Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation, pages 51–59
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Polish Lexicon-Grammar Development Methodology as an Example for Application to other Languages

Zygmunt Vetulani[1], Grażyna Vetulani[2]

[1,2] Adam Mickiewicz University in Poznań
[1] Faculty of Mathematics and Computer Science
[1] ul. Uniwersytetu Poznańskiego 4, 61-614, Poznań, Poland
[2] Faculty of Modern Languages and Literatures
[2] al. Niepodległości 4, 61-874, Poznań, Poland
{vetulani, gravet}@amu.edu.pl

Abstract

In the paper we present our methodology with the intention to propose it as a reference for creating lexicon-grammars. We share our long-term experience gained during research projects (past and on-going) concerning the description of Polish using this approach. The above-mentioned methodology, linking semantics and syntax, has revealed useful for various IT applications. Among other, we address this paper to researchers working on "less" or "middle-resourced" Indo-European languages as a proposal of a long term academic cooperation in the field. We believe that the confrontation of our lexicon-grammar methodology with other languages – Indo-European, but also Non-Indo-European languages of India, Ugro-Finish or Turkic languages in Eurasia – will allow for better understanding of the level of versatility of our approach and, last but not least, will create opportunities to intensify comparative studies. The reason of presenting some our works on language resources within the Wildre workshop is the intention not only to take up the challenge thrown down in the CFP of this workshop which is: "To provide opportunity for researchers from India to collaborate with researchers from other parts of the world", but also to generalize this challenge to other languages.

Keywords: language resources, lexicon-grammar, wordnet, Indian languages, non-Indoeuropean languages

1. Introduction

In the linguistic tradition a crucial role in language description was typically given to dictionaries and grammars. The oldest preserved dictionaries were in form of cuneiform tablets with Sumerian-Akkadian word-pairs and are dated 2300 BC. Grammars are "younger". Among the first were grammars for Sanskrit attributed to Yaska (6th century BC) and Pāṇini (6-5th century BC). In Europe the oldest known grammars and dictionaries date from the Hellenic period. The first one was *Art of Grammar* by Dyonisus Thrax (170-90 BCE), in use in Greek schools still some 1,500 years later. Until recently, these tools were used for the same purposes as before - teaching and translation, and *ipso facto* were supposed to be interpreted by humans. The formal rigor was considered of secondary importance. The situation changed recently with development of computer-based information technologies. For machine language processing (as machine translation, text and speech understanding, etc.) it appeared crucial to adapt language description methodology to the technology-imposed needs of precision. Being human-readable was not enough, new technological age required from grammars and dictionaries to become machine-readable. New concepts of organization of language description for better facing technological challenges emerged. One among them was the concept of lexicon-grammar.

This paper addresses two cases. First – languages with a rich linguistic tradition and valuable preexisting language resources, for which the methods described in this paper will be easily applicable and may bring interesting results. Among Indian languages this will be the case of Sanskrit, Hindi and many other. On the other hand, a multitude of languages in use on the Indian subcontinent do not dispose of such a privileged starting position. In this case, in order to benefit from the methodology we describe in this paper, an effort must first be done to complete existing gaps. This is a hard work, and the paper, we hope will give some idea on the priorities on this way. Still, an important basic research effort will be necessary[1].

2. Why Lexicon-Grammars?

Development of computational linguistics and resulting language technologies made possible passage from the fundamental research to the development of real-scale applications. At this stage availability of rigorous, exhaustive and easy to implement language models and descriptions appeared necessary. The concept of lexicon-grammar answers to these needs. Its main idea is to link an important amount of grammatical (syntactic and semantic) information directly to respective words. Within this approach, it is natural to keep syntactic and semantic information stored as a part of lexicon entries together with other kinds of information (e.g. pragmatic). This principle applies first of all to verbs, but also to other words which "open" syntactic positions in a sentence, as e.g. certain nouns, adjectives and adverbs. Within this approach, we include into the lexicon-grammar all predicative words (i.e. words that represent the predicate in the sentence and which open the corresponding argument positions).

[1] We do not believe that basic linguistic research is avoidable on the base of technological solutions only. (See the historical statement addressed by Euclid of Alexandria (365 BC – 270 BC) to Ptolemy I (367 BC – 282 BC): "Sir, there is no royal road to geometry".)

The idea of lexicon-grammar is to link predicative words with possibly complete grammatical information related to these words. It was first systematically explored by Maurice Gross (Gross 1975, 1994), initially for French, then for other languages. Gross was also – to the best of our knowledge – the first to use the term lexicon-grammar (fr. *lexique-grammaire*)).

3. GENELEX project (1990-1994)

The **EUREKA** GENELEX[2] was a European initiative to realize the idea of lexicon-grammar in form of a generic model for lexicons and to propose software tools for lexicons management (Antoni-Lay et al., 1994). Anoni-Lay presents two reasons to build large-size lexicons as follows. "The first reason is that Natural Language applications keep on moving from research environments to the real world of practical applications. Since real world applications invariably require larger linguistic coverage, the number of entries in electronic dictionaries inevitably increases. The second reason lies in the tendency to insert an increasing amount of linguistic information into a lexicon. (…) In the eighties, new attempts were made with an emphasis on grammars, but an engineering problem arose: how to manage a huge set of more or less interdependent rules. The recent tendency is to organize the rules independently, to call them syntactic/semantic properties, and to store this information in the lexicon. A great part of the grammatical knowledge is put in the lexicon (…). This leads to systems with fewer rules and more complex lexicons." (ibid.).

The genericity of the GENELEX model is assured by:

- "theory welcoming", what means openness of the GENELEX formalism to various linguistic theories (respecting the principle that its practical application will refer to some, well defined linguistic theories as a basis of the lexicographer's research workshop). It should allow encoding phenomena described in different ways by different theories[3];
- possibility to generate various application-oriented lexicons;
- capacity of generation of lexicons apt to serve applications demanding a huge linguistic coverage.

The second important property of GENELEX besides genericity was the requirement of high precision and clarity of GENELEX-compatible lexicon-grammars.

GENELEX was first dedicated to a number of West-European languages, among other French, English, German, Italian. Although Polish[4] was not directly addressed by GENELEX, it was covered together with Czech and Hungarian by two EU projects (COPERNICUS projects CEGLEX – COPERNICUS 1032 (1995-1996) and GRAMLEX – COPERNICUS 621 (1995-1998))[5] whose objective was testing the potential of the extension of the novel GENELEX-based LT solutions to highly inflectional (as Polish) and agglutinative (as Hungarian) languages. Positive results obtained within this project demonstrated potential usefulness of the lexicon-grammar approach for so far less-resourced languages, Indo-European or not. In particular, the case of Polish demonstrated the need to take into account, within the lexicon-grammar approach, the specificity of highly inflected languages, like Lain or Sanskrit, with complex verbal and nominal morphology.

4. Lexicon-Grammar of Polish

Already in our early works on question-understanding-and-answering systems (Vetulani, Z. 1988, 1997) we capitalized the advantages of the lexicon-grammar approach. In addition to information typically provided in

[2] GENELEX was followed by several other EU projects, such as LE-PAROLE (1996-1998), LE-SIMPLE (1998-2000) and GRAAL (1992-1996).

[3] The GENELEX creators make a clear distinction between independence with respect to language theory, and the necessity for any particular application to be covered by some language theory compatible with the GENELEX model (this is in order to organize correctly the lexicographer's work).

[4] Polish, like all other Slavic languages, Latin and, in some respect, also Germanic languages, has a developed inflection system. 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. 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. The second of the two main paradigms (called declension) is the nominal one. It is based on the case and number oppositions. The declension system of Polish strongly marks Polish syntax; as the declension case endings characterize 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, 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. 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 paradigm has 14 positions, the length of the verbal paradigm is 37 and the length of the adjectival one is 84 (Vetulani, G. 2000).

[5] Some of the outcomes of these project are described in (Vetulani, G. 2000).

dictionaries we managed to explore structural, as well as morpho-syntactic-and-semantic information directly stored with predicative words, i.e. words which are surface manifestation of sentence predicates. In Polish, as in many (all?) Indo-European languages, these are typically verbs, but also nouns, adjectives, participles and adverbs. The content of lexicon-grammar entries informs about the structure of minimal complete elementary sentences supported by the predictive words, both simple and compound. This information may be precious in order to substantially speed-up sentence processing[6] (see e.g. Vetulani, Z. 1997). Taking this into account, the text processing stage requires a new kind of language resource which is electronic lexicon-grammar. In opposition to small text processing demo systems developed so far, this requirement appears demanding when starting to build real size applications within the concept of predicate-argument approach to syntax of elementary sentences that we applied in our rule-based text analyzers and generators. The rule-based approach dominating still at the turn of the centuries remains important in all cases where high processing precision is essential.

Concerning digital language resources Polish was clearly under-resourced at those days, however with a good starting position due to well-developed traditional language descriptions. For example, since 1990s the high quality lexicon-grammar in the form of Generative Syntactic Dictionary of Polish Verbs (Polański 1980-1982) was to our disposal. This impressive resource of 7,000 most widely used Polish simple verbs, being addressed first of all to human users, was hardly computer-readable. As simplified example of an entry we propose the description of the polysemic predicative verb POLECIEĆ (meaning *to fly*). One of its meanings is represented by the following entry (lines a – d):

(a) POLECIEĆ (English: FLY)[7]
(b) $NP_{Nominative}$+NPI+($NP_{Ablative}$)+($NP_{Adlative}$)
(c) $NP_{Nominative}$ [human]; $NP_{Instrumental}$ [flying object]; $NP_{Ablative}$ [location]; $NP_{Adlative}$ [location]
(d) Examples:
..., Ja(NP_{N}) z Warszawy (NP_{Abl}) do Francji (NP_{Adl}) POLECĘ samolotem (NPI),...
..., I (NP_{N}) WILL FLY from Warsaw(NP_{Abl}) to France(NP_{Adl}) by plane(NP_{I})),... ,
where:

(a) is the entry identifier (verb in infinitive)
(b) is the *sentential scheme* showing the syntactic structure and syntactic requirements of the verb with respect to obligatory and facultative (in brackets) arguments (it may be considered as a simple sentence pattern)
(c) is the specification of semantic requirements of the verb for obligatory and facultative arguments (ontology concepts in brackets)
(d) provides some use examples

The formalism ignores details of the surface realization of meaning, such as case, gender, number, etc. of words. The pioneering and revelatory work of Polański was limited to simple verbs but both method and formalism perfectly support compound constructions. What follows is an example of an entry for a verb-noun collocation composed of a predicatively empty *support verb (light verb* in the terminology used by Fillmore (2002) together with a predicative noun which plays the function of compound verb in the sentence *Orliński and Kubiak odbyli lot z Warszawy do Tokio samolotem in a Breguet 19 w roku 1926 (In 1926, Oliński flew/made a flight from Warsaw to Tokyo in a Breguet 19).*

The dictionary entry for ODBYĆ LOT in the above format will be:
(a') ODBYĆ LOT (English: FLY)[8]
(b') NP_N +NP_I+(NP_{Abl})+(NP_{Adl})+(DATE)
(c') NP_N [human]; NP_I [flying object]; NP_{Abl} [location]; NP_{Adl} [location]; DATE [year].

Information contained in lexicon-grammar entries appeared very useful in various NLP tasks. For example, an important part of information useful for simple sentence understanding may be easily accessed through basic forms of words identified in the sentence. Parts (b) and (c) of the dictionary entries for the identified predicative word will help to make precise hypotheses[9] about the syntactic-semantic pattern of the sentence.

Despite their merits, the traditional syntactic lexicons, as is the above presented Syntactic Generative Dictionary, are not sufficient to supply all necessary linguistic information to solve all language processing problems. The case of highly inflected Polish (but also other Slavonic languages, Latin, German etc.) demonstrates the need of precise and complete description of morphology. For Polish we delivered within the project POLEX (1994-1996) a large

[6] E.g. in heuristic parsing in order to limit the grammar search space explored by the parser (Vetulani, Z. 1997).
[7] "We do not claim that the set of semantic features we propose is exhaustive and final. Besides features commonly accepted we considered necessary to introduce such distinction words as nouns designing plants, elements, information etc.", cf. (Polański 1992).
[8] "We do not claim that the set of semantic features we propose is exhaustive and final. Besides features commonly accepted we considered necessary to introduce such distinction words as nouns designing plants, elements, information etc.", cf. (Polański 1992).
[9] The concept of *syntactic hypothesis* is crucial for our methods of heuristic parsing making a right choice of hypothesis about the sentence structure may considerably reduce the parsing cost (in time and space). With good heuristics, in some cases it is possible to reduce the grammatical search space considerably and as a result turning the nondeterministic parser into a *de facto* deterministic one. We explored this idea with very good effects in our rule-based question-answering systems POLINT (see e.g. section Preanalysis in (Vetulani, Z. 1997).

electronic dictionary (Vetulani, Z. et al. 1998 ; Vetulani, Z. 2000) of over 120,000 entries.[10] This resource is easily machine treatable and was used as Polish Lexicon-Grammar complement.

5. Citing « PolNet – Polish Wordnet » as lexical ontology

Within our real-size application projects[11] we extensively used a lexical ontology to represent meaning of text messages. Absence on the market of ontologies reflecting the world conceptualization typical of Polish speakers pushed us to build from scratch PolNet – Polish Wordnet, lexical database of the type of Princeton WordNet[12]. In Princeton WordNet like systems basic entities are classes of synonyms (synsets) related by some relations of which the most important are hyponymy and hyperonymy. Synsets may be considered as ontology concepts with the advantage of being direcly linked to words.

We started the PolNet project in 2006[13] at the Department of Computer Linguistics and Artificial Intelligence of Adam Mickiewicz University and its progress continues. The resource development procedure was based on the exploration of good traditional dictionaries of Polish and the use of available language corpora (e.g. IPI PAN Corpus; cf. Przepiórkowski, 2004) in order to select the most important vocabulary, for the purpose of the application expanded with the application specific terminology[14]. Development of PolNet was organized in an incremental way, starting with general and frequently used vocabulary[15]. By 2008, the initial PolNet version based on noun synsets related by hyponymy/hyperonymy relations was already rich enough to serve as core lexical ontology for real-size application developed in the project (POLINT-112-SMS system cf. Vetulani, Z. et al. 2010). Further extension with verbs and collocations, operated after the 2009, contributed to transform PolNet into a lexicon-grammar intended to ease implementation of AI systems with natural language competence and other NLP-related tasks.

```
<SYNSET>
<ID>PL_PK-518264818</ID>
<POS>n</POS>
<DEF>instytucja zajmująca się kształceniem; educational institution </DEF>
<SYNONYM>
<LITERAL lnote="U1" sense="1">szkoła</LITERAL>     % szkoła=school
<LITERAL lnote="U1" sense="5">buda</LITERAL>
<LITERAL lnote="U1" sense="1">szkółka</LITERAL>
.....
</SYNONYM>
<USAGE>Skończyć szkołę</USAGE>
<USAGE>Kierownik szkoły</USAGE>
.....
<ILR type="hypernym" link="POL-2141701467">instytucja oświatowa:1</ILR>
<RILR type="hypernym" link="POL-2141575802">uczelnia:1,szkoła wyższa:1,wszechnica:1</RILR>
<RILR type="hypernym" link="POL-2141603029">szkoła średnia:1</RILR
.....
<STAMP>Weronika 2007-07-15 12:07:38</STAMP>
<CREATED>Weronika 2007-07-15 12:07:38</CREATED>
</SYNSET>
```

Fig. 1. The PolNet v.0.1 entry for (school *szkoła*) (the sysnset{szkoła:1,buda:5, szkółka:1,….}; indices 1, 5, … refer to the particular sense of the word *szkoła*) (Vetulani, Z. 2012)

[10] POLEX dictionary is distributed through ELDA (www.elda.fr) under ISLRN 147-211-031-223-4.

[11] For detailed description of language resources and tools used to develop POLINT-112-SMS system (2006-2010) and the specification of its language competence see (Vetulani, Z. et al., 2010).

[12] Princeton WordNet (Miller et al., 1990) was (and continue to be) widely used as a formal ontology to design and implement systems with language understanding functionality. In order to respect specific Polish conceptualization of world, we decided to build PolNet from scratch rather than merely translate Princeton WordNet into Polish. Building from scratch is more costly, but the reward we get in return was an ontology well corresponding to the conceptualization reflected in the language. We do not recommend "translation-based" construction of a wordnet for languages socio-culturally remote with respect to the source wordnet language, in particular for language pairs spoken by socio-culturally different communities.

[13] Another large wordnet-like lexical database for Polish started at about the same time at the Technical University in Wrocław (Piasecki et al. 2009). It was however based on different methodological approach.

[14] Lack of appropriate terminological dictionaries forced us to collect experimental corpora and extract missing terminology manually (Walkowska, 2009; Vetulani, Z. et al. 2010).

[15] See (Vetulani, Z. et al., 2007) for the PolNet development algorithm.

6. From PolNet to a Lexicon-Grammar for Polish

6.1 First step – simple verbs

Integration of the lexicon-grammar approach to syntax with the word-based approach to ontology was the idea behind the evolution from the PolNet 1.0 (2011) to the PolNet 3.0 (and further). This idea was implemented through expansion of the initial PolNet of nouns with other parts of speech, first of all with simple verbs, second by predicative multi-word constructions.[16]

The first step was extension of PolNet with simple verbs. This extension was operated relatively fast due to high quality of the Polański's Generative Dictionary. However, as a machine-readable version of this dictionary did not exist, the work of building verb synsets was to be done fully manually by experienced lexicographers.

In (Vetulani, Z. & Vetulani, G., 2014) we presented the concept of a verb synset as follows: "In opposition to nouns, where the focus is on the relations between concepts (represented by synsets), and in particular on hyperonymy/hyponymy relations, for verbs the main interest is in relating verbal synsets (representing predicative concepts) to noun synsets (representing general concepts) in order to show what connectivity constraints corresponding to the particular argument positions are. (…) Synonymous will be only such verb+meaning pairs in which the same *semantic roles*[17] take the same concepts as value (necessary but not sufficient). In particular, the valency structure of a verb is one of indices of meaning (members of a sysnset share the valency structure)."

Synsets for simple verbs appeared already in the first public release of PolNet in 2011 (PolNet 1.0) (Vetulani, Z. et al. 2016). (See Fig. 2, below). Already this first extension steps in turning PolNet into a lexicon-grammar permitted us to make a smart use of PolNet enriched with lexicon-grammar features to control parsing execution by heuristics[18] in order to speed-up parsing due to additional information gathered at the pre-parsing stage. The effect of substantially reducing the processing time was due to the reduction of search space.

6.2 Next step – compound verbs

The next steps consisting in expanding the initial PolNet-based lexicon-grammar with compound verbs were more demanding. The first reason for that was scarcity of dictionaries of compound words (phrasemes, or special multi-word constructions like collocations), both for general vocabulary and for domain-specific terminology (with exception of some domains). Another problem for almost all languages is insufficiency of serious research concerning syntax, semantics and pragmatics for compound words. These two problem remain to be solved by the concerned teams.

6.2.1 Lexicographical basic research on verb-noun collocations

Systematic studies[19] of Polish verb-noun collocations were initiated in the 1990s by Grażyna Vetulani. In the first phase they consisted in "manual" examination of over 40,000 of Polish nouns on "Słownik Języka Polskiego PWN" (Szymczak, 1995). This operation resulted with selection of over 7,500 abstract nouns liable to predicative use. Among them, a subset of over 2850 typical predicative nouns was identified as of primary importance to start extending a verbs-only initial lexicon-grammar with verb-nouns collocation (Vetulani, G. 2000). This class is the most important, but also the most heterogeneous, thereby hard to processing.[20] It is composed of names of activities and behavior, names of actions, techniques, methods, operations, states, processes, human activities, nature of objects of various kinds, etc. All these predicative nouns select their (predicatively empty) support verbs, simple or compound, and arguments. It is typical of this class that predicative nous accept more than one (sometimes many) support verbs to form compound verbs (verb-noun collocations) with the valency structure of the noun. In most cases these collocations will be synonyms and therefore will belong to the same synset. However, the difference between collocations due to the selection of different support verbs will be visible at the pragmatic level.

The initial step consisting in dictionary-based acquisition of collocations was concluded by the publication of the first version of Verb-Noun Collocation Dictionary (Vetulani, G. 2000) of over 5,400 entries. The main efforts have been made to retrieve collocations from the traditional dictionary, to elaborate a human-and-machine processible format of entries and to produce dictionary entries.

What follows is an example of a dictionary entry in the format described in (Vetulani, G. 2000) :

[16] For Polish, construction of PolNet entries for simple nouns and verbs was relatively easy because of availability of good quality dictionaries, however for many of the so called *less-resourced languages* this task will be challenging.

[17] See (Palmer 2009).

[18] A well-constructed heuristic permits – on the basis of morphological and valency information combined with the switch technique of Vetulani, Z. (1994) – to reduce the complexity of parsing down to linear in an important number of cases.

[19] Cf. (Vetulani, G. and Vetulani, Z. 2012) for this paragraph.

[20] Other identified classes of predicative nouns are: feature names (over 2,800), names of frequent diseases (over 250), names of occupations or popular professions (over 1,400), and other (e.g. nouns supported by *event verbs*); notice that particular nouns may be polysemic and may belong to more than one class (Vetulani, G. 2000).

aluzja, f/ *(allusion)*

- czynić(Acc,pl)/N1do(Gen), *(to make ~s to sth)*
- robić(Acc,sing)/N1do(Gen), *(to make an ~ to sth)*
- pozwalać sobie na(Acc,pl)/N1do(Gen)*(to dare to make ~s to sth)*

where Acc and Gen stand respectively for declension cases of respectively accusativus and genitivus.

The second step was operated between 2000 and 2012. Its starting point was the dictionary of some 5,400 entries for more than 2850 predicative nouns (in what follows we call it *basic resource; BR*). Its main objective was a substantial improvement of the earlier results on the basis of large text corpus of Polish and appropriate processing tools. Analysis of the results obtained by the year 2000 brought to evidence insufficiency of methods used so far, as traditional dictionaries were not sufficiently large to contain all frequently used collocations. To obtain a satisfactory balanced coverage it was necessary to make use of corpora. Machine-assisted investigation of the text IPI PAN corpus (Przepiórkowski 2004) permitted to triple the number of collocation entries for the same basis of slightly more than 2,800 predicative words. To get this result we first proposed an algorithm for computer-assisted corpus-based acquisition of new collocations (Vetulani, G. et al., 2008), where by "new" collocations we mean those attested in a corpus, but absent in the BR. The main idea of this algorithm is to transform the rough corpus data in a way to substantially reduce the collocation-retrieval time with respect to fully manual retrieval procedure.

The input resources for the algorithm were:
1) Basic Resource of 2878 predicative noun entries (Vetulani, G. 2000).
2) The public available part of the IPI PAN corpus (Przepiórkowski, 2004) without morphological (and any other kind of) annotations.
The algorithm was organized into four parts:
- preparatory steps on the input data (preparation of search patterns)
- the main part which is a concordances generator to retrieve fragments of texts which match the patterns,
- clustering of text fragments obtained from the concordancer part with respect to predicative nouns (BR) and returning "support-verb candidates" (SVC) to be identified or refused as support verbs.
- manually processing (cleaning) support-verbs-candidates (SVC) in order to eliminate worthless selections (large majority).

This algorithm was further improved by the same team (Vetulani, G. et al. 2008; Vetulani, G. 2010) and applied to the input data. This modified algorithm is composed of the following five steps (to be run consecutively).

Step 1. Extraction from the corpus of contexts with high probability to contain verb-noun collocations and detection of verbs-candidates to be qualified as support verbs (automatically).

Step 2. Manual analysis by lexicographers of the list of verbs-candidates obtained in the Step 1 in order to eliminate apparently bad choices.

Step 3. Automatic extraction of contexts in form of concordances containing verb-noun pairs (selected through steps 1-3) as concordance centers.

Step 4. Reading of the concordances by lexicographers, qualification of verb-noun pairs as collocations and their morpho-syntactic descriptions (manual).

Step 5. Verification and final formatting.

The method we used permitted to reduce (~ 100 times) the processing cost (estimation on a 5% sample).

As result of the application of this algorithm we obtained an electronic dictionary of over 14,600 entries for over 2,878 predicative nouns.

6.2.2 Introduction of verb-noun collocations to PolNet

During the period from 2009 (PolNet 0.1) to 2011(PolNet 1.0) PolNet grew as a result of addition of some 1,500 synsets for 900 simple verbs corresponding to approximately 2,900 word+meaning pairs (Vetulani, Z. and Vetulani, G. 2014). Further extension from PolNet 1.0 to PolNet 2.0 consisted in addition of 1,200 new collocation synsets corresponding to 600 predicative nouns[21].

POS: v ID: 3441
Synonyms: {pomóc:1, pomagać:1, **udzielić pomocy**:1, **udzielać pomocy:1**} (*to help*)
Definition: "*to participate in sb's work in order to help him/her"*
VALENCY:

- Agent(N)_Benef(D)
- Agent(N)_Benef(D) Action('w'+NA(L))
- Agent(N)_Benef(D) Manner
- Agent(N)_Benef(D) Action('w'+NA(L)) Manner

Usage: Agent(N)_Benef(D); "Pomogłam jej." (*I helped her*)
Usage: Agent(N)_Benef(D) Action('w'+NA(L)); "Pomogłam jej w robieniu lekcji." (*I helped her in doing homework)*
Usage: Agent(N)_Benef(D) Manner Action('w'+NA(L));
"Chętnie udzieliłam jej pomocy w lekcjach." (*I helped her willingly doing her homework*)

[21] Notice that the number of collocations is higher than the number of predicative nouns, this is due to the fact that the same predicative noun may be supported by several support verbs.

Usage: Agent(N)_Benef(D) Manner;
"Chętnie jej pomagałam." (*I used to help her willingly*)
Semantic_role: [Agent] {człowiek:1, homo sapiens:1, istota ludzka:1, …} (*{man:1,...,human being:1,...}*)
Semantic_role: [Benef] {człowiek:1, homo sapiens:1, istota ludzka:1, …} (*{man:1,...,human being:1,...}*)
Semantic_role: [Action] {czynność:1} (*{activity:1}*)
Semantic_role: [Manner] {CECHA_ADVERB_JAKOŚĆ:1} (*qualitative adverbial*)

Fig. 2. Simplified DEBVisDic[22] presentation of a PolNet synset {pomóc:1, pomagać:1, **udzielić pomocy**:1, **udzielać pomocy:1**} containing both simple verbs (*pomóc*) and collocations (udzielić pomocy) (Vetulani, Z. and Kochanowski, 2014).

Fig. 2. presents the way PolNet makes use of the idea of semantics adapted after Filmore (1977) and Palmer (2009), and shows the semantic roles Agent, Beneficient, Action, Manner together with their values being noun synsets.

The passage to PolNet 2.0 opened up new application opportunities but also pushed us to re-consider the fundamental problem of synonymy for predicative words and to base it on the concept of valency structure. As valency structure of a verb is one of the formal indices of meaning, it should be considered as an attribute of a synset, i.e. all synset's members should share the valency structure. Strict application of this principle results in relatively fine granularity of the verb section of the PolNet (Vetulani, Z., Vetulani, G. 2015).

PolNet 3.0 is the last documented version of the resource. In order to obtain this new version, PolNet 2.0 was submitted to refining and cleaning operations. For the refinement operation it was assumed that the category of language register is a part of the meaning. The totality of PolNet 2.0 synsets was revised in order to split these synsets into register-uniform sub-synsets. Inclusion of register related information, up to our best knowledge until now not practiced in other wordnets, opens new application possibilities e.g. for refinement of text generation quality.

The version PolNet 3.0 has already been user-tested as a resource for modeling semantic similarity between words (Kubis, 2015).

	PolNet 0.1 (2009)	PolNet 1.0 (2011)	PolNet 2.0 (2013)	PolNet 3.0 (2016)
Nouns	10,629	11,700	11,700	12,011
Simple verbs	---	1,500	1,500	3,645
Collocations	---	---	1,200	1,908

Table. 1. Growth of the PolNet's main parts (in numer of synsets) (Vetulani, Z. et al. 2016). Notice: This table does not represent the effort invested in the development of PolNet as an important deal of work was engaged in the wordnet cleaning operations.

7. Conclusion and further work

Undoubtedly English constitutes an absolute reference point for languages classification in terms of adaptation of their description to technological needs as well as in terms of richness of tools and language resources necessary for industries to develop language technologies. At the very bottom of the hierarchy we find a significant number of languages for which it is not needed (or realistic) to develop such technologies. In the middle we locate quite a big number of languages set down as "less-resourced". Until recently the Polish language was categorized within this group. Currently we locate there some European minority languages as well as some languages from countries, by the way highly technologically developed, such as India for instance. Among other, we address this paper to researchers working on "less" or "middle-resourced" Indo-European languages as a proposal of a long term academic cooperation in the field, within which we will share experience with our partners in the area explored in this article. We believe that the confrontation of our methodology with other languages, also non-Indo-European languages of India, Ugro-Finish or Turkic in Europe and Asia, will allow for better understanding of the level of versatility of our solutions and, last but not least, will create conditions for a close cooperation.

The PolNet enlargement with verbal components required an important investment of lexicographers' work. Lexicon-Grammar for Polish is still in progress, but what has been done until now is largely sufficient to give a good insight in the nature of linguistic and engineering problems to be done by the project executors or by people aiming to undertake similar tasks for other languages. In the course of the above-mentioned works on Lexicon-Grammar for Polish we have identified a range of factors that appeared necessary to be taken into account in order to realize our project. At the beginning of our works we could dispose only of the following resources:

- traditional or in some cases electronic language resources such as: dictionaries, thesauri, lexicons
- representative and large texts corpora,
- traditional or formalized grammatical descriptions,
- IT tools for processing the above-mentioned resources.

In the lack of adequate resources the project began with building-up the lexical database of wordnet type, initially

[22] DEBVisDic is a tool we used for edition and maintenance of PolNet entries (Pala, K. et al.).

from nouns only, through inclusion in the next phase simple predicative verbs, and finally verb-noun collocations (with a predicative noun).

The enlargement of the initial, noun-based version of the PolNet database consisted in introducing predicative elements by, *inter alia*:

- identifying predicative simple and complex words: description of the predicate-valency structure for simple and complex predicative expressions and proposing a format for predicative synsets,
- generating predicative synsets and linking with respective arguments (noun synsets).

Problems to be solved / elaborated are as follow:

- synonymy, granularity,
- aspects,
- meaning shift, diachrony,
- other relations:
  - hyponymia/hyperonmy
  - meronymy
  - passive and active verb opposition
- morphology,
- pragmatic issues:
  - language registers,
  - regionalisms.

Another hot issue for existing lexicon-grammar systems of different languages is to align them with each other. It is a demanding task, often hardly feasible due to different conceptualization of the world in various communities, and reflected in respective languages.

The reason of presenting some our works on language resources within the Wildre workshop is the intention to encourage taking up the challenge thrown in the CFP of this workshop which is: "To provide opportunity for researchers from India to collaborate with researchers from other parts of the world".

8. Acknowledgements

The succesful achievement of projects reported in this overview would be impossible without the collaboration of numerous institutions and notable individuals. We would like to express our gratitude to the many academic collaborators, student volunteers, and colleagues across Academia and various partner, who generously invested their time and expertise to make this research possible.

9. Bibliographical References

Antoni-Lay, M-H. Francopoulo, G. Zaysser, L. (1994). *A Generic Model for Reusable Lexicons: The Genelex Project*, In *Literary and Linguistic Computing*, vol. 9, no 1, Oxford University Press, 47-54.

Fillmore, Ch.J. (1977). *The need for a frame semantics in linguistics. Statistical Methods in Linguistics*. Ed. Hans Karlgren. Scriptor.

Fillmore, Ch.J. (2002). Seeing Arguments through Transparent Structures, *Proceedings of Third International Conference on Language Resources and Evaluation, Proceedings*, Vol. III, Las Palmas, 787–791.

Gross, M. (1975). *Méthodes en syntaxe,* Paris: Hermann.

Gross, M. (1994). Constructing Lexicon-grammars. In Atkins and Zampolli (eds.) *Computational Approaches to the Lexicon*, Oxford University Press, p. 213-263.

Kubis, M. (2015). A semantic similarity measurement tool for WordNet-like databases. In Z. Vetulani and J. Mariani (Eds), *Proceedings of the 7th Language and Technology Conference, Poznań, Poland, 27-29 November 2015.* FUAM, Poznań, pp. 150 – 154.

Miller, G. A, Beckwith, R., Fellbaum, C.D., Gross, D., Miller, K. (1990). WordNet: An online lexical database. Int. J. Lexicograph. 3, 4, 235–244.

Palmer, M. (2009). Semlink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference. Sept. 2009, Pisa, Italy.*

Polański K. (Ed.) (1980-1992). *Słownik syntaktyczno-generatywny czasowników polskich,* vol. I-IV, Ossolineum, Wrocław, 1980-1990, vol. V, Instytut Języka Polskiego PAN, Kraków, 1992.

Piasecki M., Szpakowicz S., Broda B. (2009). *A Wordnet from the Ground Up*, Oficyna Wydawnicza Politechniki Wrocławskiej,Wrocław.

Przepiórkowski, A. (2004). *Korpus IPI PAN. Wersja wstępna (The IPI PAN Corpus: Preliminary version).* IPI PAN, Warszawa.

Szymczak, M. (red.) (1995). *Słownik Języka Polskiego*, PWN, Warszawa.

Vetulani, G. & Vetulani, Z. & Obrębski, T. (2008). Verb-Noun Collocation SyntLex Dictionary - Corpus-Based Approach, In: *Proceedings of 6th International Conference on Language Resources and Evaluation, May 26 - June 1, 2008, Marrakech, Morocco (Proceedings)*, ELRA, Paris.

Vetulani, G. (2000). *Rzeczowniki predykatywne języka polskiego. W kierunku syntaktycznego słownika rzeczowników predykatywnych na tle porównawczym,* Adam Mickiewicz University Press: Poznań.

Vetulani, G. (2010). *Kolokacje werbo-nominalne jako samodzielne jednostki języka polskiego. Syntaktyczny słownik kolokacji werbo-nominalych języka polskiego na potrzeby zastosowań informatycznych. Część I.* Adam Mickiewicz University Press: Poznań.

Vetulani, G. and Vetulani Z. (2012). Dlaczego Leksykon-Gramatyka? In: Anna Dutka-Mańkowska, Anna Kieliszczyk, Ewa Pilecka (ed.), *Grammaticis unitis. Mélanges offers à Bohdan Krzysztof Bogacki,* Wydawnictwa Uniwersytetu Warszawskiego. 308-316.

Vetulani, Z. (1988). PROLOG Implementation of an Access in Polish to a Data Base, w: Studia z automatyki, XII, PWN, 1988, pp. 5-23.

Vetulani, Z. (1994). SWITCHes for making Prolog more Dynamic Programming Language, Logic Programming, The Newsletter of the Association for Logic Programming, vol 7/1, February 1994, pp. 10.

Vetulani, Z. and Jassem, K. (1994) Linguistically Based Optimisation of a TDDF Parsing Algorithm of the NL system POLINT. In: Dieter W. Halwachs (Eds.), Akten des 28 Linguistischen Kolloquiums, Graz - 1993 (Linguistische Arbeiten 321), Max Niemeyer Verlag, Tübingen, 1994, pp. 321-326.

Vetulani, Z. (1997). A system for Computer Understanding of Texts, in: R. Murawski, J. Pogonowski (Eds.), *Euphony and Logos*. Poznań Studies in the Philosophy of the Sciences and the Humanities, vol. 57. Rodopi, Amsterdam-Atlanta, 387-416; ISBN: 90-420-0382-0; ISSN 0303-8157.

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*, Adam Mickiewicz University Press, Poznań.

Vetulani, Z. (2000); Electronic Language Resources for POLISH: POLEX, CEGLEX and GRAMLEX. In: M. Gavrilidou et al. (Eds.) *Second International Conference on Language Resources and Evaluation, Athens, Greece, 30.05.-2.06.2000, (Proceedings)*, ELRA, pp. 367-374.

Vetulani, Z., Walkowska, J., Obrębski, T., Konieczka, P., Rzepecki P., Marciniak, J. (2007). PolNet - Polish WordNet project algorithm, in: Z. Vetulani (ed.) *Proceedings of the 3rd Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, October 5-7, 2007, Poznań, Poland*, Wyd. Poznańskie, Poznań, ISBN 978-83-7177-407-2, pp. 172-176.

Vetulani, Z., Marcinak, J., Obrębski, T., Vetulani, G., Dabrowski, A., Kubis, M., Osiński, J., Walkowska, J., Kubacki, P., Witalewski, K. (2010). *Zasoby językowe i technologie przetwarzania tekstu. POLINT-112-SMS jako przykład aplikacji z zakresu bezpieczeństwa publicznego (in Polish) (Language resources and text processing technologies. POLINT-112-SMS as example of homeland security oriented application)*, ISBN 978-83-232-2155-5, ISSN 1896-379X, Adam Mickiewicz University Press: Poznań.

Vetulani, Z. (2012). Wordnet Based Lexicon Grammar for Polish. *Proceedings* of the Eight International Conference on Language Resources and Evaluation (*LREC 2012*), May 23-25, 2012. Istanbul, Turkey, (Proceedings), ELRA: Paris. ISBN 978-2-9517408-7-7, pp. 1645-1649. http://www.lrec-conf.org/ proceedings/ lrec2012/index.html

Vetulani, Z., Kochanowski, B. (2014). "PolNet - Polish Wordnet" project: PolNet 2.0 – a short description of the release, in: Heili Orav, Christiane Fellbaum, Piek Vossen (eds.) *GWC 2014. Proc. of the Seventh Global Wordnet Conference 2014, Tartu, Estonia*, Global Wordnet Association, pp. 400-404.

Vetulani, Z., Vetulani, G. (2014). Through Wordnet to Lexicon Grammar, in. Fryni Kakoyianni Doa (Ed.). *Penser le lexique grammaire: perspectives actuelles*, Editions Honoré Champion, Paris, pp. 531-543.

Vetulani, Z., Vetulani, G., Kochanowski, B. (2016). "Recent Advances in Development of a Lexicon-Grammar of Polish: PolNet 3.0". In: Nicolette Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, European Language Resources Association (ELRA), Paris, France, ISBN 978-2-9517408-9-1, pp. 2851-2854, hal-01414304.

Walkowska, J. (2009). *Gathering and Analysis of a Corpus of Polish SMS Dialogues*. [in:] Kłopotek, M. A., et al. (Eds.), Challenging Problems of Science. Computer Science. Recent Advances in Intelligent Information Systems Academic Publishing.

10. Language Resource References

Zygmunt Vetulani (2014). POLEX Polish Lexicon. ISLRN 147-211-031-223-4, distributed via ELRA, http://catalog.elra.info/en-us/repository/browse/ELRA-L0074/.

Zygmunt Vetulani (2016). PolNet – Polish Wordnet. ISLRN 944-121-942-407-9. To be found at http://www.islrn.org/resources/944-121-942-407-9/.