# CONSTRUCTION OF AN HPSG TREEBANK FOR POLISH

**Małgorzata Marciniak, Agnieszka Mykowiecka,**
**Adam Przepiórkowski, Anna Kupść**[*]

## Résumé - Abstract

Cet article présente les aspects conceptuels et techniques concernant la construction d'un corpus annoté pour le polonais. Le corpus contient des phrases du polonais écrit représentées comme des structures AVM dans le cadre de formalisme HPSG. En plus, chaque phrase est annotés pour les types des phénomènes linguistiques illustrés. Car le corpus est aussi utilisé comme une base pour tester des grammaires (un test-suite), des phrases bien et mal formées sont inclues également. On décrit l'organisation technique de la base des données de corpus et des opérations sur cette base.

The paper presents both conceptual and technical issues related to the construction of an HPSG Treebank for Polish. The treebank consists of sentences of written Polish encoded in HPSG-style AVM structures. Additionally, each sentence is annotated with a list of linguistic phenomena it illustrates. Since the treebank serves also as a test-suite, both grammatical and ungrammatical sentences are provided. We describe also a technical organization of the database which contains the treebank as well as possible operations on this database.

## Mots Clefs - Keywords

corpus annotés pour la syntaxe, test-suites, HPSG, polonais

treebank, test suit, HPSG, Polish

[*] Institute of Computer Science, Polish Academy of Sciences. E-mail: {mm, agn, adamp, aniak}@ipipan.waw.pl

## INTRODUCTION

The aim of this paper is to describe the construction of a treebank of written Polish sentences, created as part of the European Union CRIT-2 project. At the moment, the project is past the design phase, at the end of tools-creation phase, and at the beginning of actual data entering stage.

## 1. AIMS AND DESIGN CONSTRAINTS

The treebank described here was designed as a test-suite; in fact the design was based on existing test-suites for various European languages created within the TSNLP project (Lehmann S. *et al.* 1996; Oepen S. *et al.* 1998). As a test-suite, the treebank for Polish contains also ungrammatical sentences, violating various linguistic rules.

The most common use of test-suites is evaluating computational grammars (i.e., parsers) of a given language: if a parser is able to parse most or all of the correct sentences and none of the ungrammatical sentences in the test-suite, then its coverage is very extensive. Thus, the empirical adequacy of parsers can be quantitatively evaluated by examining how they deal with respect to the data in the test-suite. They can also be qualitatively evaluated by comparing the parses they produce to the exhaustive annotations contained in the treebank.

Similarly, reference grammars, textbooks and even particular syntactic analyses can be evaluated against the data in the treebank.

The immediate aim of the treebank is the evaluation of a grammar of a fragment of Polish, which was written within the Head-driven Phrase Structure Grammar (HPSG) (Pollard C. & Sag I. A. 1994) formalism and which is currently being implemented. This aim has a number of design consequences:

1. the treebank contains sentences of written Polish;
2. it will have very extensive empirical coverage (see §2 below);
3. it is manually-annotated by a group of linguists and computational linguists, to some extent unaware of the solutions adopted in the grammar;
4. sentences are annotated with HPSG-style Attribute-Value Matrices (AVMs) (see §3 below).

As to 1, sentences are elicited instead of, e.g., being extracted from a text corpus. This allows us to represent in the treebank also less common phenomena which rarely occur in real corpora and to reduce the number of lexical entries used in examples.

As to point 2 above, once the treebank represents syntactic phenomena of Polish near-exhaustively, the existing grammar of Polish, and other such formal grammars, can be *quantitively evaluated* by comparing its coverage to that of the treebank. Thus, the treebank will also play the role of a test-suite; in fact, the design of this treebank was inspired by the TSNLP project (Lehmann S. *et al.* 1996; Oepen S. *et al.* 1998).

As to 3, the sentences are hand-annotated, as far as possible, without any concern for the often computationally-motivated solutions adopted in the existing HPSG grammar of Polish. The comparison of the annotations in the treebank with the parses of the grammar will allow for *qualitative evaluation* of the latter.

As far as 4 is concerned, the aim of the treebank (i.e., evaluation of an HPSG grammar) does not, strictly speaking, enforce an HPSG annotation scheme. Such a scheme was nevertheless adopted for the following reasons:

- it facilitates comparing parses with treebank annotations, i.e., it facilitates the evaluation of the HPSG grammar for Polish;
- HPSG mechanisms, i.e., feature structures and multiple inheritance type hierarchy, provide a uniform means for representing various types of linguistic information, including syntactic and morphosyntactic structures;
- HPSG is one of the leading formalisms used in computational linguistics; hence, the annotation format may be readily understandable to computational linguists;
- HPSG is a linguistic formalism with a large body of literature; hence, analyses of various phenomena can be modeled on those in the literature;
- the final version of the treebank will contain some semantic information, and HPSG structures allow for encoding such information and for modelling interactions between syntax and semantics.

## 2. LINGUISTIC PHENOMENA

Each sentence in the treebank is annotated with a list of linguistic phenomena (so-called indices) illustrated by this sentence. The classification of syntactic phenomena of Polish is constructed on the basis of similar classifications proposed for German, English and French described in (Lehmann S. *et al.* 1996) and (Oepen S. *et al.* 1998), but it has been elaborated specifically for Polish. Altghough the treebank will contain only a restricted number of clauses, they will reflect a large number of syntactic interralations characteristic for Polish.

The main groups of phenomena taken into consideration are: general types of utterances (declarative, interrogative, etc.); tense, aspect and modality; valency types; diathesis; types of modification; agreement; coordination; negation; word order. Each of these groups is subdivided into more specific phenomena of various levels of specificity, thus forming a hierarchy of linguistic phenomena of Polish. Each of these specific phenomena is illustrated with both grammatical and ungrammatical utterances. Below, we briefly characterize these main groups.

## 2.1. Phenomena groups

### 2.1.1. Types of utterances

We divide utterances into declarative, imperative and questions. Questions are split into *wh-* and *yes/no* questions, questions *in situ*, infinite and verbless questions. Imperative utterances are divided into those which contain imperative verb forms, those beginning with the special word *niech* 'let' and declarative sentences with the exclamation mark. Declarative utterances are finite verb clauses which may be syntactically compounded with subordinated clauses: indirect questions, relative clauses and clauses beginning with complementizers (*że, żeby*).

### 2.1.2. Tense-Aspect-Modality

This class groups verbs according to their aspect and describes how tensed or passive forms are obtained. Aspect in Polish is lexically (morphologically) encoded. The formation of tenses differs if a perfective or imperfective verb form is used. Perfective verbs lack present tense forms and they can be used only as past and future tense forms. Imperfective verbs have all tensed forms but the future tense form requires the auxiliary verb *być* 'to be'.

### 2.1.3. Complementation

This class provides possible complementation frames of nouns and verbs. We enumerate different realizations of the subject as well as complementation of auxiliary verbs. We provide also a classification of words' valency with respect to the number of arguments (zero, one or more) and their type (nominal, prepositional, adverbial, numeral or verbal phrases).

### 2.1.4. Diathesis

Diathesis describes changes in predicate-argument structure of a verb. This group represents verb voices, i.e., passive, active and reflexive verb forms.

### 2.1.5. Modification

We describe noun, verb and adjective modification. We classify modification types with respect to the type of a modifier (a noun, an adjective, an adverb, etc.) as well as the type of a modified phrase. Agreement principles (if any) which must hold between a modifier and a modified phrase are described within the agreement group.

### 2.1.6. Agreement

We describe two basic types of agreement: agreement within NP and subject-predicate agreement. We distinguish various types of agreement within NP according to the syntactic category of NP components, i.e., nouns, adjectives, pronouns, numerals, etc. Also subject-predicate agreement depends on the form of the subject.

### 2.1.7. Coordination

This group describes types of phrases which can be coordinated and types of conjuncts. We specify separately elliptical and nonelliptical constructions.

### 2.1.8. Negation

This class describes sentential and constituent negation. In particular, we distinguish idiosyncratic negation of the existential copula *być* ‘to be’. We represent also the so-called genitive of negation, i.e., the obligatory change to the genitive case of an accusative complement if the verb is negated. In Polish, the presence of an *n*-word, e.g., *nikt* ‘nobody’, *nigdzie* ‘nowhere’, triggers verbal negation. This phenomenon, the so-called negative concord, is also reflected in the classification.

### 2.1.9. Word Order

Polish is a relatively free order language but linear order in Polish is not unconstrained. This class captures several general facts of Polish linear order. Prepositions and numerals must precede their nominal complements. Relative clauses have to follow noun phrases they modify while a conjunction has to be placed between conjuncts. We represent also for the placement of the negative marker as well as verbal clitics.

## 2.2. Hierarchy of phenomena

The linguistic phenomena are organized into a hierarchy. The name of each phenomenon contains the name of its supertype. Hence, it is possible to refer to an entire group of phenomena just by using a prefix included in all appropriate names. A fragment of the hierarchy which represents types of questions is given below:

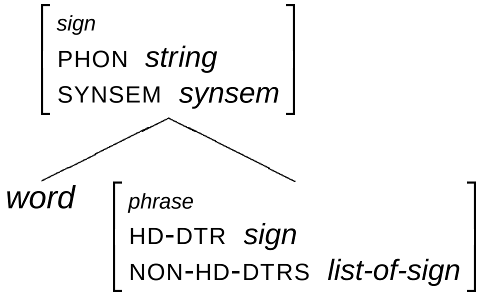
- C-Question
    - C-Question-wh
        - C-Question-wh-initial
            - C-Question-wh-initial-fin
            - C-Question-wh-initial-infinitive
        - C-Question-wh-insitu
            - C-Question-wh-insitu(repr)
            - C-Question-wh-insitu(nonrepr)
    - C-Question-yn
        - C-Question-yn-particle
            - C-Question-yn-particle-ind
            - C-Question-yn-particle-dep
        - C-Question-yn-intonation
            - C-Question-yn-intonation-fin
            - C-Question-yn-intonation-infinitive
    - C-Question-verbless

### 2.3. Complexity

Sentences included in the treebank are divided into several groups, reflecting their grammaticality and complexity. One group consists of ungrammatical sentences annotated with names of the phenomena which are violated in a particular sentence. Grammatical examples are divided into different test sets. At the moment, we use three test sets representing, respectively, basic constructions, more complex constructions, and very complex or peripheral constructions.

## 3. ANNOTATION SCHEMA

The sentences (as well as wordforms, see §4 below) are annotated with Attribute-Value Matrices (AVMs), as used in HPSG.[1] In particular, each AVM is of a certain type, where possible types constitute a multiple inheritance type hierarchy. This type hierarchy specifies, for each type, its immediate subtypes and supertypes, as well as attributes appropriate for this type (and possible values of these attributes). A small part of the type hierarchy adopted is given below. It says that the type *sign* has two immediate subtypes, *word* and *phrase*, that there are two attributes appropriate for *sign* (and all its subtypes), i.e., PHON (with values of type *string*) and SYNSEM (with values of type *synsem*), and there are two additional attributes appropriate for *phrase*, i.e., *sign*-valued HD-DTR and *list-of-sign*-valued NON-HD-DTRS.



Each sentence is annotated with an AVM of type *phrase*, with the orthography of the sentence represented by the value of PHON,[2] the morphosyntactic, etc., information represented by SYNSEM and the constituency structure encoded (for headed phrases) via HEAD-DTR and NON-HEAD-DTRS. Deeper levels of AVM structures are consistent with current HPSG theorizing, e.g., SYNSEM values are divided between LOCAL and NONLOCAL attributes, the former further divided into CATEGORY, CONTENT and CONTEXT, etc.

However, 1) not all attributes assumed in current HPSG are represented in the current version of the treebank, and 2) values of some attributes are adapted to Polish. For example, pragmatic (CONTEXT) information is ignored, while semantic (CONTENT) information is represented only provisionally at this

[1] The standard reference for AVMs, as used in HPSG, is (Carpenter B. 1992).

[2] The name of this attribute is a misnomer in the present context, but it was retained for consistency with standard HPSG (Pollard C. & Sag I. A. 1994).

stage (although it will be extended at later stages). On the other hand, the values of the morphosyntactic attributes such as GENDER and CASE had to be extensively modified (Czuba K. & Przepiórkowski A. 1995).

An example of the (partial) annotation (for *Janek widzi Marysię*, lit.: 'Janek$_{nom}$ sees Mary$_{acc}$ ') is given in Fig. 1.

## 4. TECHNICAL ISSUES

The HPSG Treebank for Polish is a database[3] of Polish sentences (the HPSG Treebank proper), with another, auxiliary, database containing Polish wordforms (a dictionary). Each sentence is annotated with a correctness marker, a list of linguistic phenomena illustrated by this sentence (so-called indices), and a list of its syntactic analyses in the form of HPSG structures.

There are two text files restricting the content of the database and its interpretation. One of them contains an HPSG signature, i.e., the multiple inheritance hierarchy of types, and names of attributes appropriate for each type, as well as possible values of these attributes. The other file contains the hierarchy of linguistic phenomena of Polish covered by the Treebank.

Correct sentences are augmented with their (one or more) HPSG analyses in the form of AVM structures, constructed according to an HPSG signature, given as a separate text file. This signature is converted into a database description. This signature should be created prior to the creation of the database but some modifications of the signature are possible also afterwards.

The dictionary is a separate part of the database. It consists of the AVM structures of inflectional forms used in sentences contained in the Treebank. Each inflectional form is linked to the base form of the word. If the base form of some inflectional form is not present in the dictionary, the user is asked to enter it.

The most important two groups of operations on the Treebank are entering and searching data.

### 4.1. Entering operations

- It is possible to enter sentences, their correctness markers, and their indices (phenomena names).
- Entering sentence parses (AVMs) are facilitated. Parts of AVMs are generated automatically, values of attributes are filled in manually: after giving a type name, attributes appropriate for this type are generated. The correctness of the information thus entered is partly verified, e.g., the appropriateness of attribute values and the consistent use of AVM's labels (so-called tags). During the edition of the structure, it is possible to view the parse tree in another window.
- It is possible to modify the data in the Treebank through the following

[3] This database is implemented in Delphi(Borland) in Microsoft Windows NT environment.

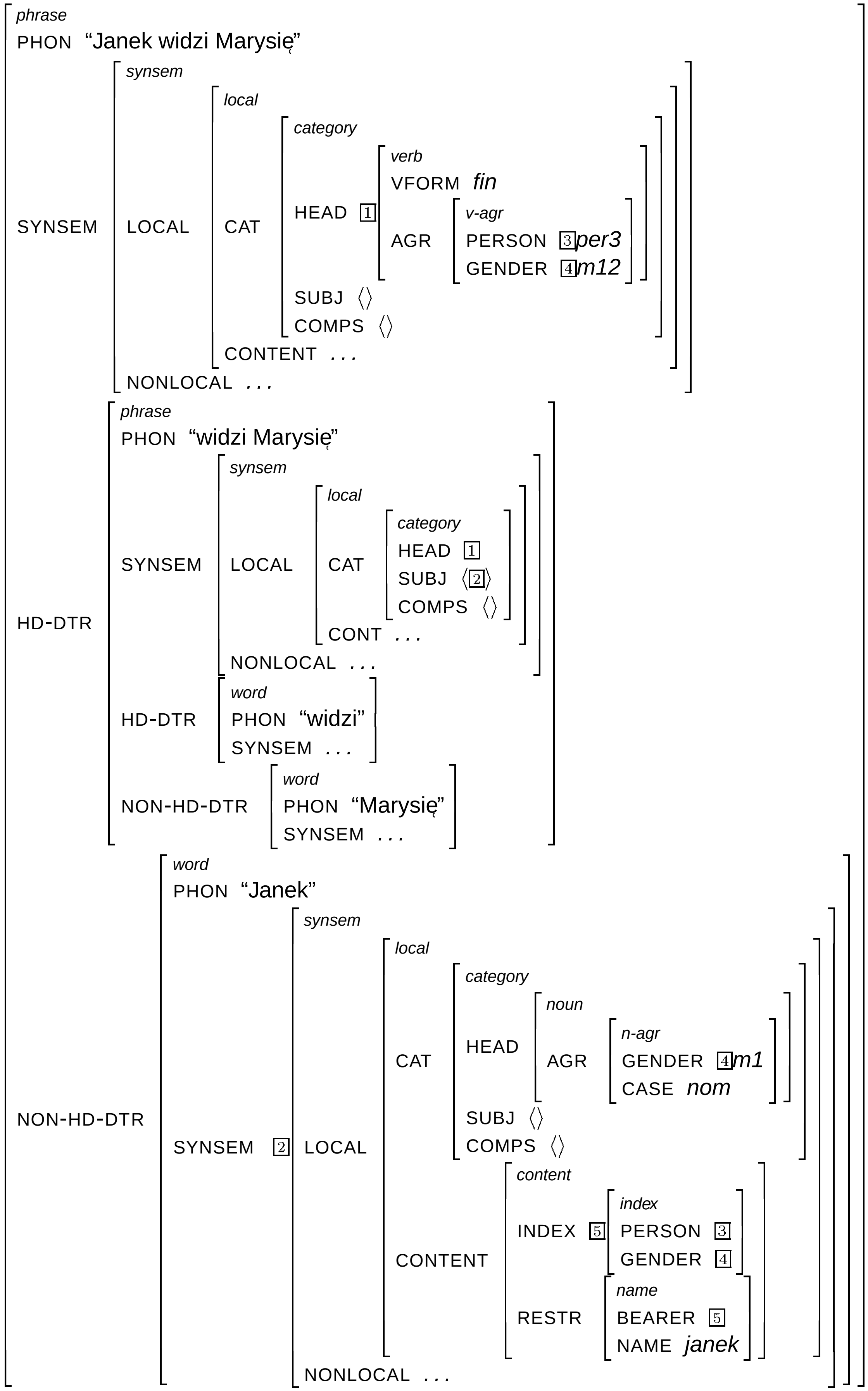

phrase
PHON “Janek widzi Marysię”
SYNSEM
synsem
LOCAL
local
CAT
category
HEAD 1
verb
VFORM fin
AGR
v-agr
PERSON 3 per3
GENDER 4 m12
SUBJ ⟨⟩
COMPS ⟨⟩
CONTENT . . .
NONLOCAL . . .
HD-DTR
phrase
PHON “widzi Marysię”
SYNSEM
synsem
LOCAL
local
CAT
category
HEAD 1
SUBJ ⟨2⟩
COMPS ⟨⟩
CONT . . .
NONLOCAL . . .
HD-DTR
word
PHON “widzi”
SYNSEM . . .
NON-HD-DTR
word
PHON “Marysię”
SYNSEM . . .
NON-HD-DTR
word
PHON “Janek”
SYNSEM 2
synsem
LOCAL
local
CAT
category
HEAD
noun
AGR
n-agr
GENDER 4 m1
CASE nom
SUBJ ⟨⟩
COMPS ⟨⟩
CONTENT
content
INDEX 5
index
PERSON 3
GENDER 4
RESTR
name
BEARER 5
NAME janek
NONLOCAL . . .

Fig. 1

operations:

- removing one of the indices assigned to a sentence,
- adding a new index to a sentence,
- changing an attribute value in one of the parses of a sentence,
- adding a new parse to a sentence.

### 4.2. Search operations

It is possible to search the Treebank according to the types of phenomena they represent, their correctness markers and according to the information present in AVM structures. In all three cases, the result of the search will be a list of sentences together with their parses.

It is possible to search the database for sentences illustrating syntactic phenomena and combinations of such phenomena. It is possible to search by a prefix of the phenomenon's name.

It is possible to search the Treebank according to information included in the parses (AVM structures), e.g. searching for inflectional forms, their base forms and types of constructions. Minimally, the user is allowed to ask about:

- parses containing structures of the specified type,
- the inflectional form of a word,
- all forms of the specified word.

The parses of a sentence can be shown on the screen in two formats: as trees and as AVM structures. It is possible to fold and unfold substructures, to hide selected attributes, to show the structure corresponding to a tag. It is also possible to output parses (AVMs) to a file, both in pure text and in LATEX formats.

## RÉFÉRENCES

CARPENTER, Bob (1992) : *The Logic of Typed Feature Structures*, Cambridge University Press, *Cambridge Tracts in Theoretical Computer Science*.

CZUBA, Krzysztof ; PRZEPIÓRKOWSKI, Adam (1995) : *Agreement and Case Assignment in Polish: An Attempt at a Unified Account*, Rapport technique n° 783, Institute of Computer Science, Polish Academy of Sciences.

LEHMANN, Sabine ; OEPEN, Stephan ; REGNIER-PROST, Sylvie ; NETTER, Klaus ; LUX, Veronika ; KLEIN, Judith ; FALKEDAL, Kirsten ; FOUVRY, Frederik ; ESTIVAL, Dominique ; DAUPHIN, Eva ; COMPAGNION, Hervé ; BAUR, Judith ; BALKAN, Lorna ; ARNOLD, Doug (1996) : "TSNLP — test suites for natural language processing", *in Proceedings of COLING 1996*, Kopenhagen.

OEPEN, Stephan ; NETTER, Klaus ; KLEIN, Judith (1998) : “TSNLP — test suites for natural language processing”, *in Linguistic Databases,* J. Nerbonne (eds.), Stanford, CSLI Publications.

POLLARD, Carl ; SAG, Ivan A. (1994) : *Head-driven Phrase Structure Grammar*, Chicago, Chicago University Press.