NEOPHILOLOGICA 2023, T. 35, 1–32
ISSN 2353-088X
DOI: https://doi.org/10.31261/NEO.2023.35.22

Zygmunt Vetulani
Adam Mickiewicz University, Poznań
Poland
https://orcid.org/0000-0003-4833-8601

Grażyna Vetulani
Adam Mickiewicz University, Poznań
Poland
https://orcid.org/0000-0002-2138-3704

Towards Lexicon-Grammar Verbnets Through Lexical Ontologies

Abstract
In this article, we present research directly inspired by the Princeton WordNet lexical *ontology* project (Miller, Fellbaum), which was a response to the real need for ontologies corresponding to the *natural conceptualization* common to all language users, within a given natural language, or within a specific *sublanguage*. Lexical ontologies for a given language or language subsystem determined by the scope of communication needs turn out to be useful and even necessary for constructing formal models of linguistic competence and, consequently, for designing and implementing AI systems with linguistic communicative competence, both passive and active. An important milestone of the research program presented in this work is the acquisition of tools in the form of extensive lexical ontologies of a new type, referred to in this work as *Lexicon-Grammar Verbnets*. In the article, we refer to the works of authors such as: Alain Colmerauer, Charles Fillmore, Christiane Fellbaum, Gaston Gross, Maurice Gross, Thomas R. Gruber, Richard Kittredge, George A. Miller, Martha Palmer, Kazimierz Polański, and Piek Vossen.

Keywords
lexical ontology, synonymy, valency structure, wordnet, PolNet, Lexicon-Grammar Verbnet, IT systems with language competence

The basis of all human mental activity is the formation of abstract concepts. At first, this enabled humans to understand a situation and make decisions. Later, the need to organize themselves in communities forced *communication* with other individuals of the group. As individuals began to identify themselves and distinguish themselves from other members of the community, there was a need to *plan and implement* collective activities within the group based on a *conceptual model of the surrounding world.* The model of the environment in which a human being functions is necessary to undertake rational actions consistent with the adopted goals. *Conceptualization* as a process of creating abstract concepts was of interest to the philosophers of the late nineteenth and early twentieth centuries. In the field of computer science, the term "conceptualization"[1] was popularized by Tomasz R. Gruber (1993) by aptly linking it with the understanding of the term "ontology," introduced by him to computer science. Gruber characterized the term "ontology" in a compact form as follows: "An ontology is an explicit specification of a conceptualization. […] Ontology is a systematic account of Existence. For knowledge-based systems, what *exists* is exactly that which can be represented." (1993, p.199).

Information technologies, including AI technologies, which are the subject of this article, refer to the issues of modeling *human language competence* in order to use the models obtained for the design and implementation of *systems with linguistic communicative competence*.

The research presented below was directly inspired by the Princeton WordNet lexical ontology pioneering project implemented by G. A. Miller and C. D. Fellbaum (see e.g., Miller et al., 1990), which was a response to the real need for ontologies corresponding to the *natural conceptualization* common to all users of a given language, or a sublanguage used in a strictly defined field. The Princeton WordNet has been an inspiration for many lexical ontologies for various languages (including PolNet).

Lexical ontologies for a given language system or subsystem (determined by specific and well-defined communication needs) turned out to be useful, and even necessary for constructing formal models of linguistic competence, and consequently for designing and implementing *AI systems* with language communication competence, both passive and active.

[1] We use the term "conceptualization" in the sense given by Thomas R. Gruber, who wrote: "A conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose. Every knowledge base, knowledge-based system, or knowledge-level agent is committed to some conceptualization, explicitly or implicitly" (Gruber, 1993, p. 199). In nineteenth-century philosophy, the term was used to refer to the formation of abstract concepts.

Knowledge processing and the ability to build a model of knowledge about the environment in which language users participating in the speech act function (people, devices, systems) are two key components necessary to achieve the goal of creating new generation AI systems at a level significantly exceeding current *chatGPT* systems. We will show, among other things, a number of results, including our owns, which make up a methodologically coherent whole, and which bring us – step by step – closer to the above-defined goal. An important stage of the research program outlined here is obtaining tools in the form of complex lexical ontologies of a new type, referred to as *Lexicon-Grammar Verbnets*.

The title of this paper, "Towards Lexicon-Grammar Verbnets Through Lexical Ontologies," illustrates the course (of a part) of our work in the field of Human Language Technologies from the 1980s until now. We consider obtaining a lexicon-grammar of the verbnet type (*Lexicon-Grammar Verbnet for Polish*) with a rich conceptual coverage as a solid basis for further R&D and implementation works in the field of IT.

In this article, we present the results scattered across a number of our publications containing essential elements and ideas. They form the backbone of a long-term, ongoing research program.[2]

The review of the results begins with early works, conducted in the conditions of scarcity of digital language resources, both lexical and grammatical. These are prototypes of systems constituting the BPII (Basic Polish for Information Interchange) family, as well as results in the field of digital lexical data and digital grammatical data formats obtained as part of national and European projects (POLEX and the EU projects PECO-COPERNICUS CEGLEX and PECO-COPERNICUS GRAMLEX); see section Early Works. Section WordNet Like Lexical Ontologies focuses on the development of a wordnet lexical ontology. The part concerning basic research mainly deals with the problems of synonymy, while the practical part presents the implementation of a lexical ontology of the WordNet type (for the Polish language) PolNet v1. Section From PolNet 1.0 to Lexicon-Grammar VerbNets is the main part of the work and concerns the transformation of the PolNet v1 lexical database into a lexical ontology, which is a VerbNet type Lexicon-Grammar. The most important challenge at the current stage of development of WordNet systems with Lexicon-Grammar features turned out to be the extension of synonymy relations and homonymy/hyperonymy relations to predicative synsets. This section discusses, among other things, the currently performed tasks.

[2] Theoretical and practical studies reported here relate directly to Polish, but are largely localizable to other language systems, primarily from the Indo-Aryan language family (Vetulani et al., 2021).

Early Works

The research referred to in the title of the article is the direct result of previous projects that made us aware of the shortages of basic resources and IT tools for processing the Polish language. The beginnings of our work on systems with linguistic competence in the 1980s and 1990s, and partly their continuation, were characterized by the lack of access to digital linguistic resources (dictionaries, grammars) in a form that would enable their direct use in IT applications. Nevertheless, the Polish language belongs to a small elite group of languages with a long tradition of linguistic work, which turned out to be a solid theoretical base for our research.

The successful development of systems with language competence became possible thanks to the work we started on a grammatical description of the Polish language, suitable for IT use in parsing algorithms, that is, in algorithms that perform syntactic analysis, which is a preparatory stage in the process of calculating the meaning of a text.[3] The result of this work were the POLINT grammars developed since 1980s. Our source of inspiration was the question-answering system ORBIS implemented in PROLOG for English and French by A. Colmerauer and R. Kittredge (using DCG) (Colmerauer & Kittredge, 1982), later extended by Vetulani with a Polish module (Vetulani, 1988).

The first POLINT programs (see Vetulani, 1988) focused on modeling question-answer dialogues, were created in order to demonstrate the application potential in terms of language coverage in BPII systems[4] and to obtain practical knowledge of linguistic resources necessary to meet the needs of application systems. This potential was positively tested in the confrontation with the empirical material in the form of a corpus of empirically generated dialogues (Vetulani, 1990), and finally confirmed in the POLINT-112-SMS system (Vetulani & Osiński, 2017).

The first successful attempts to parse sentences of the Polish language already allowed us, in accordance with our expectations and with postulates of Antonio

[3] It should be noted here that this research was carried out under simplifying assumptions, namely the compositionability and computability of the language. These assumptions have been discussed among philosophers of language and linguists since at least the Enlightenment period, but they seem essential for creating IT-useful, precise models of language, pushing the boundaries of deterministic language modeling.

[4] The subset of Polish corresponding to the grammatical coverage of POLINT prototypes from the late 1980s is referred to in this period as BPII (Basic Polish for Information Interchange).

Zampolli,[5] creator and promotor of the concept of *Language Industry* to identify priorities in terms of technological needs of Language Engineering. One of the first of our ventures was the POLEX dictionary project (1994–1996).[6]

Polish is a language with a complex inflection system and has a relatively free word order. Therefore, simple adaptation of processing algorithms efficient for languages like English or French appeared hard to apply, as in Polish the basic information concerning the function of a word in the sentence is typically being encoded in the word form, independently of its linear position in the sentence. Dictionaries are a suitable place to store this information. At the time we started our research, good-quality grammatical descriptions of Polish existed only in the form of traditional dictionaries and grammars addressed to traditional customers. However, these resources, typically addressed to human readers, appeared to be of low usefulness for automatic processing because of lack of precision.

Our solution, the POLEX Polish Lexicon, is an electronic morphological dictionary which includes the core Polish vocabulary of general interest acquired from the traditional paper dictionary (Szymczak, 1983–1985).[7] POLEX is based on a precise machine-interpretable format (coding system), the same for all grammatical categories (Vetulani et al., 1998a).

The POLEX format we propose is uniform for all grammatical categories (parts of speech) and does not apply exceptions to the rules, which makes creating algorithms for generating and lemmatizing text much easier than when using traditional language descriptions, which place high demands on programmers due to the excessive complexity of the description. The POLEX dictionary entries take the following form:

BASIC_FORM+LST_OF_STEMS+PARADIGMATIC_CODE+DISTRIBUTION_OF_STEMS

[5] Antonio Zampolli considered the lack of resources in the form of IT-processable corpora, dictionaries, and digital grammars necessary to build language models to be a critical obstacle in the development of utility IT systems. See *Language Resources. Overview* by J. J. Godfrey and Antonio Zampolli (1996) and (Zampolli, 2006).

[6] The first public release of the resource contained over 42,000 nouns, 12,000 verb, 15,000 adjectives, 25,000 participles, and about 200 pronouns.

[7] Supplemented by the basic swear words not found in this dictionary and the most frequently used elements of jargon, regional and colloquial vocabulary. Several paper editions of *Słownik Języka Polskiego PWN* [Polish Language Dictionary PWN] by Mieczysław Szymczak were edited between 1978 and mid 1990s. For our purposes we used the three volume version published from 1983 to 1985; see (Szymczak, 1983–1895).

For example, dictionary entries for two inflected variants of the word *sucker*[8] look 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

The *paradigmatic inflection code* contains full paradigmatic information about *inflection*, that is, the way of associating endings with stems in order to obtain the desired word form. The inflection code (here *N110*) includes full information on morphology and inflection, in particular a list of endings appropriate for all paradigmatic positions. The distribution parameter (*distribution_of_stems*) relates stems (here *frajer*, *frajerz*) to paradigmatic positions. The information stored in a dictionary entry is complete and unambiguous, and inflection classes are constructed in such a way that there is no need to consider exceptions.

The other two projects discussed in this section were of a different nature.

The main goal of the CEGLEX consortium (Vetulani et al., 1998b) was to test the EU EUREKA project GENELEX that proposed a reusable generic model for lexicons assumed to respond to IT needs. GENELEX was implemented (between 1990 and 1994) for a number of Western European languages, such as French, English, German, Italian. Within CEGLEX three Central-European languages, Polish and Czech (Slavic) and Hungarian (Finno-Ugric) were used as testbeds to verify genericity of the GENELEX model.

It is worth noting that the final Polish module developed in CEGLEX/GENELEX went further than original GENELEX, which focused on morphological and syntactic layers while the semantic layer was addressed only marginally.

The three layers of the CEGLEX/GENELEX model were confronted with linguistic data of the languages under consideration with generally positive results. For the Polish module of the project this confrontation consisted in adapting the model GENELEX to Polish language data.

The CEGLEX project resulted in a successful attempt to test (on representative linguistic data) the feasibility of an IT-oriented lexicon-grammar covering all three basic layers (morphological, syntactic, and semantic) of grammatical description.

[8] The word *frajer* (en. *sucker*) is a masculine-personal noun (pol. *rzeczownik męskoosobowy*), inflected for number and case (two numbers /singular and plural/ and seven cases. The themes (*frajer* and *frajerz*) are the same in both entries. Code *N110* represents a 14-position string of endings, the same in both entries for the lexeme *sucker:* (, a, owi, a, em, e, e; y, ów, om, ów, ami, ach, y). The fourth parameter, topic distribution, describes the assignment of each paradigmatic ending to the appropriate topic. In this example, 1:1-5,9-13 means that the first stem (sucker) is combined with the endings from paradigm positions 1 to 5 and 9 to 13. Similarly, the expression *2:6-8,14* indicates that the endings from positions *6, 7, 8* and *14* are connected to the second stem (*sucker*).

The CEGLEX/GENELEX methodology together with the results of the POLEX project were the starting point for our work within the PECO-COPERNICUS GLAMLEX project, carried out from 1995 to 1998. The main goal of this project was to build, in accordance with the GENELEX methodology, morphological digital dictionaries and related IT-oriented tools. The intention of the tasks of the GRAMLEX project (Vetulani et al., 1998b) regarding the Polish language was to contribute to the improvement of the situation in the field of language engineering tools and resources. Among the main achievements of GRAMLEX was the creation of a corpus-based morphological dictionary for the Polish language encoded in SGML (in the proprietary GRAMCODE format.[9]

The GRAMLEX project turned out to be the first step towards implementing a lexicon-grammar for the Polish language.

WordNet Like Lexical Ontologies

Synonymy, Hyperonymy and Inheritance[10]

The concept of *synonymy* refers to the concept of *meaning*, which is commonly used in informal discourse and usually does not raise controversy. Consequently, it is generally used as a *primary concept*, not requiring analytical definition referring to other concepts treated as known. If the reference to the obviousness of a concept turns out to be inappropriate, then an *axiomatic definition* can be used. Definitions of this type do not enter into the ontological nature of the defined concept, but are operational in nature, specifying the way of using the concept by referring to another, assumed to be already known. A classic example is the Peano arithmetic (around 1889),[11] where primitive concepts such as additions, multiplications, natural numbers, etc., are explained by axioms that, by reference to other concepts, determine how to use these concepts. In traditional linguistics, similar methods are sometimes used to determine the meaning of a word or phrase by giving usage examples considered representative. When defining

[9] The GRAMCODE dictionary included over 22,500 dictionary entries along with related tools and applications (lemmatizer, inflectional generator, concordance generator and others).

[10] This section is based on our paper "Synonimie et granularitė dans les bases lexicales du type WordNet" (Vetulani, Z. and Vetulani, G., 2015) and "EuroWordNet General Document" (Vossen, 2002). In particular, we follow Vossen in using the term "hyperonymy" for nouns and verbs.

[11] See Giuseppe Peano (1889) at https://en.wikipedia.org/wiki/Giuseppe_Peano.

synonymy, reference to meaning may be appropriate if well-defined procedures are used to compare word meanings, for example, by applying context-of-use analysis (see Vossen, 2002).

Our initial work on lexical ontologies was motivated by the desire to obtain a basic ontology for the Polish language inspired on the one hand by Linnaeus' systematics, and on the other hand by the pioneering work of cognitive scientists and linguists from Princeton (Miller, Fellbaum and others) on the WordNet system. The Princeton WordNet was an ontology directly linked to lexical material in the form of abstract nouns grouped into classes of synonyms called *synsets*. The inspiration turned out to be accurate and led to the creation of the linguistic ontology, PolNet – Polish Wordnet (version 1.0), which satisfactorily corresponds to the conceptualization reflected in the nouns of the Polish language.

The Problem of Defining the Concept of Synonymy

In natural languages, concepts (understood as mental equivalents of complex or simple entities) are represented by words. *Synonymy* is commonly understood as a binary relation holding between two words (terms, expressions) if and only if its arguments have *the same or similar meaning*. The need to define the relation of synonymy more precisely leads us to distinguish the three cases where the term synonymy will be used.

Case 1. If the concepts represented by their names (being simple or compound nouns) are *extensional* (that is, when they can be fully described by specifying which of the entities fall under the given concept and which do not), then by *synonymy* of some two names we understand that both names refer to *the same set of entities*. In a similar way, we can construct an extensional definition of *synonymy* of verbs: two verbs are said to be synonymous when they both refer to the same set of states and/or events.

Case 2. In turn, when the meaning of each of the two words compared to each other can be unambiguously determined by *a specific* set of features and their values, then their *synonymy* means that both can be uniquely described by *the same* set of features (attributes) taking the same values.

Case 3. If neither of the above two cases occurs, then it remains to refer to definitions of the nature of procedures referring to the circumstances of the use of each of the compared terms.

Let us compare three of the frequently discussed solutions:

1) Leibnitz's proposal,
2) Princton WordNet proposal (Miller-Fellbaum),
3) EuroWordNet proposal (Vossen).

Ex. 1) We quote, after Vossen (2002, p.18), a very strong definition of synonymy given by Leibnitz:

> "two expressions are synonyms if the substitution of one for the other never change the truth value of a sentence in which the substitution is made."

Note that when using this definition, *synsets* are generally very small, or even composed of one element only. This means a significant flattening of the hierarchy based on the hypernymy relationship, which in turn reduces the potential benefits of the inheritance mechanism of attributes associated with synsets and the values of these attributes. The advantage of the Leibnitz's proposal is that synonymy is an equivalence relation and thus marks a partition in a set of words.

Ex. 2) George A. Miller and Christiane Fellbaum (see Vossen, 2002, p. 18) proposed a less restrictive approach to synonymy, encapsulated in the formula:

> "two expressions are synonymous in a linguistic context C, if the substitution of one for the other in C does not alter the truth value."

In the literal sense, it means that to conclude that these expressions are *not synonymous*, it is enough to refer to just one selected context C in which the replacement of one expression with another will change the logical value of the whole sentence.[12] *De facto*, this procedure (correctly) indicates as synonyms only those words for which the fixed context C can be considered *representative* of a particular meaning. In dictionary practice, the condition of representativeness of examples (containing the context of use) for illustrating a typical meaning is not strictly observed (see e.g., Polański, 1980–1992), which in practice may significantly hinder the creation of WordNet-type systems based on the above definition of synonymy.

Ex. 3) Piek Vossen proposes the synonymy tests for various parts-of-speech (including noun-noun, verb-verb, noun-verb, etc.) implemented in the EuroWordNet project (Vossen, 2002).

[12] "The weak point of Miller's approach is the synonymy criterion (above) which – alone – is not sufficient to form synsets because it does not guarantee transitivity when the C context changes. To remedy this defect, the initial criterion of synonymy must be reinforced by imposing the reference to the same context C." (Vetulani, Z. and Vetulani, G., 2015, p. 117). [Translation from French by Z. Vetulani: "Le point faible de l'approche de Miller est le critère de synonymie (ci-dessus) qui – seul – ne suffit pas pour former les synsets car il ne garantit pas la transitivité quand le contexte C change. Pour remédier à ce défaut il faut renforcer le critère initial de la synonymie en imposant la référence à un même contexte C."]

What follows is an example of an EuroWordNet context-based tests (applied to English) for noun-noun synonymy (Test 1) (Vossen, 2002, p. 19).[13]

Test 1. Synonymy between nouns

yes	a	if it is (a/an) X then it is also (a/an) Y
yes	b	if it is (a/an) Y then it is also (a/an) X
Conditions:		X and Y are singular or plural nouns
Example:	a	if it is a fiddle then it is a violin
	b	if it is a violin then it is a fiddle
Effect:		synset variants {fiddle, violin}

Hierarchical Organization of Concepts in PolNet

The classic wordnet organization for nouns is based on a hierarchy of concepts referring to the relation of hyponymy/hyperonymy *for nouns*. This hierarchy has a *tree structure*. More general concepts are higher in this hierarchy and those more specific are lower down. Tree organization is intended to allow *inheritance of properties*, essential for knowledge representation and inference (see *Linnaean systematics*[14]). The extension of the PolNet lexical ontology to *predicative synsets*[15] introduces relations between predicative synsets and other ontology entities (synsets or not). Of particular importance is the introduction of relations that connect the predicative synset with arguments that are assigned attributes called *semantic roles* (which are synsets or other objects of the PolNet ontology). Assigning semantic roles to the argument positions opened in predicative expressions serves to determine links or connectivity constraints between these expressions and arguments.[16] Expanding PolNet with predicative synsets requires special caution when extending the hyponymy/hyperonymy relationship to predicative synsets.

PolNet Development Incremental Algorithm (Nouns)

In this section we present an algorithm of creating synsets and hierarchical relations based on hyponymy/hyperonymy relations between nouns (simple and compound). This algorithm was directly used by lexicographers in the first phase

[13] Notice however, that context-based tests ignore differences due to the pragmatic factors.

[14] Carl von Linné (1707–1778), *Systema Naturae* (1770); see https://en.wikipedia.org/wiki/Systema_Naturae.

[15] By *predicative synsets* we mean synsets whose typical elements are predicative nouns or verb-noun collocations.

[16] The set of these assignments constitutes the *valency structure* of the synset.

of building the PolNet v.1 database. The DebVisDic platform developed at the Masaryk University Brno was used (Pala et al., 2007).

Application of the algorithm requires:

- the Visdic or DEBVisDic platform (or any functionally equivalent tool),
- on-line access to Princeton WordNet,
- a *good* monolingual lexicon[17] (called *reference dictionary* in the algorithm description), preferably accessible on-line (we used *Uniwersalny słownik języka polskiego PWN*[18] (Dubisz, 2006) as the basic reference lexicon and *Słownik języka polskiego PWN* (Szymczak, 1995) as a complementary one),
- a team with both language engineering and lexicographical skills.

The algorithm input consists of a list of words (lexemes). The output is a WordNet code segment for: a) synsets, b) the ISA relation between synsets (detemined by the hyponymy/hyperonymy relation).

The general procedure for expanding PolNet consists in performing a sequence of operations, step by step:

1. Looking through the reference dictionary, we search for *word-meanings*[19] that are synonyms.
2. We create synonymity classes using appropriate definition criteria. These classes are called *synsets*.
3. For created or modified synsets, we search for candidates for hyponyms and hyperonyms using our own language competence, dictionaries, LSR list and knowledge of the Princeton WordNet structure.
4. For pairs of synsets selected in step 3, we perform hyponym and hyperonym definition tests.

Short example:

Let us take the Polish word *zamek* as an example. The list of the word-meanings identified at in step 1 will be:

- zamek-1 (zamek I-1 in the dictionary): *a lock*
- zamek-2 (separated from the zamek-1 meaning, where the phrase *zamek błyskawiczny* is mentioned): *a zip fastener*
- zamek-3 (zamek I-2): *a machine blocking lock, e.g. a valve lock*
- zamek-4 (zamek I-3): *a gun lock*
- zamek-5 (zamek II-1): *a castle*

Zamek-2, *zamek-3* and *zamek-4* will all be hyponyms of *zamek-1*.

[17] By a *good dictionary* we mean one where different word-meanings are explicitly distinguished.

[18] PWN is the name of a Polish publishing house.

[19] By *word-meaning* we mean a meaning of the *literal* together with its reference-dictionary-meaning number.

Language Resources Used: Dictionaries and Tools

The research, the main results of which are summarized in this article, make up a description of the research path leading to a coherent methodology enabling the design and implementation of large AI systems[20] with language competence. One of the most important milestones of the long-term research program discussed here is the implementation of a prototype of a large AI system used for practical verification of decisions regarding the selection and/or development of appropriate tools and methods for natural language engineering.[21]

In addition to standard tools and methods commonly recognized as elements of the canon of IT and linguistic knowledge, during our research (until the implementation of the testing system /POLINT-112-SMS/), we considered it appropriate to use two classes of resources:

A) specialized resources and publicly available tools – necessary or useful in the project,

B) own resources and tools developed in the project, which turned out to be needed to implement the milestones of our work.

Class (A) includes:

- IPI PAN National Corpus of Polish Language (on a limited scale) (Przepiórkowski, 2004),
- *PWN Polish Language Dictionary* (version edited by M. Szymczak, 1995),
- *Universal Dictionary of the Polish Language* (edited by S. Dubisz, 2006),
- *Syntactic-Generative Dictionary of Polish Verbs* (Polański, 1980–1992),
- Internet dictionary SJP.PL, more on this topic in (Vetulani et al., 2010, p. 158–159),
- Tools for generating WordNet lexical networks – VisDic and DebVisDic (Masaryk University Brno) (Pala et al., 2007).

Group (B) includes:

- formats and vocabularies created in the POLEX, GRAMLEX and CEGLEX projects (Vetulani et al., 2010),
- a corpus of private SMS records – collected and made available by Justyna Walkowska,
- a corpus of experimental SMSs (collected and described by Justyna Walkowska in her PhD dissertation[22] (see, e.g., Vetulani et. al. 2010),
- a corpus of legal texts (compiled from open sources),

[20] By “large AI system” we mean a utility application at the stage of at least pre-commercial testing.

[21] The appropriate system called POLINT-112-SMS has been described in a collective monograph (Vetulani et al., 2010) and its brief characteristics are in the annex to this work.

[22] See (Walkowska, 2012).

- a corpus of recordings from the emergency telephone 997/112 (confidential recordings, not intended for sharing),
- verb-noun collocations for the Polish language: methodology, data formats, predicative-nouns /basic resource/ (Vetulani, G., 2000), basic noun-synsets-creation algorithm (Vetulani et al., 2007), algorithms for expanding the collocation dictionary (Vetulani, G., Vetulani, Z., Obrębski, T., 2008), a digital dictionary of verbal-nominal collocations (Vetulani, G., 2012),
- coding algorithms for valency dictionaries,
- various algorithms for expanding the PolNet database (as of 2010).

Inspirations. Princeton WordNet

Creating advanced systems with language competence, such as AI systems, requires knowledge processing, and thus referring to abstract concepts. For this purpose, ontologies (as defined by T. R. Gruber) are used (see the opening paragraph of the article). Ontologies, which on the one hand correspond to the natural conceptualization of the world – real or fictitious, and on the other hand are formal entities subject to IT processing, are *WordNet-type systems.*[23] The WordNet lexical ontology (also known as Princeton WordNet /PWN/) is an implementation, in the 1980s by G. A. Miller and colleagues at Princeton University's Cognitive Sciences Laboratory, of a new method for describing semantic vocabulary that has proven particularly useful for searching information on the Internet. The key idea of this method is to present the lexicon described by referring to the concepts of synonymy and hyperonymy. PWN is composed of classes of synonyms called *synsets* and is organized hierarchically by the relation of hyponymy/hyperonymy between synsets. Some other semantic relations between synsets (as meronymy, antonymy, etc.) are implemented as well. WordNet-like systems have an advantage over traditional ontologies because they explicitly account for the relationships between the words of the language and the concepts of the ontology /represented by synsets/.

Lexical Ontology PolNet v1

Our research initiated in the early 2000s was inspired by the work of George A. Miller and his team on the WordNet lexical ontology, as well as by the work

[23] This term is used to describe ontological systems modeled on PWN.

led by Piek Vossen in the EuroWordNet project. At later stages of work on the PolNet system, we also relied on the pioneering research of Maurice Gross on the concept of Lexicon-Grammar, initially implemented for the French language (Gross, M., 1975; 1994; 1981) and independently conducted work (in the same period) by Kazimierz Polański, and crowned with the implementation in 1980–1992 of the *Syntactic-Generative Dictionary of Polish Verbs*, as well as on the results of the FrameNet (Fillmore et al., 2002) and VerbNet (Palmer, 2009) projects, close to the assumptions of Lexicon-Grammar.

The launch in 2006 of the construction of PolNet (a lexical ontology intended to be a wordnet-type lexical database) was a response to the need for a language processing module for implementation of an stand-alone, large-scale IT system with language competence (POLINT-112-SMS) (Vetulani, Z., 2014). While the concept and structure of the PolNet database was modeled on the solutions adopted for the Princeton WordNet system (Miller and Fellbaum, 2007), the methodology for creating the PolNet database was developed from scratch by a team of Polish computer scientists and lexicographers.[24] The adopted methodology assumed the use of existing dictionaries of Polish in order to maintain the conceptualization appropriate for users of the Polish language.

The PolNet database is a structure built from synonym classes and relations between these classes. Synonym classes (*synsets*) represent concepts identifiable in natural language, thanks to which PolNet can be used as a *lexical ontology* corresponding to the conceptualization reflected in the Polish language. PolNet v1 was built on the basis of high-quality traditional dictionaries of the Polish language and the study of available language corpora (such as IPI PAN Corpus (Przepiórkowski, 2004) and small domain corpora). Resource creation is done incrementally, starting with high-frequency vocabulary[25] and words that are (for various reasons) considered important.

While the initial work on PolNet was conducted towards a system with a structure similar to the Princeton WordNet and intended to serve as an ontology naturally associated with the language, over time, the PolNet project, influenced by theoretical work carried out independently by Maurice Gross and Kazimierz Polański and implementation-oriented works (in particular by Alain Colmerauer and Charles Fillemore), evolved into a Lexicon-Grammar by gradually incorpo-

[24] Mainly from the Department of Computer Linguistics and Artificial Intelligence of the Adam Mickiewicz University and the Faculty of Modern Languages and Literatures of the Adam Mickiewicz University.

[25] A departure from this principle, made for methodological reasons in order to enable early testing of the developed resource in applications for which the condition of lexical completeness must be met, was the inclusion of terminology specific to these applications.

rating simple and compound verbs. This evolution coincided with the progress of theoretical work on the development of a formalized dictionary of verbal-nominal collocations initiated in the 1990s by Grażyna Vetulani (see Vetulani G. 2000; 2012), and with Gaston Gross's independent research on the category of object classes (fr. *classes d'objets*) (see e.g., Gross, G., 1994).

The first versions of the PolNet database, made available to a limited extent before 2012, included mainly nouns and the most important verbs. It was also during this period that *verb-noun collocations* began to be included in the PolNet database. The addition of simple and complex verbs (verb-noun collocations) along with syntactic information was the first step towards giving the PolNet lexical database the character of Lexicon-Grammar (as understood by Maurice Gross and Kazimierz Polański).

What follows is a (simplified) example of a noun synset (code).[26]

```
<SYNSET>
   <ID>PL_PK-28557</ID>
   <POS>n</POS>
   <DEF>drobniutkie, sproszkowane ziarenka ziemi, piasku, różnego rodzaju
    rozkruszonych lub bardzo rozdrobnionych ciał, unoszące się w powietrzu
    i osiadające na powierzchni przedmiotów; kurz</DEF>
   <SYNONYM>
   <LITERAL lnote="U1" sense="1">pył</LITERAL>
   <LITERAL lnote="U3" sense="3">proch</LITERAL>
   <LITERAL lnote="U1a" sense="1">pyłek</LITERAL>
   </SYNONYM>
   <USAGE>Cząsteczki pyłu wirują w powietrzu.</USAGE>
   <USAGE>Po wyburzeniu kamienicy wszystko spowijał pył.</USAGE>
   <USAGE>Pył cementowy, wapienny, krzemowy, węglowy, azbestowy.
   </USAGE>
   <USAGE>Pył śnieżny, wodny, pustynny.</USAGE>
   <USAGE>Tumany, kłęby pyłu.</USAGE>
   <USAGE>Pył opada, osiada, wznosi się, wciska się w usta.</USAGE>
   <USAGE>Zetrzeć z czegoś pył.</USAGE>
   <USAGE>Otrzepać, otrząsnąć, omieść coś z pyłu.</USAGE>
   <ILR type="hypernym">POL-2141601944</ILR>
   <SNOTE>--kurz</SNOTE>
   <SNOTE>--próchno</SNOTE>
   <BCS></BCS>
```

[26] An example of a noun synset from: (Vetulani Z. et al., 2010), p. 192.

```
    <NL>false</NL>
    <STAMP>przemekr 2007-06-14 18:34:21</STAMP>
    <CREATED>przemekr 2007-06-14 18:34:21</CREATED>
</SYNSET>
```

Synset description (simplified):

Synset (set of synonyms)	{pył1,pyłek1a,proch3}
Synset ID	PL_PK-28557
	% *pył1* oznacza *słowo ‘pył’* w pierwszym znaczeniu słownikowym, % *pył1a* oznacza zdrobnienie dla *pył1* % *proch3* oznacza *słowo ‘proch’* w jego trzecim znaczeniu słownikowym
Definition	Drobniutkie, sproszkowane ziarenka ziemi, piasku, różnego rodzaju rozkruszonych lub bardzo rozdrobnionych ciał, unoszące się w powietrzu i osiadające na powierzchni przedmiotów; kurz.
Use example	Cząsteczki pyłu wirują w powietrzu.
Use example	Z daleka widać było tumany, kłęby pyłu.
Use example	Po wyburzeniu kamienicy wszystko spowijał cementowy pył.
Hypernim ID	POL-2141601944

Usefulness of WordNet Lexical Networks for IT Application Development

The usability of the PolNet network as a lexical ontology in specific applications (e.g., in AI systems with language competence) is primarily determined by the properties of the concepts of *synonymy* and *hyperonymy*, as well as the features of *lexical coverage* (more on the prospects for the development of lexical ontologies of the PolNet/Lexicon-Grammar VerbNet type later in the article).

From PolNet 1.0 to Lexicon-Grammar VerbNets

The extent to which WordNet lexical networks will be useful in IT applications is determined by the properties of the concepts used in defining these networks. These concepts include, above all, the notion of *synonymy*, as well as *relations* defined on synsets. Of the latter, the relation of *hyperonymy* between the synsets representing particular concepts is the most important. Hyperonymy plays the role of the backbone for organizing the structure of the synset network. In the network, synsets can also enter into relationships with entities other than synsets, e.g., with attribute values or metadata.

Already the first attempts to extend the PolNet system with simple and complex verbs prompted us to in-depth reflection on synonymy and homonymy. The aim was to propose definitions that would correspond to the intuitive understanding of these concepts by linguists and at the same time be of a procedural nature, facilitating the writing of algorithms for creating synsets and extending the homonymy/hyperonymy relationship for the purposes of knowledge management using mechanisms of inheritance of the features of objects represented by synsets.

In order for the search for appropriate solutions to be tested on the basis of language material in applications of a practical nature (on a real scale), it was first necessary to supplement those language resources that were used to complete the first stage described in section WordNet Like Lexical Ontologies, as well as to acquire or create new resources. In this respect, a pioneering task was the development of dictionaries of predicative nouns and verbal-nominal collocations (Vetulani, G., 2000; 2012), as well as the proposal of a model for encoding and implementing grammatical information assigned to verb synsets for collocations. The most important of these tasks are listed in section Synonymy, Hyperonymy and Inheritance (see also Vetulani, Z. et al., 2010).

New Inspirations

Kazimierz Polański (1929–2009), parallel to Maurice Gross (1934–2001), was a precursor of the idea of Lexicon-Grammar. In his formalized dictionary of Polish verbs Polański includes entries with morphological, syntactic, and semantic information related to the chosen word form, which is also the ID of the entry (Polański, 1976; 1980–1992). The dictionary was developed and published in the years 1980–1992 and included 7,000 entries for the most important Polish verbs.

At the same time and independently of Polański, Maurice Gross was working on the formal description of the French verbs. Gross's concept is similar to Polański's in that the word form of the verb is directly related to the relevant lexical and semantic information. Gross held the opinion that the determinants of the meaning of words are elementary sentences characterizing their typical uses. Both of the above approaches are also found in the idea of the WordNet lexical network implemented under the direction of Charles Miller at Princeton University and organized around the concept of *synonymy*, which makes WordNet legitimately considered a lexical ontology.

All three approaches in the initial phase were implemented independently for significantly different languages: English, French and Polish (in alphabetical order). These languages are characterized by a different grammatical, and, in particular, dictionary tradition, which (probably) explains the fact that the initial research was not mutually quoted.

An important reason for the wide-spread adoption of these ideas is their significant application potential, supported by insightful theoretical work aimed at strengthening the lexical and grammatical coverage of important data repositories and tools needed for the development of language engineering (including multilingual aspects).

The EU-funded EuroWordNet project led by Piek Vossen (Vossen, 2002) went in this direction. The excellent theoretical documentation of EuroWordNet, was an important source of inspiration for the Lexicon-Grammar Verbnet for Polish.

Creating Lexical and Grammatical Resources and Their Digitization: Predicative Nouns and Verb-Noun Collocations

The Need for Lexical and Grammatical Resources

Initial work on the PolNet system was motivated by the desire to obtain ontologies sufficient to meet the basic needs[27] in the field of knowledge representation. At this stage, an ontology that well reflects the conceptualization typical of the language that people use every day seemed to be sufficient. Thus, in the initial period, limiting our work to the noun category was justified. However, this state turned out to be insufficient when there was a need to represent knowledge about situations, states, and events in AI systems, typically expressed in language

[27] The need for large lexical resources was not significant in the initial period of work discussed in Section Early Works, because until the end of the 1980s in Poland there were no favorable conditions for practical work in the field of natural language technology.

by predicative-argument structures, inspired by computer logic and knowledge engineering. Hence the need to extend PolNet with language constructions used to express relational content.

The basic lexical categories for this role are simple (one-word in Polish) or complex verbs of various types (see Vetulani, G., 2000). Among the latter, in Polish and a number of other languages, the most important are verb-noun collocations, composed of a supporting verb and a predicative noun, belonging to the category of abstract nouns. The support verb (Vsup) primarily plays a syntactic role, but sometimes also a different one (e.g., pragmatic or sociolinguistic), while the predicative noun (Npred) is associated with syntactic and semantic attributes. The latter are organized in valency structures expressing (through the attribute values) requirements or constraints of connectivity with arguments in a sentence structure. (More on the predicate-argument model in Vetulani, Z., 1998 and 2004 and Karolak, 1984).

Since the class of verb-nominal constructions is much more flexible and evolutionarily open than the class of simple verbs, we considered it reasonable to treat this class as a priority in the development of the PolNet database. The first step was to develop a careful methodology for recognizing the use of a compound structure as a verb-noun collocation acting as the center of a sentence. We will devote the rest of section Valency Structures in PolNet Lexical Ontologies to the acquisition of verbal-nominal collocations.

At the beginning of our research in this field, we focused on the description of the noun capable of playing the role of predicate in the verb-noun construction (Vetulani, G., 2000). In the 1970s and 1980s, the first important work on the predicative noun in the verb-noun construction appeared in the French literature; see (Giry-Schneider, 1978), (Danlos, 1980), (Vivès, 1983), (Gross, G., 1987). It is customary for a predicative noun (Npred) to appear in an analytic construction, partly fixed (frozen), forming with its accompanying verb (Vsup) a verb-noun collocation (Vsup + Npred) which plays the role of the sentence center (in simple sentences). The first tangible result of implementing the assumptions described above was the development of a digital dictionary covering over 14,600 Polish verb-noun collocations that can act as a predicate in a sentence (Vetulani, G., 2012).

Valency Structures in PolNet Lexical Ontologies

When the main goal of the first phase of the PolNet project, which ended with the implementation of PolNet v1, was to develop noun synsets and the relations between them (induced by semantic relations between the elements of synsets), the extension of PolNet to *verb synsets*, and more generally *predicative synsets*, posed a significant challenge that forced redefinition of the concept of synonymy. The modification required the introduction of relations aimed at enabling the formulation of conditions of connectivity between verb and noun synsets (representing predicate and arguments, respectively). This function is played by valency structures.

The key to making the right decisions regarding the development of linguistic ontologies for grammatical categories other than abstract nouns is to follow the idea already successfully validated for nouns at the stage of PolNet v1 implementation. What we mean here is that the structure of a formal ontology is consistent with the natural categorization of knowledge, so that Gruber's[28] maxim, which has worked for noun ontologies, does not lose its validity for other grammatical categories. This was the direction of research by a number of linguists working on the formal description of the semantics of natural languages.

In this field the most active among Polish linguists was Kazimierz Polański, while for other languages, pioneering research was conducted by Charles Fillmore, Martha Palmer (English), Maurice Gross and Gaston Gross (French), Piek Vossen (Dutch) and others. It is important to take account of the work of mathematicians and logicians such as Gottlob Frege, Alfred Tarski, Richard Montague and Kazimierz Ajdukiewicz, who had an essential impact on the formation of the model of thinking about natural language in the pre-informatics period.

Grouping together verb synsets and noun synsets according to the semantic-syntactic connectivity constraints imposed by the argument positions opened in a sentence by the predicate gives the PolNet system the status of a lexicon-grammar. The key to extending synonymy to predicative phrases (simple predicative verbs, predicative nouns, verbal-nominal predicative collocations and other grammatical categories) in such a way as to respect compatibility with the idea of Lexicon-Grammar is the concept of valency structure (Vetulani Z. and Vetulani, G., 2014).

By *valency structure* we mean here information about all argument positions opened by a predicative word, taking into account both semantic constraints on arguments, as well as, morpho-syntactic constraints on text elements filling these positions (case, number, gender, etc.) (Vetulani, Z. and Vetulani, G., 2015). In particular, we require synonyms to have the same valency structure and the same assignments of *semantic role values*. Thus, the valency structure is one of the formal exponents of *meaning* and, *ipso facto*, imposes strong granularity constraints on the synonymy of predicative expressions.

[28] "An ontology is an explicit specification of a conceptualization" (Gruber, 1993, p. 199).

Extending the Synset Definition.

Traditional descriptions of the vocabulary of natural languages generally distinguish words in terms of the *meaning* that is assigned to them. By *meaning* we understand the reference of a word[29] by the user to reality (real or fictional). This reference may associate a word with an object, a class of objects, or a bundle of relevant semantic features.

From the point of view of knowledge representation, particular importance is attached to the noun and verb categories. Both of these categories are composed of simple and complex forms. Typical *meanings of nouns* are entities (physical or abstract) or their descriptions. Typical *meanings of verbs* (simple or compound) are relationships between entities (physical or abstract), as well as states and events relating to entities (as well as other states, events, etc.).

Extension of the dictionary with synsets containing predicative expressions (predicative verbs and nouns, predicative collocations, etc.) is done with the use of predicative uses obtained from text corpora. Analysis of usage contexts provided the necessary syntactic and semantic information used for further work.

As in the lexicographical tradition (traditional dictionaries), during the development of the current versions of the PolNet system, examples were grouped to illustrate related uses, but shoving differences in surface implementation (see, for example, Polański, 1980–1992). For verbs and other predicative expressions, the key property defining their meaning (in the above sense) is the way they function in the structure of the sentence, in which they play a central (organizing) role (according to the widespread opinion of many linguists).

The description of the function of the entry in the structure of the sentence specifies the conditions of connectivity between the predicative expression (verb) and noun groups (as arguments). Connectivity conditions are described in the *valency structure*, which consists of appropriately selected *syntactic frames* obtained from the analysis of empirical material (corpus).

Implementation of Valency Structures.

The (simplified) example below is supposed to give a rough idea of the implementation of simple valency structures for predicative words. This is a code generated for the synset that includes four synonyms with a common meaning that may be translated to English as *to help.* This synset is composed of two Polish predicative simple verbs (Vpred) (*pomóc* and *pomagać*) and two (predicative) verb-noun collocations (Vsup+Npred) (*udzielić pomocy* and *udzielać pomocy*). (Each of the two

[29] Or other linguistic signs.

pairs contains the perfective and the imperfective forms, *pomóc-udzielić pomocy* and *pomagać-udzielać pomocy*, respectively.)

VALENCY STRUCTURE IMPLEMENTATION EXAMPLE (simplified)

The code below is a simplified representation of a verb synset (in the DEBVisDic system notation) containing a simple verb (*to help* in perfective and imperfective forms) and its synonym in the form of a collocation (help in perfective and imperfective forms).

POS: v ID: 3441
Synonyms: {pomóc:1, pomagać:1, **udzielić pomocy**:1, **udzielać pomocy**:1} (*to help*)
Definition: "wziąć lub brać udział w pracy jakiejś osoby (zwykle razem z nią), aby ułatwić jej tę pracę"(*"to participate in sb's work in order to help him/her"*)
Frame: Agent(N)_Benef(D)
Frame: Agent(N)_Benef(D) Action('w'+NA(L))
Frame: Agent(N)_Benef(D) Manner
Frame: Agent(N)_Benef(D) Action('w'+NA(L)) Manner
Usage: Agent(N)_Benef(D); "Pomogłem jej." (*I helped her*)
Usage: Agent(N)_Benef(D) Action('w'+NA(L)); "Pomogłem jej w robieniu lekcji." (*I helped her in doing homework*)
Usage: Agent(N)_Benef(D) Manner Action('w'+NA(L)); "Chętnie udzieliłem jej pomocy w lekcjach." (*I helped her willingly doing her homework*)
Usage: Agent(N)_Benef(D) Manner; "Ja chętnie jej pomagałem." (*I used to help her willingly*)
Semantic_role: [Agent] {człek:1, człowiek:1, homo sapiens:1, istota ludzka:1, … } (*{man,human,…}*)
Semantic_role: [Benef] {człek:1, człowiek:1, homo sapiens:1, istota ludzka:1, … } (*{man,human…}*)
Semantic_role: [Action] {czynność:1} (*{activity}*)
Semantic_role: [Manner] {ADVERB_FEATURE_QUALITY} (*an upper-ontology concept*

Legend
The row "Synonyms" – members of the synset identified as "POS: v ID: 344" (POS stands for "Part Of Speech")

Rows “Frames” – rows representing *semantic-syntactic frames* assigning attributes to argument positions / (slots) (in the basic /canonical/ surface order of a sentence)
The row “Definition” – (non-formalized) definition of the term corresponding to the synset /in Polish and English/
Rows “Usage” – typical examples corresponding to the intended meaning of the synset
Rows “Semantic_role” – semantic role values (synsets or upper ontology concepts)

In the example above, the rows containing predicate-argument schemes of sentences (tagged as ‘frames’) are *compatible* with each other. We talk about the compatibility of predicative-argument schemes in a valence structure when any two rows of this structure tagged as frames have a common extension that is a frame in this structure. (For example, for ‘Agent(N)_Benef(D) Manner’ and ‘Agent(N)_Benef(D) Action(‘w’+NA(L))’ such common extension is ‘Agent(N)_Benef(D) Action(‘w’+NA(L)) Manner’.) In sentences that match this extension, such as ‘Ja chętnie pomogłem jej w pracy’ /’I gladly helped her in work’/, one can distinguish sub-sentences, as for example, ‘Ja chętnie pomogłem jej’ /’I gladly helped her’/ and ‘Ja pomogłem jej w pracy’ /’I helped her in work’/, corresponding to the respective sub-frames. In subsequent work, we will (primarily) use such valency structures in which all predicative-argument schemes are mutually compatible.

Intra-Synset Variations.

Adoption of the concept of *meaning* as the starting point for the construction of lexical ontologies has important consequences for the utility aspects of the use of ontologies in engineering practice. The critical point is the *scope* of the commonly used, and thus vague, term *meaning*.

The carriers of intra-synset distinctions are often, but not exclusively, support verbs (Vsup) in predicative compound constructions, such as verb-noun collocations (Vsup+Npred).

Vsup plays an important role in the interpretation of complex predicative structures because:

1) in many cases it allows one to abolish the polysemy of the predicative form (here: a predicative noun) which is important from the point of view of applications, and also
2) brings information about register and aspect.

Grażyna Vetulani in her recent work on meaning-related aspects of support verbs (Vsup), exhaustively analyzes the role of Vsup in complex predicative expressions. She observes that despite its apparently subordinate non-predicative role, the support verb brings important semantic and grammatical information to the meaning of the whole predicative expression. In particular support verbs often serve to determine *register* and *aspect* of a collocation and, *ipso facto*, to concretize the *meaning* of the predicative noun (Npred) of the collocation (Vetulani, G., 2022).

Valency Structures in Lexicon Grammar VerbNets.
Storing the valency structure together with synsets – as part of their description – brings a number of benefits from the point of view of use in NLP applications (parsing, computer understanding, text processing) and is consistent with the idea and practice of Lexicon-Grammar as a tool dedicated to the broadly understood area of Language Technology for real utility applications (Gross, M., 1979).

An example of using a lexicon-grammar for the Polish language (built on the basis of PolNet v3) is the prototype of the POLINT-112-SMS system intended to support information management and decision making in emergency situations (see e.g., Vetulani et al., 2010; Vetulani & Marciniak, 2011; Vetulani & Osiński, 2017). The system is able to interpret SMS texts messages, as well as understand and process information provided by the human user.

POLINT-112-SMS has also proven itself to be an environment for testing the usefulness of grammar lexicons in the creation of utility applications. In particular, easy access to valency information facilitated the creation of simple heuristics allowing for effective (smart) search space reduction in syntactic and semantic analysis (parsing). This feature enables computationally cheap creation and testing of prototypes of utility systems or their replaceable modules, as well as the development of systems with multilingual competence.

Verb-Noun Collocation Gathering.
The family of verb-noun collocations is (in Polish and many Indo-European languages) an important group of compound verbs typically built of 1) an abstract predicative noun (Nsup) with a semantic and semantic-syntactic function, and 2) a support verb (Vsup), the main role of which is to introduce the predicative component (e.g., Nsup) and (often) to convey the pragmatic aspects (Vetulani, G., 2022). In some cases, the support verb is omitted from the surface structure (ellipsis). The semantic-syntactic function is primarily realized by the valency structure which fixes the conditions for the connection of the predicate with the arguments. In contrast to simple words, both nouns and (even more so) verbs, compound words are less well described than single-word forms for most languages. This is largely due to the scarcity of empirical research

based on representative corpora of texts, whether in written or spoken form. This circumstance is particularly important in the design and implementation of IT applications.

Digital Dictionary of Verb-Noun Collocations (Algorithm and Examples of Collocation Dictionary Entries).

The dictionary developed by Grażyna Vetulani (2012) is the result of the IPI PAN corpus exploration (Przepiórkowski, 2004). The following five step algorithm served the lexicographers to set up the last public release[30] of the version of the lexicon.[31]

Algorithm.

The 5-steps algorithm of corpus exploration and description includes the following steps:

- Step 1. Extraction from the corpus of the contexts with a high probability of containing verb-noun collocations, as well as detection of verb-candidates to be qualified as support verbs (automatically);
- Step 2. Manual analysis by lexicographers of the list of verb candidates obtained in the Step 1 in order to eliminate the evidently bad choices;
- Step 3. Automatic extraction of contexts in the form of concordances containing *verb-noun* pairs (selected through steps 1–2) as concordance centers;
- Step 4. Reading of the concordances by lexicographers, qualification of verb-noun pairs as collocations and providing their morpho-syntactic descriptions (manual);
- Step 5. Verification and final formatting.

The method used permitted the reduction (approx. 100 times) of the processing cost (estimation on the 5% sample).

We decided to put in the collocation lexicon all verb-noun collocations found in the corpus, because the inclusion in an electronic dictionary of a huge number of items is not a problem, as it would be for a traditional one. This also means that the lexicon contains, together with the well-known and currently commonly used collocations, a large number of less frequently used ones. The verb-noun lexicon demonstrates the dynamic and open character of the domain of nominal predication in Polish.

[30] The 2012 edition includes, after their examination and supplementation based on the IPI PAN Corpus, collocations collected until 2000.

[31] The carriers of intra-synset distinctions are often, but not exclusively, auxiliary verbs (Vsup) in complex predicative constructions such as verb-noun collocations (Vsup+Npred).

Various stages of work on the lexicon of verb-noun collocations were described in (Vetulani, G., 2000) and (Vetulani, G., 2012) and with the 2012 publication the lexicon was made available in digital form. The dictionary resource obtained was used in the implementation of valency structures in PolNet v3. For this reason, PolNet v3 may be considered the *first mature version* of Lexicon-Grammar Verbnet (Vetulani, Z. & Vetulani, G., 2016).

Collocation Dictionary Entries – Examples.
Verb-noun dictionary entries. Extract from the dictionary (Vetulani, G., 2012).

=>agresja, ż
czuć agresję/ czuć(B)/N1_do(D);wobec(D);w stosunku do(D),
dokonać agresji/ dokonać(D)/N1_na(Ms),
dokonać agresji/ dokonać aktu(D)/N1_na(Ms),
dokonywać agresji/ dokonywać(D)/N1_na(Ms),
dopuścić się agresji/ dopuścić się(D)/N1_na(Ms),
dopuszczać się agresji/ dopuszczać się(D)/N1_na(Ms),
doświadczać agresji/ doświadczać(D)/N1_ze strony(D),
doświadczyć agresji/ doświadczyć(D)/N1_ze strony(D),
kierować agresję/ kierować(B)/N1_przeciw(C),
odczuwać agresję/ odczuwać(B)/N1_do(D);wobec(D);w stosunku do(D),
popełnić agresję/ popełnić(B)/N1_wobec(D),
przejawiać agresję/ przejawiać(B)/N1_wobec(D),
przejawić agresję/ przejawić(B)/N1_wobec(D),
reagować agresją/ reagować(N)/N1_wobec(D),
zareagować agresją/ zareagować(N)/N1_wobec(D),
skierować agresję/ skierować(B)/N1_przeciw(C),
wybuchać agresją/ wybuchać(N)/N1_wobec(C),
wybuchnąć agresją/ wybuchnąć(N)/N1_wobec(C),
wykazać agresję/ wykazać(B)/N1_wobec(D),
wykazywać agresję/ wykazywać(B)/N1_wobec(D),
zareagować agresją/ zareagować(N)/N1_wobec(D).

Examples of selected contexts confirming the use of collocations in the text corpus

*** **dokonać**	
po tym jak podobnie jak swego	[* dokonało_ono_**agresji** *] na Kuwejt,
ostrzegając: jeżeli ktoś gdy	[* dokonałby_**agresji** *] na Polskę w czasie,
Wprowadzając stan wojenny, siłowy sposób	[* dokonano_**agresji** *] w brutalny, bo
*** **dopuścić**	
dzieckiem ojcu, który przemoc wobec matki	[* dopuścił_się_**agresji** *] i stosował
chcieli przecież nie	[* dopuścić_do_takiej_**agresji** *]
*** **doświadczać**	
w swej historii	[* doświadczała_obcej_**agresji** *].
*** **reagować**	
czynności fizjologicznych, nawiązania	[* reagowało_**agresją** *] i krzykiem na próby
*** **skierować**	
w Hucie Jedność prezydentowi miasta	[* skierowali_swoją_**agresję** *] przeciwko
*** **wybuchać**	
niezadowolona z siebie,	[* wybuchała_**agresją** *].
*** **wykazać**	
To nie policja sowali bezprawne	[* wykazał_**agresję** *], to związkowcy zasto-
*** **wykazywać**	
obserwowany, to znaczy wobec osoby słabszej, a	[* wykazuje_dużo_**agresji** *], brutalności
zgromadzenia, którzy	[* wykazywali_szczególną_**agresję** *].
przez okno albo	[* wykazuje_**agresję** *] wobec innego dziecka.
na inne dziecko, z okna	[* wykazuje_**agresję** *] albo chce wyskoczyć

Final Comments

The transition from the PolNet v1 phase to PolNet v2 was a significant step towards the Lexicon-Grammar, when the concept of the *valency structure* was launched. Starting from PolNet v2, *valency structures* were used in PolNet systems as the basic exponent of the meaning of *collections of predicative expressions* (simple or complex) organized in synsets. The requirement of *mutual compatibility of syntactic patterns for all elements of the synset* adopted for PolNet determines that the valency structure is *ipso facto* a determinant of the *meaning* of the predicative synset. The work currently being carried out aims to significantly increase the lexical and linguistic coverage of the class of complex predicative expressions, as well as to expand the scope of research covering the pragmatic layer of the Polish language.

Positive results in terms of practical usefulness of the lexical ontology model[32] indicate directions of natural continuation of previous work. These will be:

- at the grammatical description level: work covering the syntactic and semantic levels corresponding primarily to the needs generated by emerging application perspectives; this work will require further acquisition of empirical data from representative corpora certifying the use of units,
- at the pragmatic level (currently in the initial phase): extension of the model with new factors that may allow the internal structure of synsets (intra-synset relations) to be taken into account,
- at the tool level: development and implementation (or adaptation of existing ones) of the most effective systems collecting the necessary empirical data.

References

Colmerauer A., & Kittredge, R. (1982). ORBIS, *9th International Conference on Computational Linguistics*, *COLING*.

Danlos, L. (1980). *Représentation d'informations linguistiques : constructions N être Prép X*. Thèse de 3 cycle. Paris, L.A.D.L., Université Paris VII.

Dubisz, St. (Ed.). (2006). *Uniwersalny słownik języka polskiego PWN* [Universal Dictionary of Polish; PWN], 2nd edition. Wydawnictwo Naukowe PWN.

Fillmore, Ch., Baker, C. F., & Sato, H. (2002). Seeing arguments through transparent structures. *Third International Conference on Language Resources and Evaluation, Proceedings*, Vol. III (pp. 787–791). ELRA.

Giry-Schneider, J. (1978). *Les nominalisations en français : l'opérateur « faire » dans le lexique*. Librairie DROZ.

Godfrey, J. J., & Zampolli, A. (1997). Language resources. Overview. In R. Cole (Ed.), *Survey of the state of the art in human language technology* (pp. 381–408). Cambridge University Press.

[32] The POLINT-112-SMS system was used as a platform for testing the practical usability of wordnet-type lexical ontologies for implementing IT applications.

Gross, G. (1987). *Les constructions converses en français*. Librairie DROZ.

Gross, G. (1994). Classes d'objets et description des verbes. *Langages, 15*, 15–30.

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

Gross, M. (1979). On the failure of generative grammar. *Language, 55*(4) (Dec. 1979), 859–885.

Gross, M. (1981). Les bases empiriques de la notion de prédicat sémantique. *Langages, 63*, 7–52.

Gruber, Th. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*(2), 199–220.

Karolak, S. (1984). Składnia wyrażeń predykatywnych. In Z. Topolińska (Ed.), *Gramatyka współczesnego języka polskiego. Składnia* (pp. 11–30). Wydawnictwo Naukowe PWN.

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

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM,38* (11), 39–41.

Miller, G. A., & Fellbaum, Ch. (2007). WordNet then and now. *Lang. Resour. Evaluation, 41*(2), 209–214.

Pala, K., Horák, A., Rambousek, A., Vetulani, Z., Konieczka, P., Marciniak, J., Obrębski, T., Rzepecki, P., & Walkowska, J. (2007). DEB Platform tools for effective development of WordNets in application to PolNet. 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* (pp. 514–518). Wyd. Poznańskie.

Palmer, M. (2009). Semlink: Linking PropBank, VerbNet and FrameNet. *Proceedings of the Generative Lexicon Conference.* Sept. 2009, Pisa, Italy. GenLex-09.

Polański, K. (1976). Słownik syntaktyczno-generatywny czasowników polskich, zeszyt próbny. *Prace Naukowe Uniwersytetu Śląskiego, 124.* Wydawnictwo Uniwersytetu Śląskiego.

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

Przepiórkowski, A. (2004) *Korpus IPI PAN. Wersja wstępna*. Instytut Podstaw Informatyki PAN.

Szymczak, M. (1983–1985). *Słownik Języka Polskiego PWN* [Dictionary of Polish Language, in Polish]. Państwowe Wydawnictwo Naukowe.

Vetulani, G. (2000). *Rzeczowniki predykatywne języka polskiego. W kierunku syntaktycznego słownika rzeczowników predykatywnych*. Wydawnictwo Naukowe UAM.

Vetulani, G. (2004). Le rôle du verbe dans le réseau dérivationnel des prédicats nominaux. *Studia Romanica Posnaniensia, XXXI*, 459–467.

Vetulani, G., Vetulani, Z, & Obrębski, T. (2008). Verb-Noun Collocation SyntLex Dictionary – Corpus-based approach. *Proc. of 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco* (pp. 1561–1564). ELRA.

Vetulani, G. (2012). *Kolokacje werbo-nominalne jako samodzielne jednostki języka. Syntaktyczny słownik kolokacji werbo-nominalnych języka polskiego na potrzeby zastosowań informatycznych. Część I.* Wydawnictwo Naukowe UAM.

Vetulani, G. (2022). L'apport du Vsup au tour prédicatif verbo-nominal en polonais. *Neophilologica, 34 (2022)*, 1–18. Wydawnictwo Uniwersytetu Śląskiego. https://doi.org/10.31261/NEO.2022.34.10.

Vetulani, Z. (1988). PROLOG Implementation of an access in Polish to a data base. *Studia z Automatyki, XII.* Państwowe Wydawnictwo Naukowe (PWN), 5–23.

Vetulani, Z. (1989). *Linguistic problems in the theory of man-machine communication in natural language. A study of consultative question answering dialogues. Empirical approach.* Brockmeyer.

Vetulani, Z. (1990). *Corpus of consultative dialogues. Experimentally collected source data for AI applications.* Wydawnictwo Naukowe UAM.

Vetulani, Z., Walczak, B., Obrębski, T., & Vetulani, G. (1998a). *Unambiguous coding of the inflection of Polish nouns and its application in the electronic dictionaries – Format POLEX*. Wydawnictwo Naukowe UAM.

Vetulani, Z., Martinek, J. Obrębski, T., & Vetulani, G. (1998b). *Dictionary based methods and tools for language engineering*. Wydawnictwo Naukowe UAM.

Vetulani, Z., Obrębski, T., & Vetulani, G. (2007). Towards a lexicon-grammar of Polish: Extraction of verbo-nominal collocations from corpora (pp. 267–268). In *Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference (FLAIRS-07)*, AAAI Press.

Vetulani, Z. (2004). *Komunikacja człowieka z maszyną. Komputerowe modelowanie kompetencji językowej.* Akademicka Oficyna Wydawnicza EXIT.

Vetulani, Z., Marcinak, J., Obrębski, J., 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* [Language resources and text processing technologies. POLINT-112-SMS as example of homeland security oriented application]. Wydawnictwo Naukowe UAM.

Vetulani, Z., & Marciniak, J. (2011). Natural language based communication between human users and the emergency center: POLINT-112-SMS (pp. 303–314). In Z. Vetulani (Ed.), *Human language technology. Challenges for computer science and linguistics. LTC 2009. Revised Selected Papers.* LNAI 6562. Springer-Verlag.

Vetulani, Z. (2014). (PolNt-Polish WordNet (pp. 408–416). In Z. Vetulani & J. Mariani (Eds.), *Human language technology. Challenges for computer science and linguistics. LTC 2011 Revised Selected Papers. LNAI 8387*. Springer-Verlag.

Vetulani, Z., & Vetulani, G. (2014). Through Wordnet to Lexicon Grammar. In F. K. Doa (Ed.), *Penser le lexique grammaire: perspectives actuelles* (pp. 531–543). Editions Honoré Champion.

Vetulani, Z., & Vetulani, G. (2015). Synonymie et granularité dans les bases lexicales du type WordNet. *Studia Romanica Posnaniensia, XLII* (1), 113–127.

Vetulani, Z., Vetulani, G., & Kochanowski, B. (2016). Recent advances in development of a lexicon-grammar of Polish: PolNet 3.0 (pp. 2851–2854). In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016). European Language Resources Association (ELRA), Paris, France.

Vetulani, Z., & Osiński, J. (2017). Intelligent information bypass for more efficient emergency management. *Computational Methods in Science and Technology*, *23*(2), 105–123. https://doi.org/10.12921/cmst.2017.0000019.

Vetulani, Z., Vetulani, G., & Mohanty, P. (2021). Development of real size IT systems with language competence as a challenge for a Less-Resourced Language: a methodological proposal for Indo-Aryan languages. *Journal of Information and Telecommunication, 5*(4), 514–535. https://doi.org/10.1080/24751839.2021.1966236.

Vivès, R. (1983). *Avoir, prendre, perdre : constructions à verbe support et extension aspectuelle*. Thèse de 3 cycle. Paris, L.A.D.L. et Université Paris-VIII.

Vossen, P. (Ed.) (2002). *EuroWordNet. General Document.Version 3*. University of Amsterdam.

Walkowska, J. (2012). *Modelowanie kompetencji dialogowej człowieka na potrzeby jej emulacji w zarządzających wiedzą systemach informatycznych współpracujących z wieloma użytkownikami*. PhD Dissertation at IPI PAN Warszawa (supervised by Z. Vetulani).

Zampolli, A. (1996). Współpraca międzynarodowa w dziedzinie LR [International cooperation in the field of Language Resources, in Polish]. *Informatyka, 3*, 34–37.

Web pages

http://alain.colmerauer.free.fr/alcol/ArchivesPublications/PrologHistory/19november92.pdf (Alain Colmerauer and Philippe Roussel (1992). The birth of Prolog. GIA, Marseille). Last access: 12.09.2023.

https://pl.wikipedia.org/wiki/Słownik_języka_polskiego_(Mieczysław_Szymczak). Last access: 15.12.2023.

https://www.jstor.org/stable/412748 (Maurice Gross (1979). On the Failure of Generative Grammar). Last access: 12.09.2023.

https://en.wikipedia.org/wiki/Giuseppe_Peano (Giuseppe Peano, 1889). Last access: 12.09.2023.

http://www.illc.uva.nl/EuroWordNet/docs/GeneralDocDOC.zip (Piek Vossen. (2002). EuroWordNet. Final Document). Last access: 12.09.2023.