Electronic Language Resources for POLISH: POLEX, CEGLEX and GRAMLEX

Zygmunt Vetulani
Adam Mickiewicz University
Department of Computer Linguistics and Artificial Intelligence
ul. Matejki 48/49, PL-60769 Poznań , Poland
http://main.amu.edu.pl/~vetulani
vetulani@amu.edu.pl

Abstract

We present theoretical results and resources obtained within three projects: national project POLEX, Copernicus[1] Project CEGLEX (1032) and Copernicus Project GRAMLEX (632). Morphological resources obtained within these projects contribute to fill-in the gap on the map of available electronic language resources for Polish. After a short presentation of some common methodological bases defined within the POLEX project, we proceed to present methodology and data obtained in CEGLEX and GRAMLEX projects. The intention of the Polish language part of CEGLEX was to test formats proposed by the GENELEX project against Polish data. The aim of the GRAMLEX project was to create a corpus-based morphological resources for Polish. GRAMLEX refers directly to the morphological part of the CEGLEX project. Large samples of data presented here are accessible at http://main.amu.edu.pl/~zlisi/projects.htm.

1. Introduction

It is worthwhile to recall that Polish is likely to become soon the 6th largest language of the enlarged European Union - with almost 40,000,000 speakers in Europe (after English, French, German, Italian and Spanish). There are important Polish language communities (ca. 10,000,000) outside Poland, mainly in the USA (6,000,000), in the countries which were parts of the Soviet Union (over 1,000,000), as well as in Brazil, France, Canada, and Germany. Polish is also relatively well described using traditional, human-oriented methods. Unfortunately, until the middle-nineties, electronic resources were practically not available. The situation has started to change with national and European projects resulting in a significant amount of electronic, reusable data for Polish morphology, whereas for other crucial resource categories - like corpora, bilingual resources, syntactic or semantic layer electronic dictionaries - remains still unsatisfactory.

1.1. About the Polish Language

Polish, like all other Slavonic languages, Latin and, in some respect, also Germanic languages, has a developed inflection system. Inflected words have various morphological forms carrying grammatical information formally represented in terms of inflectional or descriptive (classifying) categories. Inflectional categories are **case** and **number** for nouns, **gender**, **mood**, **number**, **person tense**, and **voice** for verbs, **case, gender**, **number** and **degree** for adjectives, **degree** alone for adverbs, etc. Examples of descriptive categories are **gender** for nouns and **aspect** for verbs.

Polish inflection is based on two main types of inflection paradigms: verbal and nominal. The verbal inflection system (called **conjugation**) is simpler than in most Romance or Germanic languages but still complex enough to precisely situate action or narration on the temporal axis. This system is completed by the category of aspect (perfective or imperfective) which is a classifying feature of Polish verbs. The second of the two main paradigms (called **declension**) is the nominal one. It is based on the case and number oppositions. The well developed declension system of Polish strongly marks Polish syntax; as the declension case endings characterise the function of the word within the sentence, therefore the word order is more free than in, e.g., Romance or Germanic languages where the position of the word in a sentence is meaningful. Main representatives of the Polish declension system are nouns (with case and number as inflection features and gender as a classifying feature), but also adjectives, numerals, pronouns and participles.

Polish inflected forms are created by combining various grammatical morphemes with stems. These morphemes are mainly prefixes and suffixes (endings). Endings are considered as the typical inflection markers and traditional classifications into inflection classes are based on ending configurations. Endings may fulfil various syntactic and semantic functions at the same time. For example, the

[1] **COPERNICUS** is the name of the program for Research and Technical Development defined by the European Union within the Fourth Framework Program (1994-1998) and dedicated to promote co-operation with CCE/NIS countries in various domains of science and technology. **POLEX** is the name of the project (Polish Government Grant no 8S50301007) started in September 1994 and finished in December 1996 under direction of Zygmunt Vetulani. **CEGLEX** is the name of one year COPERNICUS Project 1032 (March 1995-March 1996) co-ordinated by GSI-ERLI, Charanton, France (A. Ogonowski) and involving partners from the Czech Republic (Charles University, Prague, Jan Hajic) and Hungary (Lingware, Szeged, Károly Fábricz). Tasks for Polish were carried out at the Adam Mickiewicz University, Poznań, by the team headed by Zygmunt Vetulani. **GRAMLEX** was three year COPERNICUS Project 621 (April 1995-April 1998) co-ordinated by ASSTRIL, Marne-la-Vallée, France (Eric Laporte) and involving partners from Poland (Adam Mickiewicz University, team headed by Z. Vetulani), Hungary (Hungarian Academy of Sciences, Budapest, Julia Pajz; Morphologic, Budapest, Gabor Proszeki) and Italy (Consorzio Lexicon Ricerche, Salerno, Mario Monteleone).

ending of an inflected noun may express case and number. For verbs, endings may represent a person and number. Exceptionally, we may consider a prefix as inflectional morpheme (for adjectives and adverbs in superlative).

A large variety of inflectional categories for most of parts of speech is the reason why inflection paradigms are complex and long in Polish. For example, the nominal paradigme has 14 positions, the length of the verbal paradigme is 37 and the length of the adjectival one is 84.

One of the main difficulties of the Polish inflection system is not due to the size of paradigmes but to the so-called **morphonological alternations**. By this term we mean the phenomena of stem variations in different inflected forms. Although this phenomenon may be controlled by a system of formal rules, we have proposed, within POLEX, a computationally simpler solution consisting of the description of lexemes in terms of the canonical list of inflection stems (Vetulani et al., 1998b).[2]

1.2. Classical Sources and Tools

Realisation of projects of the kind described here strongly depends on the availability of valid sources. One of the major problems for modern style research on Polish is a lack of large size text corpora. (The well known frequency lists established for Polish in 70-ties were based on small 100,000 text-word corpora (Kurcz, 1974).) This is the reason for which our main sources were classical "paper" dictionaries and grammars (cf. (Szymczak, 1995); and to some extent (Doroszewski, 1958)). These tools were supplemented, within GRAMLEX, with a medium size corpus of newspaper texts. Among the main sources of grammatical information were: syntactic dictionary of Polish verbs (Polański, 1980), Orzechowska's work about inflection of nouns (Grzegorczykowa, Laskowski & Wróbel, 1984) and Tokarski's papers (1951, 1973) about the verbal paradigm. Although the first two dictionaries mentioned above were the most important morphological sources, their utility was limited because of evident lack of precision typical of dictionaries addressed exclusively to humans. On the other hand, classifications proposed by Orzechowska and Tokarski (cf. above) were imprecise and exception-based as well. An important amount of work was invested in order to make these descriptions exception-free (cf. the next section).

2. POLEX

2.1. Objectives

The main objective of POLEX was to create morphological electronic dictionaries for the core Polish vocabulary of general interest, based on a precise machine-interpretable formalism and operational coding system. We also considered as important factors human transparency and readability of the formalism. This aspect, sometimes considered as secondary from the automatic processing point of view, is however important for maintenance and further development of electronic resources (openness). At the beginning of the project there was no commercially available electronic dictionaries of that kind for Polish. Therefore, the necessary fundamental research aiming to work out human-and-machine-readable formats and reaching required precision level of linguistic (morphological) description, had to be made practically from scratch.

As traditional morphology was human-oriented, the simplicity of descriptive rules was its main concern. Within a traditional approach we have therefore a **small** number of classes (with similar ending sets within each class) but with a **high** number of exceptions (cf. (Tokarski, 1973), Orzechowska in (Grzegorczykowa, Laskowski & Wróbel, 1984)). For current language engineering applications we look rather for exception-free classification systems. One way to do this is through refinement of traditional classifications. Consistently, in POLEX we prefer to deal with unambiguous classes with low numbers of elements (including one-element classes) but still with clear criteria of class membership derived from the traditional system.

2.2. POLEX as Electronic Morphological System

By **morphological system** we mean a whole composed of:

- morphological formalism,
- dictionary of lexemes,
- package of basic software tools.

The POLEX **morphological formalism** is the descriptive tool permitting to encode the morphological properties of words. In particular, it defines the format of codes enabling generation of all inflected word forms for the given lexeme.

Dictionary of lexemes consists of dictionary units (items) including all morphological information necessary to generation and identification of word forms (in texts).

Basic software tools of POLEX allow the following operations:

- generation of all inflected forms of the word,
- identification of all inflection-relevant features of a given word form in a text,
- morphological tagging of the text

2.3. Morphological Formalism

2.3.1. Information Contents of Dictionary Units

The morphological formalism of POLEX is not supposed to describe derivational properties of words (as of secondary importance for language engineering). We limit ourselves to considering the word segmentation into **stem** and **inflection ending** and we are not concerned with any finer segmentation of stems (referring to the notions of prefix, radical, infix, suffix, postfix):

WORD_FORM=STEM + ENDING

(where the ending may be empty)

This segmentation is compatible with the traditional, human-oriented approach, where classes correspond to tables of endings, so that classification is made with respect to the way the endings are used to express functions of particular word forms.

A dictionary entry of POLEX contains three items of information, as shown below:

[2] For synthetic information about Polish cf. (Urbańczyk, 1994).

DICTIONARY_ENTRY::=
BASIC_FORM,
PARADIGMATIC_CLASS,
STEM_ALTERNATIONS.

2.3.2. Basic Forms

The **basic forms** are selected conformably to the classical convention: the infinitive for verbs, the form in singular nominative for nouns (if exists, plural nominative otherwise) etc.

2.3.3. Paradigmatic Classes

Information about **paradigmatic class** of the word should permit unambiguous calculation of exactly one ordered set of endings. We have been proceeding to refine the already existing, linguistically motivated classifications (cf. Orzechowska, in (Grzegorczykowa et. al., 1984) for nouns, (Tokarski, 1973) and (Doroszewski, 1958) for verbs). This refinement resulted in getting a number of small size classes, some of them with only one element (which is the price paid for having an exception free system).

2.3.4. Alternations - Stem Distribution

Stem alternations are described in a synthetic way despite the fact that alternations are regular in most cases and may be described by a system of morphonological rules. The synthetic approach does not generate exception-handling problems and simplifies the processing. The idea consists in listing all stems necessary for form generation. This listing refers to the ordering (called **paradigmatical**) in which word forms appear during the inflection procedure (declension, conjugation). In order to calculate the ordered set of inflection stems, the only thing we have to do is to erase endings of all generated forms. What we obtain is the so-called v**ector of stems** (i.e. set of stems ordered as they appear when inflecting the word, including repetitions). We label stems with successive natural numbers (starting with 1) according to their first occurrence in the vector of stems. The ordering of stems obtained in this way is called **canonical**. The canonical list of stems is a part of a dictionary item. It is worth noting that within our approach no "alternation theory" is needed.

2.3.5. Format of Morphological Dictionary Units

Information about stem distribution (contained in the stem distribution vector) combined with the canonical list of stems includes, in a synthetic way, information on stem alternations. Its combination with information about paradigmatic class (vector of endings) permits easy generation of all inflected forms (word forms) of the lexeme by means of elementary string operations (similar to the simple addition of vectors).

POLEX morphological entries have the following shape:

BASIC_FORM +
LIST_OF_STEMS +
PARADIGMATIC_CODE +
STEMS_DISTRIBUTION

For example, the dictionary items for *frajer*^{I} and *frajer*^{II} will be as follows:

frajer; frajer,frajerz; N110; 1:1-5,9-13; 2:6-8,14
frajer; frajer,frajerz; N110; 1:1-5,8-14; 2:6-7

In this example, 1:1-5,9-13 means that stem #1 is applied at the paradigmatic position from 1 to 5, and from 9 to 13. In some cases, pieces of information like 0:1-7 or 0:9 may appear, meaning that forms corresponding to the paradigmatic positions from 1 to 7, or 9 do not exist. This is the case of defective words. The inflection code N110 stands for a sub-class of masculine-virile, hard stem, non-velar nouns with the corresponding vector of endings :

(0,a,owi,a,em,e,e,y,ów,om,ów,ami,ach,y).

2.4. Morphonological Inflection Classes

Our approach to the description of alternation phenomena consists in considering separately two aspects: morphemic (list of stems) and combinatorial (stem distribution). When stem distributions are considered together with endings vectors we get the partition which is finer than the one based on endings only. We will consider as belonging to the same **morphonological inflection class**[3] those lexemes which belong to the same paradigmatic class (and therefore have the same vector of endings) and which have the same vector of stem distribution. The notion of morphonological inflection class has a procedural motivation: within one inflection class all inflected forms of a lexeme will be generated using the same sequence of string operations, **the same for all classes of speech.** The system of morphonological inflection classes will usually contain a great number of small classes (with a few members only). The latter phenomenon is due to the fact that our system is exception free.

Despite *a priori* deterministic nature of hidden morphonological rules governing alternation phenomena, the complexity of these rules makes that it is practically impossible to analytically calculate the number of morphonological inflection classes in Polish. On the other hand, once having the data (i.e. dictionary items) collected in the electronic form it is trivial to obtain the evaluation by simply comparing and counting different codes. With the size of the POLEX data (for more than 41,000 nouns) we claim that the observed set of morphonological inflection classes (379) is close being complete (if not just complete) for contemporary Polish. In (Vetulani et al., 1998b) we present the complete table of morphonological inflection classes observed in the POLEX lexical data.

2.5. POLEX Resources

POLEX ended on December 31, 1996 with, as its main achievement, elaboration of inflection codes for the main inflected categories, elaboration of the POLEX morphological format, and creation of the basic resource of ca 96,000 entries (incl. ca 41,450 nouns and 11,750 verbs). This draft material continues to be developed and maintained by LEX s.c.[4] Also, the basic software for form generation, lemmatisation and tagging whose prototypes were delivered within POLEX is being developed[5]. Follows a fragment of dictionary data:

bazuna; bazun,bazuni;N411;1:1,2,4,5,7-14;2:3
bazyleus; bazyleus,bazyleusi;N112;1:1-5,8-14;2:6
bazylia; bazyli;N470;1:1-14
bazylianin; bazylian;N140;1:1-14

[3] Notion first introduced by Vetulani in (Vetulani et al., 1998b)
[4] For more information please contact the Author.
[5] (Vetulani & Obrębski, 1997).

bazylika; bazylik,bazylic;N410;1:1,2,4,5,7-14;2:3

3. CEGLEX

3.1. Objectives

The main objective of CEGLEX was verification and extension of the generic electronic dictionary model developed within the EUREKA GENELEX project to three Central European languages Czech, Hungarian and Polish.

3.2. Overall Structure of the GENELEX / CEGLEX Model

GENELEX[6] was intended as a generic electronic dictionary model organised into three layers: for Morphology, Syntax and Semantics, respectively.

The conceptual model of GENELEX/CEGLEX may be expressed in terms of the *entity-attribute-relation* methodology. This conceptual model is encoded in the Standard Generalized Markup Language (SGML) as a Document Type Definition (DTD). The CEGLEX DTD defines a universe composed of *elements.* (*Entities* become *elements* in the SGML terminology.) Elements are complex (e.g., an element may be a part of another element). We may think of element definition as specifying a collection of individuals of the same type labelled by the element name. Any given type has a characteristic set of *attributes* and a *pattern* of the *is-a-part* relation. The DTD may be considered as a set of rules describing the structure of this universe. The format of these rules respects the SGML conventions. Two kinds of rules are used in this description: ELEMENT and ATTLIST.

The example given below is taken from the CEGLEX DTD for Polish morphology. In the ELEMENT rule, 'Affix_mu' is the name of the element. Also 'Graph_mu', 'Phon_mu', ... are element names. The pattern '((Graph_mu|Phon_mu)+ & Base_cat* & Derived_cat* & Derived_gender*)' informs what may be parts of the 'Affix_mu' element.

The ATTLIST rule specifies attributes associated with the element type 'Affix_mu' and characterises their values and defaults.

Example

```
<!ELEMENT Affix_mu - O ((Graph_mu|Phon_mu)+ &
    Base_cat* & Derived_cat* & Derived_gender*)>
<!ATTLIST Affix_mu
     id           ID                  #REQUIRED
     appellation  CDATA               #IMPLIED
     attestation  CDATA               #IMPLIED
     use_values   IDREF               #IMPLIED
     etymon_l     IDREFS              #IMPLIED
     affix_type   (WITHOUT_A|PREFIX|
                  SUFFIX|INFIX)       WITHOUT_A
     sem_unit_l   IDREFS              #IMPLIED>
```

[6] For more detailed characterisation of the GENELEX/CEGLEX model cf., (Vetulani, Martinek & Vetulani, 1995) and (Martinek, Vetulani & Vetulani, 1996).

Attributes may take values in different domains. These may be :

- predefined sets of items (finite or not); finite lists of items from such sets, e.g., *'affix_type'* above; in that case we specify a default value,
- identifiers of objects of a given type ; finite lists of object identifiers for objects of a given type (cf. attributes marked IDREF, or IDREFS, above),
- arbitrary, human-oriented comments (cf. attributes marked CDATA, above).

3.3. Morphological Layer

The morphological layer specifications (as defined in the DTD) makes possible putting together into one unit (corresponding to a given word) all relevant information about spelling, graphical variations, inflection, grammatical category, etc. A link to the syntactic layer is realised via the special attribute ('syn_unit_l') which point to the syntactic structures in which this given word may appear. There is no direct link to the semantic layer: such a relationship is defined via the syntactic layer.

Below we present a fragment of a morphological unit encoded in the CEGLEX format.

```
<Simple_mu id="5745"
  appellation="rzeczownik rodzaju męskoosobowego"
  autonomy="YES"
  category="NOUN"
  subcategory="COMMON">
   <Graph_mu current_nb="0"
     prefferred="YES"
     inflection="N111">
      <Spelling>student</Spelling>
      <Gstem current_nb="0">
         <Spelling>student</Spelling>
      <Gstem current_nb="1">
         <Spelling>studenci</Spelling>
      <Gstem current_nb="2">
         <Spelling>studenc</Spelling>
</Simple_mu>

<Simple_mu id="4008"
  appellation="rzeczownik rodzaju męskonieżywotny"
  autonomy="YES"
  category="NOUN"
  subcategory="COMMON">
   <Graph_mu current_nb="0"
     prefferred="YES"
     inflection="N319">
      <Spelling>dom</Spelling>
      <Gstem current_nb="0">
         <Spelling>dom</Spelling>
</Simple_mu>
```

The morphological unit may be simple (connected with a simple word), affixal (describing a part of word; in Polish they are prefixes, suffixes or infixes), compound (for complex expressions) and contracted (to record written contractions of words). Graphical and phonemic units characterize spelling and pronunciation of words. The following auxiliary entities may be used as well: etymons, graphical and phonemic inflections, inflections of compounds, combinations of morphological features.

While adapting the original GENELEX scheme (DTD) to Polish we preserved existing entities modifying at the same time some of their attributes, i.e. for some entities attribute values were changed or additional attributes were introduced. For a more detailed description of the modifications introduced to the GENELEX model within the CEGLEX experiment (at all three layers) you may consult (Vetulani, Martinek & Vetulani, 1995) and (Martinek, Vetulani & Vetulani, 1996).

It is worth observing that we use the system of inflection codes and canonical stems for characterising inflectional properties of lexemes in fundamentally the same as it was done in POLEX (cf. above).

3.4. Syntactical Layer

At the syntactic layer we describe syntactic behaviour of entities represented by morphological units. One unit describes one type of syntactic behaviour. The GENELEX/CEGLEX model appears to be particularly well suited to formalising descriptions initially realised in terms of predicate-argument structure. This claim was verified within the project on data from Polański's syntactic dictionary (Polański, 1980).

Example

<Synt_unit id="synu_braćII"
appellation="syntactic unit for 'brać' "
example="'Dramat polski brał motywy z dorobku literatury światowej'"
comment="Polański notation: NP_N --- NP_{Acc} +(z^NP_G)
use_values="use_val_stand"
description="syn_desc_braćII"
<Syn_sem_corresp target_usem="seu_braćII">
</>
</>
<Description
id=" syn_desc_braćII"
appellation="syntactic description of sentence structure for 'brać according to Polański"
example="'Motto dla swego dzieła autor wziął z Owidiusza'"
comment="Polański notation: NP_N --- NP_{Acc} +(z^NP_G) no restrictions on self"
representing_mu="brać wziąć"
self="self_head_sentence_1"
construction="constr_braćII">
</>

3.5. Semantic Layer

Within the GENELEX/CEGLEX approach, semantics of words may be described at the two representation levels: linguistic and conceptual.

A semantic unit represents the meaning of a word in some syntactic context. A description of the semantic unit is done in accordance with two main axes: componential (analytical) and relational (differential). The componential axis breaks down the meaning into elementary components (e.g. semantic features). This description is conceptual and may imply many levels of abstraction. The relational axis defines relations between semantic units (such as semantic derivation, collocation or preference) and abstract relations between concepts or predicates.

The CEGLEX semantic description of **verbs** is based on a dictionary of Polish verbs (Polański, 1980). For one meaning of the verb one semantic unit with the unique identifier is created. The predicative representation of the meaning includes semantic requirements of the predicate, description of predicate arguments and their roles. Semantic units may be inter-related by the synonymy relation. Units may be completed by semantic features, e.g. the semantic type may be a state, transition or process.

Description of **noun** is based on the Szymczak's dictionary (Szymczak, 1995). For nouns we apply relations such as generalisation (vs. particularisation), meronymy, synonymy or antonymy. Relations point to verbs and describe the semantic role of the nouns with respect to verbs. For example the noun "pencil" may have the semantic role "instrument" for the verb "to write".

Example

<Sem_unit id = „seu_braćII"
appellation =
„(ktoś|coś)(bierze|czerpie) coś z czegoś"
example =
„Dramat polski brał motywy z dorobku literatury światowej"
wval_sem_feature_l = „wvf_aspect_imperfect">
<Predicative_rep
prefered = „YES"
predicate = „pred_braćII">
</>
</>

<Predicate
id = „pred_braćII"
type = „LEXICAL"
wval_sem_feature_l = „wvf_aspect_imperfect"
argument_l =
„arg_agent_hum_plusORclass_information
arg_patient_abstr_plus
arg_source_poss_abstr_plus"></>

3.6. CEGLEX Deliverables

The CEGLEX project produced GENELEX compatible formats for dictionary data in the form of the DTD (Document Type Definition) according to the SGML norm. A large part of the original GENELEX model was verified trough a selection of vocabulary of ca. 2800 words (including ca. 380 compounds). This vocabulary came from two sources: experimentally collected[7] corpora "Robot" and "Casino" of written road descriptions (863 items) and the set of words representing semantic primitives used by Polański in his dictionary (Polański, 1980) (ca 1940). The categories considered were: nouns (1858), compound nouns (379), adjectives (315), pronouns (21), verbs (110), participles (28), numerals (31), adverbs (13), conjunctions (19), prepositions (31),

[7] Within a Joint French-Polish Project (*Accès en langage naturel aux bases de connaissances spatiales*) involving UAM (Z. Vetulani) and LIMSI/CNRS (Orsay, France, G. Ligozat) (cf. (Marciniak & Vetulani, 1999)).

particles (15). For all these words full, 3 layer descriptions were supplied.

The CEGLEX dictionary was tested within an application compiling this dictionary into a form readable by a Polish language parser capable of analysing sentences. The parser used to verify CEGLEX data was extracted from POLINT, a natural language interface designed by Vetulani (cf. (Vetulani 1997)).

4. GRAMLEX

4.1. Objectives

GRAMLEX was intended to design and to produce data and tools (algorithms) needed to achieve lexical tagging of texts for four European languages (Hungarian, French, Italian and Polish). Its Polish part started effectively after the end of the CEGLEX project, so that some of its results, concerning morphological layer, could have been included. (More about GRAMLEX in (Vetulani et al., 1997, 1998a).)

4.2. Methodology

Several elements of GRAMLEX methodology came from our former research or from project partners' experience, especially of the LADL and ASSTRIL laboratories. The most important were the use of finite state automata and the corpus-based method of vocabulary selection. A corpus-oriented approach allows the coverage of important areas of applications with middle size dictionaries, depending of course on the appropriate corpus choice. Application of finite state automata is particularly important for storing the dictionary of forms (as the inflection factor for large-scale Polish dictionaries is greater than 16). The finite automata technology is space and time effective (if appropriately used) but not transparent at all. Therefore we developed a more human-oriented (but still directly machine-readable) format to store the generic dictionary of lexemes. This format was developed in the SGML notation and was directly inspired by the results of the projects CEGLEX. The main idea inherited from POLEX is that the inflection of words is described in terms of inflection stems and endings in a uniform way for all inflected categories which is a good solution for "alternation problems" and facilitates further processing.

4.3. Dictionary Data

4.3.1. Vocabulary Acquisition

In the situation of a lack of large scale publicly or commercially available corpora for Polish, we decided to apply a mixed approach based on both the corpus investigation and dictionary exploration.

The vocabulary of GRAMLEX consists of 4 parts. These are:

1) nouns (754), adjectives (176) and verbs (422) from the "Słownik minimum języka polskiego" (Kurzowa & Zgółkowa, 1993),

2) 505 nouns and 283 adjectives which come from CEGLEX (small corpora ROBOT and CASINO, cf. above),

3) a large part (8591) of the frequency list published by IJP PAN[8] (Kurcz et al., 1974),

4) the remaining vocabulary obtained from the corpus of articles from two regional daily newspapers "Dziennik Bałtycki" (Gdańsk, 1995) and "Głos Wielkopolski" (Poznań, 1996).

The size of the simple word list of GRAMLEX project amounts to ca 22,500 words (i.e. about 25,000 entries /lexemes/), almost all of them coming from a corpus or from text generating experiment. The exception to this principle is the inclusion of some nouns, adjectives and verbs (1352 entries) from the dictionary of basic words. This exception to the general principle of direct confrontation with a corpus was motivated by the will to avoid the case where some of the very basic words would be incidentally omitted, because of their absence in the considered corpora. The GRAMLEX dictionary of simple words is to be considered as a middle size dictionary whose lexical contents corresponds to the core (passive) linguistic competence of a native speaker.

Besides the general core GRAMLEX dictionary some small size dictionaries of special interest were produced within GRAMLEX. The first of them is the dictionary of compound terms (2500), most of them (2000) extracted from the GRAMLEX corpus with the help of the compound-term-acquisition-program called EXTRACT. Another dictionary of both simple and compound terms was concerned with the technical terminology of one chosen domain. Four sources (books) concerning mobile telephony and telecommunication networks were searched for terminology. This operation resulted with selection of ca 870 terms (of which ca 470 compounds and 320 acronyms).

Compound terms from these two dictionaries were the empirical material for the development of the appropriate format for morphological description of compounds. Two versions of such format were proposed and discussed within GRAMLEX (cf. (Vetulani et al., 1998a)).

4.3.2. Formats

GRAMLEX formats for generic dictionary data (for simple and compound words) are a variation of CEGLEX formats for the morphological layer. This format, called GRAMCODE, was tested for compound telecommunication terms acquired in the project.

Independently of presentation of generic dictionary in the SGML based format GRAMCODE, the dictionary of word forms generated from the GRAMLEX dictionary was delivered in the form of finite state automaton.

4.3.2. Coverage

By **morphological coverage** we mean the distribution of the lexicon into various morphological categories (as *nouns, verbs, ...*). The morphological coverage of GRAMLEX is almost complete. The only important category not considered in GRAMLEX is that of numerals. The corpus behind the dictionary amounts to ca 430,000 text words. The GRAMLEX morphological

[8] IJP PAN - Instytut Języka Polskiego Polskiej Akademii Nauk (Institute of Polish Language of the Polish Academy of Science)

coverage is characterised by the following table (state at the end of the project, april 1998).

MORPHOLOGICAL COVERAGE		
CATEGORY	**NB OF ENTRIES**	**EXAMPLE**
words	22,138	
lexemes	24,679	
ADJ	3897	biały
ADJPAP	1461	robiony
ADJPP	35	skostniały
ADJPRO	37	taki
ADJPRP	775	robiący
ADV	802	mało
ADVANP	22	zrobiwszy
ADVPRO	38	gdzie
ADVPRP	379	robiąc
APP	10	hura
BYC	1	być
CONJ	89	i
EXCL	19	dość
N	12786	kot
NPRO	35	on
ONO	7	miau
P	97	do
PART	75	nie
PPRO	3	doń
V	4075	biec
VM	33	można
VNI	2	wiadomo

ADJ-adjective; ADJPAP-adjectival-passive-participle; ADJPP-adjectival-past-participle; ADJPRO-adjectival pronoun; ADJPRP-adjectival-present-participle; ADV-adverb; ADVANP-adverbial-anterior-participle; ADVPRO-adverbial-pronoun; ADVPRP-adverbial-present-participle; APP-call; BYC-to-be; CONJ-conjunction; EXCL-exclamation (interjection); N-noun; NPRO-nominal-pronoun; ONO-onomatopoeia; P-preposition; PART-particle; PPRO-nominal-prepositional-pronoun; V-verb; VM-modal-verb; VNI-non-inflected-verb

4.3.3. Examples

As the format of entries for simple words is similar to that used within CEGLEX (cf. example above), we provide here only one example of encoded telecommunication compound term.

```
<Entry id="p177"
   type="compound"
   index="ind1">
      <Spelling>dynamiczny przydział kanału</>
      <Structure
        infl-code="Adj+N+N/g"
        category="noun-phrase"
        constraints="i+i+u">
            <Component nr=1
                  simple-entry-ref="g2-85745">
            <Component nr=2
                  simple-entry-ref="g2-4008">
            <Component nr=3
                  simple-entry-ref="g2-1473">
                  <Semantics
                        domain="GSM"
                        english="dynamic channel allocation">
            </>
</>
…

<Entry id="g2-1473"
   type="simple"
   index="ind1">
      <Spelling> kanał</>
      <Morphology
            category="noun"
            gender="masculine-inanimate"
            inflection="i-N310-3">
              <Stems>kanału kanał </>
      </>
</>
...
```

4.4. Tools

Production of prototypes of basic language engineering tools was among the main objectives of GRAMLEX. Three main tools produced in the project were GENFORM, LEXAN and SCON. All of them are implemented in C under LINUX.

4.4.1. GENFORM

GENFORM is a generator of inflected forms for simple and compound lexemes. Its algorithm is straightforward, due to very simple way of inflection description applied in GRAMLEX (described in terms of stem distribution vector and vector of endings). It is worth observing that the algorithm of GENFORM remains the same for all categories.

4.4.2. LEXAN

LEXAN is a dictionary-based lemmatizer and morphological tagger for Polish texts. The lexeme identifier calculated while lemmatizing the word form, together with the morphological information about the word form, constitute a structured tag used by LEXAN to mark the word form in the text. The program looks for all possible solutions and produces an output which may sometimes be multiple. LEXAN may be used alone, as a simple lemmatizer/tagger for a text (or a list of words in particular). It may also be component of more sophisticated software as, e.g., selective taggers used to mark properly specified elements only, concordance generators etc.

4.4.3. SCON

An important progress in the text processing of Polish was made due to the SCON program to built structural concordance tables. The term "structural" means here the possibility to find concordances defined by complex structures specified using a **grammatical pattern.** As LEXAN is an integral part of the SCON program it is possible to use the lexeme identifier with appropriate morphological constraints (morphological attribute values as parameters) in order to specify the required combination of inflected forms. Some examples of admissible patterns follow.

Complex pattern	Matching terms
/poczta_/polski	Poczta Polska, Poczty Polskiej
mieć//Ns+<B>+//N	ma kaca, miał wiadomość
//V_deszcz	pada deszcz, idzie deszcz
<N>+<S>+<N>	2:1, 1,23, 10-20
<N>+<W>	1200ccm, 1998r

4.4.4. Examples of applications

The GRAMLEX project tools we have been talking about above, were verified in a number of applications within the GRAMLEX project itself. In particular, programs for: structure analysis of dictionary entries (VERBAN), interactive analysis of dictionary definitions (NOUNDAN) and acquisition of terminology from dictionary definitions (NOUNAN) require LEXAN pre-processed texts as input. These programs were implemented in Prolog by Martinek (cf. (Vetulani et al., 1998a) and (Martinek, Obrębski & Vetulani, 2000)).

5. Acknowledgements

The author wishes to thank his closest collaborators within all projects presented here: Jacek Martinek, Tomasz Obrębski, Grażyna Vetulani, Bogdan Walczak are co-authors of the most of project achievments. He also thanks for fruitful co-operation within the Copernicus Projects CEGLEX and GRAMLEX all colleagues from the collaborating teams, and first of all Károly Fábricz, Eva Hajičova, Jan Hajič, Ferenc Kiefer, Eric Laporte, Mario Monteleone, Antoni Ogonowski, Julia Pajz, Gabor Proszeki. Special thanks are due to Poul Andersen from DGXIII who assisted us as Project Officer with help and advices.

References

Doroszewski, W. (1958). Słownik Języka Polskiego (The Dictionary of Polish Language), PWN, Warszawa.

Kurcz, I., Lewicki, A., Sambor, J., Woronczak, J. (1974). Słownictwo współczesnego języka polskiego, listy frekwencyjne, tom II: drobne wiadomości prasowe (Vocabulary of contemporary Polish, frequency lists, vol II: short press reports), IJP PAN, Warszawa.

Marciniak, J. & Vetulani, Z. (1999). Ontological problems related to construction of natural language interface for a mobile robot. In H. Guesgen (Ed.), Workshop on Hot Topics in Spatial and Temporal Reasoning, IJCAI'99 (proc.), Stockholm, pp. 31--36.

Martinek, J., Obrębski, T., Vetulani, Z. (2000). Dictionary-based tools for linguistic data acquisition from texts. In: Lewandowska-Tomaszczyk, J. Melia, PALC99 - Practical Applications in Language Corpora, Peter Lang GmBH, Frankfurt (in print).

Martinek, J., Vetulani, G. & Vetulani, Z. (1996). A description of Lexical Knowledge for Polish within the Genelex Model. In: K. Sroka (Ed.) Kognitive Aspekte der Sprache, Max Niemeyer Verlag, Tübingen, pp. 175--180.

Polański, K. (1980-1990,1992). Słownik Syntaktyczno-Generatywny Czasowników Polskich, t. I-IV, Ossolineum PWN, Wrocław-Warszawa-Kraków-Gdańsk, 1980-1990, t. V, Instytut Języka Polskiego PAN, Kraków, 1992.

Szymczak, M. (ed.) (1995 and 1997). Słownik języka polskiego PWN (PWN Dictionary of Polish), paper edition 1995, electronic edition 1997.

Tokarski, J. (1951). Czasowniki polskie. Formy , typy, wyjątki, słownik. PWN. Warszawa.

Tokarski, J. (1973). Fleksja polska. PWN. Warszawa.

Urbańczyk, St. (Ed.) (1994). Encyklopedia języka polskiego, Ossolineum, Kraków.

Vetulani, Z. (1997) A system for Computer Understanding of Texts. In: Murawski, R. & Pogonowski, J. (Eds.), Euphony and Logos (Poznań Studies in the Philosophy of the Sciences and the Humanities, vol. 57) Rodopi, Amsterdam-Atlanta, pp. 387--416.

Vetulani, Z. & Obrębski, T. (1997). Morphological tagging of texts using the lemmatizer of the 'POLEX' electronic dictionary. In: Lewandowska-Tomaszczyk, B. & Melia, P. J. (Eds.) Practical Applications in Language Corpora, Proceedings, Łódz: Łódz University Press, pp. 496--505.

Vetulani, Z., Martinek, J. & Vetulani, G. (1995). The CEGLEX dictionary model for Polish, Proceedings of the 4th and 5th International Conferences UKRSOFT (Lviv, 1994, 1995), SP «BaK», Lviv, pp. 144--150.

Vetulani, Z., Martinek. J, Obrębski, T., & Vetulani, G. (1997). Lexical Resources and Tools for Tagging Polish Texts within GRAMLEX. Investigationes Linguisticae, XXI:2, pp. 401--416.

Vetulani, Z., Martinek. J, Obrębski, T., & Vetulani, G. (1998). Dictionary Based Methods and Tools for Language Engineering, Poznań: Adam Mickiewicz University Press.

Vetulani, Z., Walczak, B., Obrębski, T., & Vetulani, G. (1998). Unambiguous coding of the inflection of Polish nouns and its application in the electronic dictionaries - format POLEX / Jednoznaczne kodowanie fleksji rzeczownika polskiego i jego zastosowanie w słownikach elektronicznych - format POLEX, Poznań: Adam Mickiewicz University Press.

Zgółkowa, H. (1993). Słownik minimum języka polskiego (The Dictionary-minimum of Polish Language), Kantor Wydawniczy SAWW, Poznań.