test the parallel link option between analyses of linguistic structures.

There are a few **challenging issues** which call for an inventive solution:

- The relation between the surface string of graphemes/phonemes, hierarchical syntactic structure and the ordering of meaning-bearing elements according to the degrees of **communicative dynamism** is far from straightforward. This concerns especially cases of **crossing dependency** (non-projective structures). If the representation is to accommodate descriptions on all levels in an integral form, a non-trivial solution has to be found.
- **Complex expressions** like idioms, compound words and morphological categories realized by discontinuous sequences of auxiliary words present another problem of a similar kind.
- The **integration** of all kinds of linguistic knowledge **in a single formal framework** capable of application to the widest range of language structures is a unique enterprise. Disregarding the undoubtedly immense practical profit for a moment, the project will probably bring the most precious theoretical fruit precisely in this domain.

4 The Treebank

The formalism developed within this project will be applied towards a mid-size treebank, mainly on the Czech material. There will be three layers in the treebank.

1. Morphosyntactic Tagging

   This layer will represent the text in the original linear word order with a tag assigned unambiguously to each word form occurence, much like the Brown corpus does. The tagset used will reflect the richness of the Czech inflective morphology--there is about 1500 different tags in a flat list of tags needed for appropriate morphosyntactic description of Czech word forms.

2. Syntactic Dependency Annotation

   This will be the main treebank file. It will contain the (unambiguous) dependency representation of every sentence, with features describing the morphosyntactic properties, the syntactic function, and the lexical unit itself. All words from the sentence will appear in its representation.

   This part of the treebank will be done semi-automatically, possibly using the results of syntactic analysis of Czech which is being worked on under the Grammar Checker project we are involved in (EU/PECO). We have already finished development of software which will facilitate the handwork involved (graphical tree editors for both DOS/Windows and Unix (OpenLook/Linux) environments).

The morphosyntactic tagging layer can be derived automatically from the syntactic dependency layer.

3. Tectogrammatical Representation

At this level of description we will annotate every (autosemantic non-auxiliary) lexical unit with its tectogrammatical function, position in the scale of the communicative dynamism and its grammatemes (similar to the morphosyntactic tag, but only for categories which cannot be derived from the word's function, like number for nouns, but not its case).

We expect to annotate a subset of the treebank at this layer, as it is unclear to what extent this could be done automatically starting at the syntactic layer.

5 Cooperation, Links and Previous Experience

The problems and methods outlined above are being tackled at different places in the world, and there are some previous results both at the authors' sites and elsewhere to use and/or be inspired by them.

As mentioned in the previous section, there has been a long history of theoretical linguistics in Prague. In the past 30 years, formal representations of several kinds have been developed, (parts of) which can be used as a starting point [1], [2], [3]. However, there has not (yet) been an attempt to solve the representation problem in such a complex and unified way. Also, the development of the past ten years will lead to novel approaches in the representation theory.

However, the idea of the "development cycle" involving immediate, large-scale evaluation and verification on real texts has not been exploited previously in the framework of such a theoretical issue as a formal representation of language structures undoubtedly is. There are various projects, mainly in the United States, which do use the repetitive evaluation strategy to get valuable feedback, but they are more application-oriented. We feel that an appropriate modification and proper usage of such methods would mean a **qualitative leap** in a search for a theoretical result in a non-technical discipline. We would like to cooperate as much as possible with the centers doing a lot of work in this direction, namely, the LDC (Linguistic Data Consortium) at the University of Pennsylvania, and use their materials, especially for the evaluation phase of the English side.
There are also projects the results of which (or at least some of them) would help this project: this would also make very effective use of funds spent on other grants and research activites both within and outside of the Czech Republic. We envisage the **use of some of the results obtained in the following projects**: Grammar Checking for Slavic Languages (a PECO project, funded by the EU), from which we would like to obtain some ideas about representations of ill-formed input; Czech National Corpus project (funded by GAČR), as a resource of Czech textual material; and MATRACE (also funded by GAČR), as a starting point for comparison (and later, unification) of structural representations developed for the purpose of machine translation between two typologically different languages.

6 A Summary of the Goals

There are two main goals to be achieved:

- A **specification** and thorough **description** of a single **formal representation of language structure**, integrating and enhancing the previous theoretical results, and adding new contributions at the same time (especially the representation of topic/focus, coreference, discontinuous elements relations, etc.);
- An **experimental verification** of the above, i.e. the markup of a substantial portion of diverse, real text samples using the formal specification developed under the grant. In other words, building a treebank. Two **typologically different languages** will be used for the experiments, Czech and English.

We consider the two goals mutually indispensible, as we believe that only a rigorous testing of any formal representation theory will put it on a solid ground, and it will make an immediate feedback possible.

References

[1] Petr Sgall, Alla Goralčíková, Ladislav Nebeský and Eva Hajičová, *Functional Approach to Syntax*, American Elsevier, New York, 1969

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

[3] Vladimír Petkevič, *A New Formal Specification of Underlying Representations*, Theoretical Linguistics 21, 7-61, 1995