An HPSG-Annotated Test Suite for Polish

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

[*]Institute of Computer Science, Polish Academy of Sciences
J. Ordona 21 01-237 Warsaw, Poland
email: {mm,agn,adamp,aniak}@ipipan.waw.pl
[†] University of Tübingen
Wilhelmstr. 113 72074 Tübingen Germany
[‡] The Ohio State University, Department of Linguistics
222 Oxley Hall, 1712 Neil Ave. Columbus OH 43210 USA

Abstract

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

1. Aims and Design Constraints

The aim of this paper [1] is to describe **B**aza **R**ozbiorów **G**ramatycznych (Database of Grammatical Parses) — a test-suite of written Polish sentences, created as a part of the European Union CRIT-2 project. The idea is based on the TSNLP project (Lehmann et al., 1996; Oepen et al., 1998), which resulted in creating test-suites for various European languages. As a test-suite, BRG contains not only grammatical sentences but also ungrammatical ones, violating various linguistic rules. At the moment, the project is at the final stage of data entering.

The test-suite contains sentences of written Polish. They are hand-annotated with correctness markers, lists of linguistic phenomena names and HPSG-style Attribute-Value Matrices (AVMs) (see §4. below). Sentences included in BRG are elicited instead of, e.g., being extracted from a text corpus. This allows us to represent in the test-suite also less common phenomena which rarely occur in real corpora and to reduce the number of lexical entries used in examples.

The aim of the test-suite is to evaluate computational grammars of Polish (i.e., parsers). The empirical adequacy of parsers can be quantitatively evaluated by examining how they deal with respect to the data in the test-suite. Parsers can also be qualitatively evaluated by comparing the parses they produce to the exhaustive annotations contained in BRG.

The immediate aim of the test-suite is the evaluation of an HPSG (Head-driven Phrase Structure Grammar (Pollard and Sag, 1994)) grammar of a fragment of Polish, which is currently being implemented. Choosing an HPSG annotation scheme facilitates comparing parses with test-suite annotations, but there are also more general reasons for this decision. 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 also one of the leading formalisms used in computational linguistics, so the annotation format may be readily understandable to computational linguists.

[1]Earlier versions of this paper were presented at ATALA Workshop on Treebanks in Paris, June 1999, at Third European Conference on Formal Description of Slavic Languages held in Leipzig, 1999 and at Tenth CLIN Meeting in Utrecht, 1999.

2. Correctness and Complexity Markers

Sentences contained in the database can be divided into several groups on the basis of their grammaticality and complexity (see also Bańko (1990) for a discussion of test data selection). Each group is labeled with a name of 'correctness marker'. These markers are defined and entered during the initialization of the database.

In this project we assume two-step classification consisting of 6 markers. The main classification divides sentences into correct and incorrect. In the second step we partition both correct and incorrect sentences into three subgroups: basic constructions, complex constructions and very complex or peripheral constructions. The number of sentences labeled with each correctness and comlexity marker is presented in Table 1.

During the evaluation of the formal grammar it is important to know how complicated the particular sentence is, as well as what kind of grammatical infelicity is represented in the analyzed sentence. For example, a fundamental grammatical phenomenon can be violated, e.g., the agreement between the verb and the nominative subject in *Chłopiec śpią* 'A boy sleep$_{pl}$.' Such a sentence must be rejected by any parser of Polish and, for this reason, it is important to mark a set of basic correct and incorrect sentences that any parser should get right. The following sentences illustrate our use of the complexity marker:

- correct–basic: *Jan widzi Marysię* 'John sees Mary.'
- correct–complex: *Po co i z kim chcesz tam jechać?* 'What for and with whom do you want to go there?'
- correct–peripherial: *Dzieci zjadły chleb i sera* 'The children ate bread$_{acc}$ and cheese$_{gen}$' (coordinated nouns are in different cases).

Marker	Number
correct–basic	116
correct–complex	68
correct–peripherial	9
subtotal	193
incorrect–basic	83
incorrect–complex	58
incorrect–peripheral	6
subtotal	147
total	340

Table 1: Number of sentences

- incorrect–basic: *Chłopiec śpią* 'A boy sleep$_{pl}$.'
- incorrect–complex: *Marysia jest najzwinniejsza jak kot* 'Mary is the most agile as a cat.'
- incorrect–peripheral: *Matka*$_{fem}$ *i jej ukochane dziewczę*$_{neutr}$ *poszli*$_{masc-hum}$ *razem* 'Mother and her beloved girl went together.' (In this sentence, a very untypical agreement pattern between coordinated nominative phrase and verb holds. The correct sentence is following: *Matka i jej ukochane dziewczę poszły*$_{non-masc-hum}$ *razem.*)

3. Linguistic Phenomena

Each sentence in the test-suite is annotated with a list of linguistic phenomena (so-called indices) illustrated by this sentence. An incorrect sentence is labeled with phenomena which are violated (they are marked with $(*)$), as well as with phenomena which can be observed in this sentence.

The classification of syntactic phenomena of Polish is constructed on the basis of similar classifications for German, English and French (Lehmann et al., 1996; Oepen et al., 1998), but it has been elaborated specifically for Polish (Marciniak et al., 2000). Although the database contains only a restricted number of clauses, they reflect a large number of syntactic phenomena characteristic for Polish, as well as their interrelations. In the project, we distinguish the following nine main groups of phenomena:

- Types of utterances

  In this class we distinguish three types of core clauses: interrogative (C-Que), imperative (C-Imp) and declarative (C-Decl) utterances.

  Interrogative sentences are divided into *yes/no* and *wh*-questions. *Yes/no* questions are further divided into those which are only marked with a question mark and those which begin with special question particles. *Wh*-questions are divided into regular and *in situ* questions, which in turn are subdivided into reprise and non-reprise questions.

  Imperative utterances are divided into those which contain imperative verb forms, those which begin with the special word *niech* 'let', and declarative sentences with the exclamation mark.

  Apart from core clauses, we describe also subordinate (C-Sub) clauses, for which we indicate the way they are introduced. We distinguish, thus, indirect questions, relative clauses, and clauses introduced by complementizers (e.g., *że*, *żeby*).

- Verbs types and forms

  This class of phenomena describes types of verbs (we distinguish proper verbs (Verb-V) and quasi-verbs (Verb-QV)) and various phenomena related to verbal forms (Verb-forms). This class includes also such properties of the verb as diathesis (Verb-Diathesis), aspect (Verb-Aspect) and tense (Verb-Tense).

  The subclass Verb-Diathesis represents changes in the predicate-argument structure of the verb and includes passive, active and reflexive verb forms.

  The Verb-Aspect subclass groups verbs according to their aspect, which, in Polish, is lexically (morphologically) encoded.

  The Verb-Tense comprises phenomena related to tense forms. 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 in past and future tenses. Imperfective verbs have all tensed forms but the future tense form requires the auxiliary verb *być* 'to be'.

- Adjectives and Adverbs

  This small class comprises phenomena describing adjectival and adverbial forms according to the degree and the way they are graded.

- Complementation

  This class reflects possible complementation frames of verbs (Compl-V), nouns (Compl-N), prepositions (Compl-Prep), adjectives (Compl-Adj), adverbs (Compl-Adv) and numerals (Compl-Num). We provide a classification of words' valencies with respect to the number of arguments (from zero to four) and their type (nominal, prepositional, adverbial, numeral or verbal phrases). With respect to the subject requirement, we distinguish nominative, sentential and non-typical (dummy or non-nominative) subjects.

  For lexemes which have more than one requirement, each of them is represented by a separate phenomenon name.

- Modification

  In the first step, we classify modification types with respect to the type of the modified phrase. Thus, we distinguish: Modification-NP, Modification-VP, Modification-AdjP, Modification-AdvP and Modification-Num. These classes are then further divided with respect to the category of the modifier (nominal, adjectival, adverbial modifier, etc.).

  Agreement principles (if any) which must hold between the modifier and the modified phrase are described in the Agreement class.

- Agreement

  We distinguish two basic types of agreement in Polish: agreement within NP (Agreement-NP) and subject-predicate agreement (Agreement-NomSubj). Agreement within NP is further divided according to the syntactic category of NP components, i.e., nouns, adjectives, pronouns, numerals, etc. Subject-predicate agreement depends on the form of the subject. If the subject is a coordinated phrase, usually two different verb forms are allowed.

  A separate group of phenomena (Agreement-Coord) deals with agreement between elements of coordinated verbal, nominal, adjectival and numeral phrases (cf. Coordination below).

- Coordination

  Within this group, we differentiate coordinated constructions according to the type of coordinated phrases (Coordination-NP, Coordination-VP, Coordination-AdjP, Coordination-Num, Coordination-AdvP, Coordination-Prep, Coordination-Pron-wh, Coordination-Clause) and then according to the type of the conjunction used. Conjunctions are divided into five classes: left-right, central, incorporational, serial disjunctive and serial conjunctive.

- Negation

  This class describes constituent (NP, Prep, Num, AdjP, AdvP) and sentential (C) negation.

  In sentential negation, we distinguish idiosyncratic negation of the existential copula *być* 'to be'. We represent also the so-called genitive of negation, i.e., the 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 sentential negation. This phenomenon, the so-called negative concord, is also reflected in the classification.

- Word Order

  The linear order in Polish is relatively free but it is not unconstrained. Word Order class captures several general facts of Polish linear order, e.g., relative clauses have to follow noun phrases they modify, a conjunction has to occur in a place appropriate for its type. We deal also with the placement of the negative marker as well as verbal clitics.

Each of these groups is subdivided into more specific phenomena of various levels of specificity, thus forming a hierarchy of linguistic phenomena of Polish. The number of each phenomena group is presented in Table 2.

Each of the most specific phenomena is illustrated with both grammatical and, if possible, with ungrammatical Polish utterances (we assume that ungrammatical sentences have to be incorrect in all readings). Incorrect sentences can be related in BRG to their correct versions. The sentence *Jan iść* 'John go$_{inf}$' can be correlated with the correct sentence *Jan idzie* 'John goes', where the wrong form of verb is used or with imperative sentence *Iść* 'Go', which cannot have a nominative subject. So the above incorrect sentence can be entered into BRG twice with different sets of indices.

Phenomena group	Number
Types of utterances	43
Verbstypes and forms	34
Adjectives and Adverbs	10
Complementation	77
Modification	20
Agreement	24
Coordination	20
Negation	17
Word Order	19
Total	264

Table 2: Number of linguistic phenomena

In order to illustrate the application of the hierarchy of phenomena, we present the list of phenomena for the sentence *Piotr daje piękne kwiaty Marysi* 'Peter gives beautiful flowers to Mary'. This sentence is annotated with the following indices:

C-Decl,
Compl-V-Valency-Two,
Compl-V-Comps-NP(acc),
Compl-V-Comps-NP(dat),
Compl-V-Subject-Nom-NP
Agreement-V-NPNom-reg,
Modification-NP-AdjP,
Adj-Cmp-degree-positive,
Agreemenet-NP-AdjP.

These annotations provide the following information: the sentence is declarative; it contains a verb with two complements which are nominal phrases, one in the accusative and the other in the dative case; the verb has a nominative subject, the finite verb and the subject follow the regular agreement rule; there is also an adjectival modifier in positive degree within a nominal phrase and an appropriate agreement rule between the noun and the adjective holds.

The incorrect sentence *Piotr daje piękny kwiaty Marysi* 'Peter gives beautiful$_{sing}$ flowers$_{pl}$ to Mary' is correlated in BRG with the correct sentence cited above. The incorrect sentence is labeled with the following indices:

$*$ C-Decl,
Compl-V-Valency-Two,
Comlp-V-Comps-NP(acc),
Compl-V-Comps-NP(dat),
Compl-V-Subject-Nom-NP
Agreement-V-NPNom-reg,
Modification-NP-AdjP,
Adj-Cmp-degree-positive
$*$ Agreemenet-NP-AdjP.

Two phenomena are violated: C-Decl (the whole sentence is incorrect) and Agreement-NP-AdjP (the agreement between *kwiaty piękny* does not hold).

The names of linguistic phenomena are organized into a hierarchy. The name of each phenomenon contains the

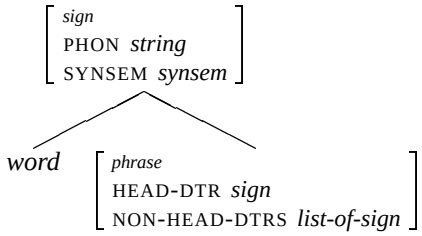

Figure 1: A part of the type hierarchy.

name of its supertype, e.g. Agreement-NP subsumes the Agreement-NP-Adj class. Hence, it is possible to refer to an entire group of phenomena just by using a prefix included in all appropriate names.

4. Annotation Schema

Sentences (as well as wordforms, see §5. below) are annotated with AVMs, as used in HPSG.[2] 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 is given in Figure 1. 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 HEAD-DTR and *list-of-sign*-valued NON-HEAD-DTRS.

Each sentence is annotated with an AVM of type *phrase*, with the orthography of the sentence represented by the value of PHON,[3] 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.

Not all attributes assumed in current HPSG are represented in the current version of the test-suite. For example, pragmatic (CONTEXT) information is ignored, while semantic (CONTENT) information is represented only provisionally. Values of some attributes are adapted to Polish, e.g., the values of the morphosyntactic attribute GENDER (Czuba and Przepiórkowski, 1995) and the hierarchy of subtypes of the *substantive* type (Przepiórkowski, 1999).

Figure 2 contains an example of a (partial) annotation for the sentence *Janek widzi Marysię*. This sentence is assumed to have the phrase structure tree as in Figure 3.

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

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

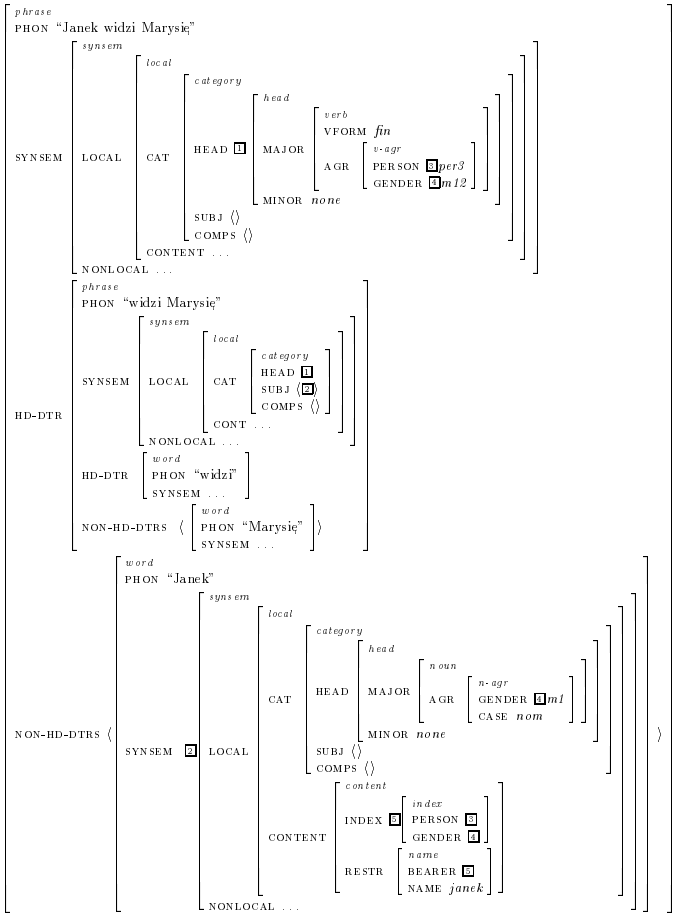

Figure 2: AVM structure of *Janek widzi Marysię*

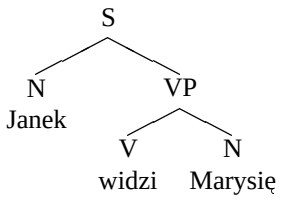

Figure 3: The phrase structure tree of *Janek widzi Marysię*

5. Implementation Issues

The HPSG test-suite for Polish is a database[4] of Polish sentences (the HPSG test-suite proper). Correct sentences are augmented with their (one or more) HPSG representations (AVM structures) constructed according to the HPSG signature. An additional, auxiliary, database contains Polish wordforms (a dictionary).

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 test-suite.

[4]This database is implemented in Delphi (Borland) in the Microsoft Windows NT environment by Wiesław Bartkowski (Bartkowski, 2000).

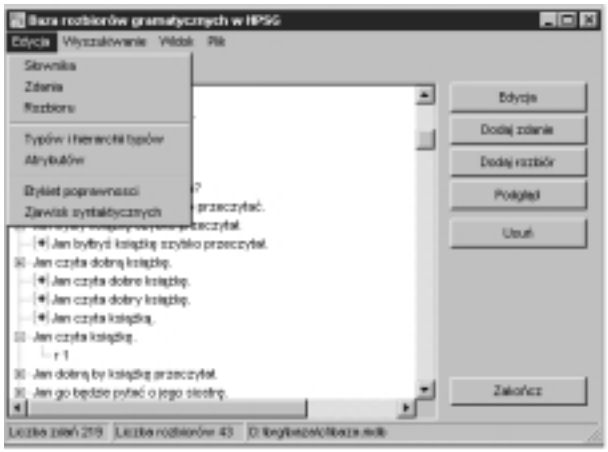

Figure 4: The main program window.

The HPSG signature is converted into a database description. This signature should be created prior to the creation of the database but some modifications of it are possible also afterwards.

The dictionary is a separate part of the database. It consists of AVM structures of inflectional forms used in sentences contained in the test-suite. 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 groups of operations on the test-suite are entering, searching and viewing data. The main program window after opening a database is shown in Figure 4.

Entering data

Sentences with phenomena annotations can be added interactively and non-interactively from a text file. Figure 5 shows the window (*Nowe Zdanie*) for entering a sentence (*Zdanie*) with the correctness marker (*Etykieta poprawności*) and the indices of linguistic phenomena (*Zjawiska syntaktyczne*). Correctness markers and names of linguistic phenomena can be entered manually or selected from the list.

Parses (AVMs) can be added only interactively via the specialized graphical interface. The interface facilitates the construction of AVMs by generating their parts semi-automatically. At every stage only a limited number of types is presented to the user. After choosing a type, a list of values of attributes appropriate for this type automatically pops up. Values of attributes must then be filled in manually, in any order. The correctness of the information entered this way is partially verified. The appropriateness of attribute values is automatically obtained. The consistent use of AVM's labels (so-called tags) is checked.

To facilitate entering of complicated structures, it is possible to view the parse tree in another window during the edition process. Figure 6 presents a window in which an AVM representing the parse of the sentence *Jan czyta książkę* is being entered.

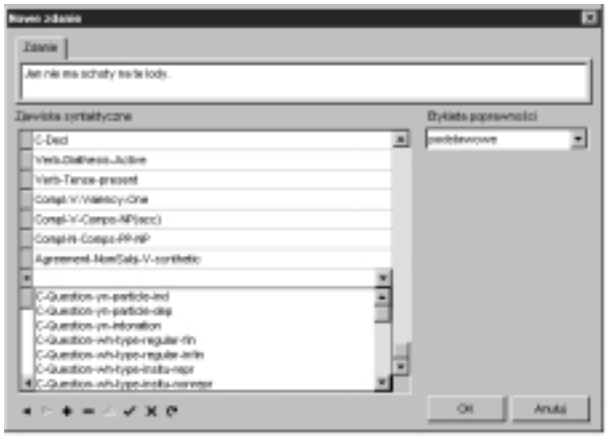

Figure 5: The entering sentences window

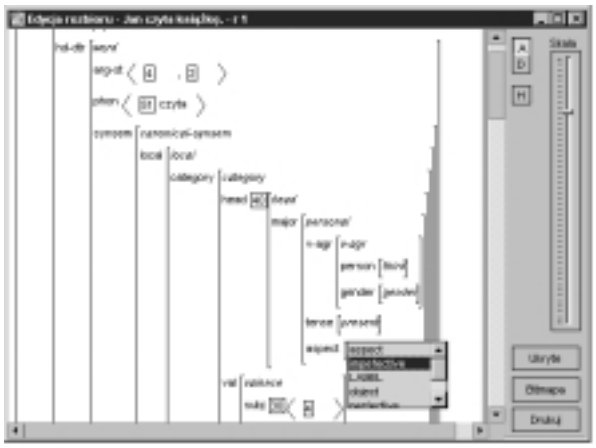

Figure 6: The entering AVMs window.

Each word (i.e., each inflectional form) not yet present in the dictionary must be entered into it before the parse is completed. Figure 7 presents a window containing the lexical entry of the word (*Słowo*) *Jana*, which is the genitive form of *Jan* and has only one description (*Opisy*) *jana1*.

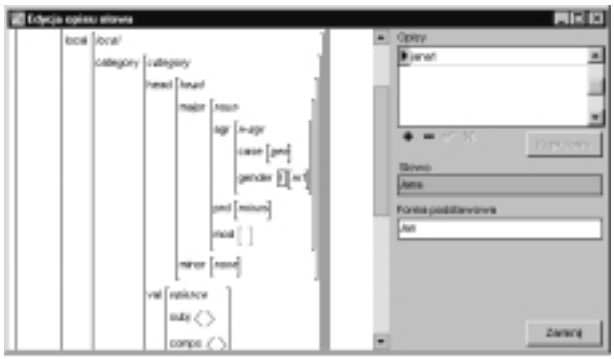

Figure 7: The lexicon window.

Figure 8: The simple query results.

Figure 9: The advance search window.

Modifying data

The data in BRG can be modified by means of the same interface which is used for entering data. Following operations are possible:

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

Search operations

There are several search operations which can be performed over the test-suite. In all cases, the search result is a list of sentences together with their parses. It is possible to output search results (sentences with indices) to a file as an ASCII text.

There are two ways of selecting data from a database. A simple query can consist of one word form, one base form of a word, one phenomena name, one correctness marker or one name of an HPSG type. In Figure 8, we present an appropriate program window together with the main window containing the result of the search.

If this simple search mechanism is not sufficient, one can formulate a regular expression query being a combining the equality and non-equality relations over word forms, phenomena names, type names and correctness markers. Figure 9 presents an exemplary question of this kind (<z> denotes a phenomenon name, <e> – a correctness marker). The result of this query is the set of sentences which contian verbs having noun phrases as their complements and any kind of negation and are annotated with the correctness marker *skomplikowane* (i.e., correct-complex).

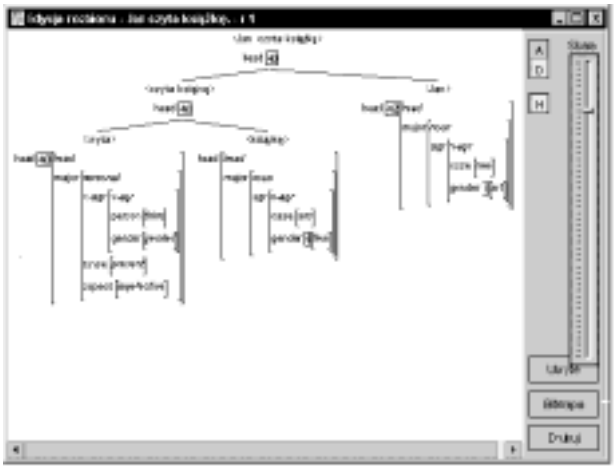

Figure 10: The tree structure.

Viewing the test-suite

The parses of a sentence can be shown on the screen in two formats simultaneously: as trees (see Figure 10) and as AVM structures (see Figure 2).

When viewing an AVM, it is possible:

- to fold and unfold substructures,
- to hide selected attributes,
- to show the structure corresponding to a tag.

6. Conclusion

Our project is the first attempt at developing tools for evaluating the coverage of formal grammars of Polish. These tools are designed specifically for HPSG grammars, but we hope that the test-suite developed here will be equally useful for the evaluation of formal grammars developed within other frameworks, e.g., the DCG-style grammars of (Szpakowicz, 1986; Świdziński, 1992) (the latter grammar was implemented in the AS parser (Bień et al., 2000)). Moreover, this project constitutes the first step towards creating a large-scale treebank for Polish.

7. References

Bańko, Miroslaw, 1990. Niektóre problemy oceny adekwatności gramatyk (na przykładzie gramatyki Szpakowicza). In *Studia Gramatyczne IX*, Prace Instytutu Języka Polskiego. Wroclaw: Zakład im. Ossolinskich.

Bartkowski, Wiesław, 2000. *Komputerowa baza analiz gramatycznych w formalizmie HPSG*. Master's thesis, Uniwersytet Warszawski, Wydział Matematyki, Informatyki i Mechaniki, Warszawa. In preparation.

Bień, Janusz S., Krzysztof Szafran, and Marcin Woliński, 2000. Experimental parsers of Polish. In *Proceedings of FDSL 3*. Lipsk. In preparation.

Carpenter, Bob, 1992. *The Logic of Typed Feature Structures*. Cambridge Tracts in Theoretical Computer Science. Cambridge University Press.

Czuba, Krzysztof and Adam Przepiórkowski, 1995. Agreement and case assignment in Polish: An attempt at a unified account. Technical Report 783, Institute of Computer Science, Polish Academy of Sciences.

Lehmann, Sabine, Stephan Oepen, Sylvie Regnier-Prost, Klaus Netter, Veronika Lux, Judith Klein, Kirsten Falkedal, Frederik Fouvry, Dominique Estival, Eva Dauphin, Hervé Compagnion, Judith Baur, Lorna Balkan, and Doug Arnold, 1996. TSNLP — test suites for natural language processing. In *Proceedings of COLING 1996*. Kopenhagen.

Marciniak, Malgorzata, Agnieszka Mykowiecka, Anna Kupść, and Maria Węgiel, 2000. Klasyfikacja zjawisk syntaktycznych na potrzeby testowego zbioru wyrażeń języka polskiego. Technical report, Institute of Computer Science, Polish Academy of Sciences. In preparation.

Oepen, Stephan, Klaus Netter, and Judith Klein, 1998. TSNLP — test suites for natural language processing. In John Nerbonne (ed.), *Linguistic Databases*, CSLI Lecture Notes. Stanford: CSLI Publications.

Pollard, Carl and Ivan A. Sag, 1994. *Head-driven Phrase Structure Grammar*. Chicago: Chicago University Press.

Przepiórkowski, Adam, 1999. *Case Assignment and the Complement-Adjunct Dichotomy: A Non-Configurational Constraint-Based Approach*. Ph.D. thesis, Universität Tübingen, Germany.

Świdziński, Marek, 1992. *Gramatyka formalna języka polskiego*, volume 349 of *Rozprawy Uniwersytetu Warszawskiego*. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.

Szpakowicz, Stanisław, 1986. *Formalny opis składniowy zdań polskich*. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.