ABSTRACT
In this paper, based on the example of our early works for Polish, we want to share our experience in the challenging task of developing NLP-based technologies in the situation of initial scarcity of digital language resources that ranked Polish among the Less-Resourced Languages. We present some of our projects aiming at language resources and tools we had to create in order to be able to process texts in Polish and develop real-scale systems with language understanding competence. The case study we present here is the rule-based system POLINT-112-SMS for improving information management in emergency situations. We argue in favour of the lexicon-grammar approach to the formal description of inflecting languages and present our current work on this grammatical paradigm. Our current work is on the implementation of the ideas presented in the first part of the paper on three prominent Indian languages, that is, Hindi, Odia, and Bengali.

1. Introduction

Using the example of our early works on Polish, we would like to share our experience of the last three decades in the challenging task of developing NLP-based technologies in the situation of initial shortage of digital language resources. Our objective behind this overview is to contribute to the discussion of possible solutions of scarcity of resources and tools for the low-resourced languages.

For a long time, that is, from the fifteenth till the eighteenth centuries, Polish was a lingua franca over a large part of Central and Eastern Europe. During the Renaissance and Baroque periods there was intensive literary production in Poland followed by the first modern dictionaries and grammars. Still until recently descriptions of the Polish language were addressed to human users for teaching and translation purposes. For this reason, they were of little use to the language industries because of lack of precision.
Polish was considered as a ‘less-resourced-language’ at least until the end of the twentieth century. The negative effects of this deficiency may still be observed. The awareness actions undertaken by the EU followed by appropriate funding measures in the 1990s were important milestones in the development of human language technologies in Central Europe. The new bridges emerging between the technologically advanced research communities and the newcomers often stimulated the appearance of new and original ideas. International grants bringing together partners from various language communities representing different linguistic traditions were often oriented to the creation of basic language resources and tools in order to serve the dynamically growing language industries in Europe including Poland. The general condition of language technologies for Polish is now good and efficient so that Polish is no longer classified as ‘less-resourced’. Nevertheless, there are still many gaps to fill.

This paper is organized into five sections, followed by conclusions and references. After the introduction and a brief reference to the beginnings of our research, we present an overview of some of our projects related to the language resources and tools we had to develop in order to be able to process texts in Polish and develop real-scale systems like POLINT-112-SMS (Section 3.3.3) with language understanding competence. In particular attention was given to electronic dictionaries, WordNet-inspired lexical ontologies, collocations, and the organization of grammatical data into a lexicon-grammar (Vetulani & Vetulani, 2020) and, last but not least, the role of human resources.

The first four sections are mainly concerned with the presentation of works on Polish language technologies aimed at applications in artificial intelligence systems. In Section 5, we draft the first steps towards applying the methodological model so far tested for Polish to the selected Indo-Aryan languages of India.

2. Early Works in Poland and in India

Among the first attempts to use computers for processing Polish texts and speech, the research conducted in the 1970s in Warsaw and Poznań was noteworthy (L. Bolc, S. Szpakowicz for text, W. Jassem, M. Steffen-Batóg for speech processing). In the late 1970s and early 1980s, the first attempts to implement systems that understood Polish were carried out (independently), for example, by W. Lubaszewski, St. Szpakowicz and Z. Vetulani.

Typical of these early works was the unavailability of real-size electronic language resources as well as of NLP dedicated tools for Polish. These early studies on NL parsing and logic-based question-answering were found inspiring two decades later when we started working on real-size applications (Vetulani, 2012). However, real-size applications also require real-scale resources.

The NLP activities started in India in a serious manner in the last quarter of the twentieth century. Realizing the importance of these activities for the development and growth of the country, the Government of India provided support to Indian universities and institutes (see Section 5.5 selection of the main outcomes of this involvement).

3. Basic resources for text processing

During our first attempts to design non-trivial rule-based question-answering systems with deep understanding (Vetulani, 1988) we identified the most urgent needs:
electronic dictionaries, computer processible grammars and corpora (for general language, but also for domain-specific applications). Several national and European projects contributed to partially fill the gaps in resources and tools.

3.1. Dictionary project POLEX (1994–1996)

Polish is a highly inflecting language and, therefore, it has a relatively free word order. Consequently, the simple adaptation of processing algorithms that are efficient for English or French appeared hard to apply. This is so, because the information concerning the function of a word in the sentence is typically encoded in the word form, independently of its linear position in the sentence. The evidence forced us to propose our own algorithms, for example, parsers (Vetulani, 1988) that require a precise description of Polish grammar. In the field of morphology, a huge amount of work by lexicographers until the 1990s did not lead to a standard description of Polish words that would not require the individual linguistic competence of users to interpret dictionaries. We proposed an unambiguous inflectional description system capable of eliminating the need for human linguistic competence to interpret word forms and, therefore, appropriate for the development of machine-processing software (Vetulani et al., 1998a). POLEX is a morphological dictionary of core Polish words of general interest included in a large traditional paper dictionary (Szymczak, 1995). It uses a precise machine-interpretable formalism (coding system), which is the same for all categories (classes of speech) (Vetulani et al., 1998a).

The dictionary entries are described as a list of four parameters: basic form, list of stems, paradigmatic inflection code, and distribution of stems. Two of these parameters, i.e. the paradigmatic inflection code and distribution of stems, describe how to associate endings to stems to obtain the required forms of the word (distribution associates stems to the paradigmatic positions).

The first public release of the resource contained (2007) over 42,000 nouns, 12,000 verbs, 15,000 adjectives, 25,000 participles, and about 200 pronouns.

3.2. First steps towards lexicon-grammars for Polish

In the 1970s, Maurice Gross (LADL, Paris 7) proposed a grammatical lexicon based on the idea of storing words together with possibly all relevant syntactic and semantic information (Gross, 1975). Lexicon-Grammars were developed first for French, then for other languages. Consequently, predicative words were studied from the point of view of their aptitude to form elementary sentences. Gross introduced the term lexicon grammar (Fr. lexique-grammaire) in the sense of the method of describing the meaning of predicative words by providing descriptions of how these words form simple sentences. What distinguishes lexicon grammars from traditional grammatical descriptions of a language is that lexicon-grammar entries contain an exhaustive grammatical descriptions (on syntactic and semantic levels) of well-defined senses of words. This property renders lexical-grammars well-suited for NLP applications.

*Numerous doctoral dissertations under the supervision of Maurice Gross were oriented towards languages other than French (his own doctoral dissertation was about comparing
French with English) ranging from other Romance languages, through Greek, Korean and Arabic. These works showed that Gross’s theoretical hypotheses are confirmed on typologically very diverse material, which gives them a strong conceptual basis.’ (Lamiroy, 2003)

3.2.1. The EUREKA Project GENELEX (1990–1994)
GENELEX was an initiative – which first addressed some Western-European languages, starting with English, French, German, and Italian – to realize the idea of lexicon-grammar in the form of a generic model for lexicons and to propose software tools for lexicon management (Antoni-Lay et al., 1994). The first of the two reasons given by Antoni-Lay for large-size lexicons was that when Natural Language applications left the research labs for the world of practical applications, they were required to cover a wide variety of language phenomena, and this implied real-scale resources such as electronic dictionaries, grammars and corpora. The second reason was the tendency observed already in the eighties (and even before) to put a large amount of grammatical information into a lexicon. This tendency finally led to grammatical systems with relatively few grammar rules and more complex lexicon instead (see (Gross, 1975), (Polański, 1992)).

Although Polish was not covered by the GENELEX project, it was studied in two EU projects, GRAMLEX and CEGLEX, aimed at testing the potential of the novel GENELEX-based LT solutions to some languages not represented in GENELEX.

3.2.2. PECO-COPERNICUS Project 1032 CEGLEX (1995–1996)
The main goal of the CEGLEX consortium (Vetulani et al., 1995) was to test the GENELEX proposal of a generic model for reusable lexicons for three official languages spoken in Central Europe: Czech, Hungarian and Polish.

The CEGLEX/GENELEX model claims to be theory-welcoming, complete, (i.e. to cover all relevant phenomena on three classical layers: morphological, syntactic, and semantic), as well as easily applicable to different languages. The three layers of the CEGLEX/GENELEX model were confronted with the data for languages under consideration with generally positive results, especially for Czech and Polish. In particular, the model had to be adapted to accept language data for the Polish language. On this occasion, some modifications were proposed concerning the representation of the inflection phenomena specific to the Polish language.

3.2.3. PECO-COPERNICUS Project 621 GRAMLEX (1995–1998)
The aim of GRAMLEX was to facilitate the initiation, coordination and standardization of the construction of morphological dictionary packages for the following European languages: French, Hungarian, Italian and Polish, including a detailed formal description of the morphology of the Polish language. The intention of the GRAMLEX tasks for Polish was to contribute to the improvement of the situation concerning language engineering tools and resources. Among the main achievements was a corpus-based morphological dictionary (for over 22,500 entries derived from POLEX) as well as the related tools and applications (lemmatizer, inflected form generator, concordance generator and others) in the GENELEX format (Vetulani et al., 1998b). The project GRAMLEX was related to the two projects mentioned above POLEX and CEGLEX lexemes.
3.3. Further steps towards lexicon grammars for Polish. PolNet 3.0 as lexicon-grammar

The IT applications with language competence we were able to develop until 1990 were all classified as toy-systems. This was caused by the lack of real-size digital, easily machine-processable electronic resources. This problem was addressed in the R&D grant ‘Text Processing Technologies for Homeland Security Purposes’.

3.3.1. POLINT – 112 – SMS: a system with natural language competence (2006–2010)

Within the above-mentioned grant, we created a prototype of the system that was designed to assist the monitoring of mass events and to enhance real-time identification of potentially dangerous situations in the crowd of fans in order to fix processes with a high degeneration risk (early prevention). This application required a powerful natural language competence to understand and process SMS messages exchanged between the security staff agents in the uncontrolled natural language (cf. Vetulani & Osiński, 2017). The understanding module of the system is rule-based because of the need to obtain a very precise representation of the utterance content that is crucial in processing sensitive information. Messages were supposed to be written in standard, correct and unconstrained Polish. Resources essential for this project, besides the machine-readable grammar and dictionaries, were ontologies for knowledge representation and logic-based reasoning and text and dialogues corpora to support the implementation of the language processing model of the application.

3.3.2. Corpora

Text and speech corpora have numerous applications in language processing. First of all, they constitute an empirical foundation on which to construct language models. The elementary application of a corpus consists in identifying and cataloguing language use phenomena important for further text processing (like parsing, understanding, etc.). Several corpora were used to implement the POLINT-112-SMS system. For purposes such as concepts acquisition for ontology building or creating frequency-based heuristics for language processing, it is common to apply large corpora. In our case, we used the open version of the Polish National Corpus (Przepiórkowski, 2004).

In the cases where project-specific linguistic information was necessary, we were obliged to collect the necessary corpora. Some of them were:

- the experimentally collected St. Claus Corpus of informative dialogues (Vetulani, 1989, 1990, Vetulani 2004),
- the corpus of 1700 private SMS messages (Vetulani et al., 2010; Walkowska et al., 2009),
- the corpus of experimental Role Playing Game dialogues (Vetulani et al., 2010)/1374 dialogues/,
- experimental Wizard_of_Oz – supported dialogs (Vetulani et al., 2010)/1198 dialogues/,
- a juridical text corpus (papers/23/, books/2/, documents/80/) (Vetulani et al., 2010),
- speech corpus – emergency line 997/12 (24 h of recording for 1818 calls, partially transcribed into text and tagged/58,935 words/).
3.3.3. PolNet – PolishWordnet as lexical ontology (since 2006)

The absence of lexical ontologies reflecting conceptualization typical of Polish speakers in the market inspired us to develop PolNet PolishWordnet – a lexical database of the type of PrincetonWordNet. We built it from scratch according to the so-called merge model methodology. Creation of PolNet started in 2006 and continues. The resource development procedure was based on the exploration of traditional dictionaries and the use of available language corpora such as the IPI PAN Corpus (Przepiórkowski, 2004). Incremental development of PolNet started with general and frequently used vocabulary. We decided to select the most widely used words found in a reference corpus of Polish language (Vetulani et al., 1997), with an important exception made for methodological reasons: even though we intended PolNet to be a resource of general interest, we assumed its possibly early validation in the real-size applications for which having a domain-specific terminology represented in the system was crucial.

By 2008, the initial PolNet version based on noun synsets related by hyponymy/hyperonymy relations was already large enough to serve as a core lexical ontology for real-size applications. However, in order to develop a POLINT-112-SMS system prototype, an extension of the core set of nouns with domain terminology was necessary. Further extension with verbs and collocations transformed PolNet into a lexicon-grammar (Vetulani & Vetulani, 2014a), bringing to PolNet various new possibilities, such as facilitating the implementation of parsers. By including the verb category in PolNet we brought ideas inspired by FrameNet (Fillmore et al., 2002) and VerbNet (Palmer, 2009) projects to PolNet. The verbal part was at the heart of the entire network (Vetulani & Obrębski, 2010). Its organizing backbone was the system of semantic roles (Palmer, 2009).

3.3.4. First step from PolNet to lexicon-grammar PolNet 1.0.

Already in the early 1980s, information typically contained in lexicon-grammar entries for predicative words, simple or compound, was considered useful for parsing and generating natural language sentences. Lexical entries used in the PROLOG code of the demonstration system ORBIS were, in fact, lexicon-grammar units describing syntactic and semantic valency of words. The syntactic/semantic valency (valency structure) appeared useful to avoid the acceptance of incorrect sentences by a parser and the production of incorrect sentences by a generation algorithm. It is also helpful to build error-correcting software. The qualitative evolution of PolNet, initially conceived as a lexical ontology, towards a lexicon-grammar of Polish took place between the release of PolNet 0.1 (2009) and PolNet 3.0 (2014). The pragmatic reason to substantially enrich PolNet was the need of an efficient parsing engine to support the understanding module. In addition to noun synsets that make of PolNet 0.1 (2009) a lexical ontology, we decided to enrich PolNet with verb synsets containing syntactic/semantic information in the form of valency structure. The valency structure of a predicative word provides the morpho-syntactic and semantic constraints on the acceptable fillers of the argument positions opened by this word (like case, number, gender, preposition, register, etc. for morpho-syntactic constraints and semantic roles (Palmer, 2009) like agent, patient, beneficiary, etc. for semantic ones). Later, we decided to refine our idea of verb synsets taking into account the granularity issues related to synonymy. We focused on verb–noun constraints corresponding to the particular argument positions of the verb. For our purposes, we proposed the
following definition: verb + meaning pairs are only synonymous when they take the same semantic roles and the same concepts as values (see more in (Vetulani & Vetulani, 2015)). In particular, the valency structure of a verb is one of the indices of meaning (i.e. all members of a synset share the valency structure). Verb synsets appeared already in the first public release of PolNet in 2011 (PolNet 1.0) (Vetulani et al., 2016). This extension opened a new generation of PolNet systems that we call now ‘PolNet – Polish Lexicon-Grammar systems’. The expansion of PolNet to the Lexicon Grammar of Polish was based on the results of research on predicative verbs assembled in the Dictionary of Polish Verbs (Polański 1980–1992) where linguistic descriptions were provided for 7000 Polish predicative verbs. The valency information permitted us to make a smart use of PolNet as it was enriched with lexicon-grammar features when implementing the POLINT-112-SMS system. In addition to using PolNet as a lexical ontology in the World Knowledge and Situation Analysis Modules, we made use of the valency information to enhance the efficiency of the parser as part of the Text Understanding Module. In this module, syntactic/semantic valency information stored in the lexicon-grammar rules was used to control parsing execution by heuristics in order to speed up parsing due to additional information gathered at the pre-analysis stage. The effect of substantially reducing the processing time was due to the reduction of the search space.

3.3.5. Collocations in PolNet 2.0 – PolNet 3.0
As the usefulness of PolNet as lexical ontology in the development of real-size systems was positively verified by the above-mentioned application, we decided to take the next steps towards the full lexicon-grammar. The versions PolNet 2.0 and PolNet 3.0 were important milestones in this process. The passage from PolNet 1.0 to PolNet 2.0 (Vetulani & Vetulani, 2014b) was marked by inclusion of a set of verb–noun collocations from the syntactic dictionary of verb–noun collocations (Vetulani, 2000, 2012) or directly from corpora described in (Vetulani et al., 2010). Verb–noun collocations, typically composed of a noun (as predicate) and a predicatively empty support verb (Vetulani, 2000), play in the sentence the same role as simple predicative verbs in other cases, as for example, for conduct an evaluation/evaluate or carry out an assessment/assess. However, such a connection does not always exist, and some verb-nominal collocations do not have an equivalent in the given language system in the form of a single (synthetic) predicate-verb, as it is the case in Polish for mieć nadzieję (to hope) and czynić dobro (to do good). It is, therefore, all the more advisable that such dictionary data (i.e. the entire class of verbo-nominal collocations) should be acquired and entered into dictionaries and to PolNet. As the logical centre of the sentence, collocations assign certain properties to the sentence arguments (subject and complements) and establish relationships between the subject and its complements. Thus, the role of predicative collocations in the lexicon-grammar is equally as important as the role of simple predicative verbs, especially when for purposes of further processing, information carried by predicative elements appears useful. Being predicatively empty, the support verb in Polish still plays a significant auxiliary function in the interpretation of the predicates. This is due to its possible emotional and metaphorical aspects. A support verb may provide information on the language register, aspect (perfective/imperfective), action mood (inchoative, terminative, progressive etc.) or pragmatic/situational meaning (as politeness or informality). The ‘predicatively empty’ support verb contributes to the sense of the collocation
and to its disambiguation. Despite its non-predicative function in the sentence, the support verb plays an important role in the semantic analysis of the whole sentence. In fact, it is impossible to determine the meaning of a noun predicate without a detailed analysis of the accompanying (support) verb.

As for the role of the support verb in the description of noun predicates:

- firstly, because the class of nouns is heterogeneous, particular problems arise when it is necessary to distinguish predicative words in this class. Supportive verbs play a fundamental role in distinguishing predicative nouns from among all nouns. The syntactic rule of necessity to use a predicate noun together with a verb devoid of its basic meaning, turns out to be helpful. Thus, one can see in a supporting verb the unifying factor for the entire class of predicative nouns, that is, both those predicates that have their equivalents belonging to other grammatical categories (such as, for example, in Polish *odnieść zwycięstwo*/**zwyciężyć** (in English *defete*)) and those that in a given language system are considered autonomous, that is, they do not have a synthetic verb equivalent.

- secondly, the support verb, in many cases, helps to avoid the polysemy of the predicative form (here: the predicative noun) – which is important in terms of applications. In different linguistic realizations (applications) the same lexical forms display different semantic features: for example, in Polish many concrete nouns [CONCR] take the abstract character [ABSTR] in some use and become a predicate (then they coexist with an support/auxiliary verb) – compare, for example, the noun ‘telefon’ as the name of a specific object and its abstract use with the expression ‘wykonał telefon’ (meaning ‘made a call’). Further: it turns out that the noun, being semantically marked as [ABSTR], most often remains ambiguous, and hence there is a need for its further (more precise) description. Referring to a verb that coexists with a noun predicate in many cases allows to recognize the meaning of the predicate unequivocally. In other words, in order to eliminate the polysemy of the predicative form, it is enough to use it with the appropriate support verbs (corresponding to different senses). We call these forms, together with the corresponding support verbs, ‘verbo-nominal collocations’.

- thirdly, the support verb is an element that specifies the meaning of the entire verb – nominal phrase. These are cases in which the replacement of a verb component does not change the basic meaning of the predicate, but makes this meaning concrete. For example, *nałożyć/nakładać cenę* [impose a price]; *określić/określać cenę* [set/define a price]; *ustalić/ustalać cenę* [set/fix a price], *wyznaczyć/wyznaczać cenę* [fix a price]. In all these cases, we are dealing with the same basic (deep) sense. Verbs that accompany the predicate ‘price’ bring the so-called semantic ‘surplus’ and their use depends on linguistic register (loose, official, etc.) or the type of discourse (specialist, scientific, etc.) or – in Polish (as in other Slavic languages) with a grammatical aspect (perfective/imperfective) (cf. Ewa Jędrzejko in: *Słownik polskich zwrotów werbo-nominalnych*. Zeszyt próbny /Dictionary of Polish verbo-nominal phrases. Sample book; in Polish/ (Jędrzejko, 1998:49)).

Verbal-nominal collocations based on the same predicative form do not have to be semantically related and may differ significantly in terms of use. Compare two classes of collocations of different meanings, both based on the same predicative form, the first of which is related to interpersonal contacts and the second is related to reporting.
The formal distinguishing feature that characterizes them is a set of support verbs as well as syntactic and semantic requirements that determine the function of individual collocations in a sentence.

| Class 1: | relacja – interpersonal contacts |
|---------|----------------------------------|
| być w relacji/ być w(Ms)/N1(D);MOD/N2(do(D), /* be in relationship with */ |
| nawiązać relację/ nawiązać(B)/N1z(N), /* establish a relationship with */ |
| pozostawać w relacji/ pozostawać w(Ms)/N1(D);MOD/N2(do(D), /* stay in relations with */ |
| układać relację/ układać(B,lmn)/MOD/N1z(N), /* arrange a relationship with */ |
| zachować relację/ zachować(B,lmn)/MOD/N1z(N), /* keep relationships */ |

| Class 2: | relacja – reporting |
|---------|---------------------|
| nadawać relację/ nadawać(B)/N1z(D), /* broadcast a report */ |
| przedstawić relację/ przedstawić(B)/N1(C)/N2z(D), /* present a report */ |
| przekazać relację/ przekazać(B)/N1(C)/N2z(D), /* provide a report from */ |
| zdać relację/ zdać(B)/N1(C);przed(N)/N2z(D), /* to report on */ |

The study of predicative collocations is equally important for language technology as for the traditional domains of language teaching and translation. This is because of the conventional nature of collocations, whose structure and lexical selection are unpredictable. This is one more reason to have all the varieties of predicative collocations included in the lexicon-grammar with as complete a description as possible. Adding verb–noun collocations to PolNet appeared a non-trivial task because of specific morpho-syntactic phenomena related to collocations such as syntactic synonymy24 (Vetulani et al., 2016), as well as the problem of (optimal) granularity of verbal synsets. In Vetulani and Vetulani, (2014b), we observed that granularity is directly related to synonymy which is the basis of the organization of the wordnet database in synsets, (i.e. synonyms should belong to the same synset; see above). Our approach differs from that of Miller and Fellbaum (see in (Miller et al. 1990) and (Vossen et al., 1998)), who postulate a very weak understanding of this concept (based on an invariability test with respect to just one linguistic context), often leading to very large synsets.

Version 3.0 of PolNet was meticulously cleaned and extended25 with respect to the Version 2.0. It has been user-tested as a resource for modelling semantic similarity between words (Kubis, 2015).

4. Last but not least: human resources
The possibility of gaining the status of a High-Resourced Language depends on the existence of a research community of experts qualified to create NLP resources and tools. For
Polish, this difficult requirement was satisfied from the very beginning due to the existence of well-qualified computer science manpower and very competent descriptive linguists. This fact made possible the creation of crucial language resources independently in several research centres throughout the country.\textsuperscript{26}

The increase in the volume of R&D work in Poland is correlated with the progress of Polish computational linguistics. The Language and Technology Conferences (LTC) organized in Poznań mobilized since 1995 over 1,360 authors from over 60 countries (see http://www.ltc.amu.edu.pl). For 25 years, LTC was for Polish authors a natural forum to present research about Polish. More than 300 LTC contributors representing national and private institutions of fundamental research, R&D groups and companies testify to the creation in Poland of a strong community of HLT developers. This means the promotion of Poland to the small number of countries with solid human capital in the area of Language Technologies who are well prepared to develop language technologies for their mother tongue.

5. Application of lexicon-grammar approach to Indian languages: case studies of Hindi, Odia, and Bengali

5.1. Current mid-term collaborative project

The intention behind the present collaboration between UAM (Research Group on Computational Linguistics) and the GLA University (Institute of Applied Sciences and Humanities) is to propose a generic model for the elaboration of theoretical bases for Language Industries (presented in this paper) for Indo-Aryan Languages. This model, developed and tested for the Polish language, is described in this article.

The long-term perspective of this cooperation aims at meeting the different needs of those languages of the Indian subcontinent that do not have the sufficiently required language resources and tools and, as a result, are Less-Resourced Languages.

India has 22 constitutionally recognized scheduled languages\textsuperscript{27}, including Sanskrit and Hindi, whereas it has two official languages, that is, Hindi and English. Out of these 22 scheduled languages, 15 are Indo-Aryan, 4 Dravidian, 2 Tibeto-Burman, and 1 Austroasiatic (Munda). It will be an excellent testing ground for validation of our approach on a big scale if we can apply it on some selected Indo-Aryan Languages. There are several reasons why this work should be considered an important contribution to the development of the still missing basic linguistic resources for several of these languages (middle-resourced languages, see Vetulani & Vetulani, 2020). Some of these reasons are:

(1) Polish is an important language of the Indo-European language family which may be considered as a bridge between the major modern West-European languages (all Romance and Germanic languages) and most of languages of India because of its syntactic and semantic properties.

(2) Over the last 25 years, the Polish language has advanced from the group of low-resourced languages to the group of technologically advanced languages with important language resources and tools.

(3) The choice of a group of Indian languages to test the usability of the model presented here results from several premises:
- Existence of numerous similarities on the morphological, syntactic and semantic level between the Polish language and the Indo-Aryan languages; these similarities are often absent in the case of the English language, which makes it difficult to apply the methodological patterns tested for the English language; the same is true for other important West-European languages, such as French and German.
- Existence of a well-trained workforce of experts – both linguists and language engineers – of these languages,

(4) The support of the Government of India on the basis of its sensitivity to the social and economic benefits of the development of the Language Industries.

5.2. Existing resources, tools and know-how for Hindi, Odia, and Bengali to reuse in current and forthcoming work

The NLP activities on Indian languages began much later than similar works on the languages of the Western world. Due to generous support and continuous encouragement from the Government of India, in a short span of time, a number of universities and institutes developed resources and created tools for various activities, such as WordNets, Indian Language to Indian Language Machine Translation Systems, Parallel Corpora, Shallow Parser Tools, Tree Banks, and monolingual as well as parallel corpora in specified domains in the constitutionally recognized languages. Some of these are available on the following sites:

- IndoWordNet: https://www.cfilt.iitb.ac.in/indowordnet/
- Indian Language to Indian Language Machine Translation: https://ltrc.iiit.ac.in/
- Indian Languages Corpora Initiative: http://sanskrit.jnu.ac.in/index.jsp
- Shallow Parser Tools in Indian Languages: http://calts.uohyd.ac.in/calts/tools.php

One can also find resources and tools on the following site of the Ministry of Electronics and Information Technology, Government of India: Technology Development for Indian Languages (TDIL): https://tdil.meity.gov.in. It must be noticed, however, that the impressive work that has been done so far does not cover all 22 local official languages, nor does it concern the majority of the country’s population. In any case, the obtained tools, resources, and acquired know-how can be used and further developed for more effective and efficient applications on the selected Less-Resourced Indian languages. Some important publications that deal with various aspects of these languages in the pan Indian context are: as follows: Gender-agreement in Oriya (Mohanty, 1989), Dravidian Substratum and Indo-Aryan Languages, International Journal of Dravidian Linguistics (Mohanty, 2008), WordNets for Indian Languages: Some Issues. (Mohanty, 2010), Oriya: A Confluence of Aryan, Dravidian and Munda (Mohanty, 2011), Brutt e mo poshe kuTumba: oDia: bha:sha:baigya:nika carca:ra nutana diganta (in Odia) (Mohanty, 2016), Issues in the Creation of Synsets in Odia WordNet (Mohanty, Malik, Bhol, et al., 2017).

5.3 Some specific properties of Hindi, Odia, and Bengali

According to the Sanskrit grammatical tradition, the verb plays the most important role in a sentence. Besides the simple verbs that have only a root along with the inflections,
conjunct and compound verbs\textsuperscript{28} are two important characteristic features of most Indian languages. The conjunct verbs have the $N + V$ or $\text{Adj.} + V$ structure, whereas the compound verbs consist of more than one verb. These compound verbs were almost absent in Sanskrit and are an innovation in the modern Indian languages (like Hindi, Odia and Bengali). The $V_1$ in a compound verb conveys the semantic content, whereas the following ones contribute to the grammatical aspects.

Let us take examples from Odia:

(1) rāma madhu-ku Tankā de-lā
Ram Madhu-to money gave

*Ram gave money to Madhu.*

(2) rāma madhu-ku sikhyā de-lā
Ram Madhu-to teaching gave

*Ram taught a lesson to Madhu.*

In sentence (1), the verb /de-lā/ ‘gave’ has two objects, that is, Madhu is the indirect object and /Tankā/ is the direct object. But in sentence (2), /sikhyā/ ‘teaching’ is not the direct object of the verb /delā/ ‘gave’; instead, the two-word expression /sikhyā delā/, as a whole, is a non-compositional verbal unit with ‘Madhu’ as its direct object. The fact that the same verb form /delā/ ‘gave’ behaves differently in (1) and (2) is clearly reflected in their meanings. A similar trend can be found in the use of compound verbs in most Indian languages.

Let us take examples from Odia again to show that traditional speech-to-text transcription can be ambiguous (and \textit{ipso facto} causes text understanding problems if no context is provided).

In sentence (3), /soi/ and /gala/ are two different verbs which form a multi-verb construction\textsuperscript{29}.

(3) rāma dui ghaNTā soi galā
Ram two hour having slept went

*Ram slept for two hours and went.*

In turn, in the following example (4), /soi-galā/ is a predicative two-word verbal unit composed of the verbs $V_1$ and $V_2$ (/so-i/ and /ga-la/ respectively), where $V_1$ is the predicative \textit{polar verb}, and it expresses the meaning ‘to sleep’, whereas the $V_2$ /ga-la/ which is a semantically void \textit{vector verb} plays the role of the support verb for $V_1$, that is, the act of sleeping. Although $V_2$ is semantically empty, it determines the aspect of the whole predicative construction.

(4) rāma dui ghaNTā soi-galā
Ram two hour sleep-went

*Ram slept for two hours.*

We should mention here that the replacement of the vector verb by another one may change the aspect of the whole construction. For example, another $V_2$ vector verb /paDilā/ ‘fell down’ may assign the aspect of suddenness to $V_1$ whereas the $V_2$ vector verb /cālilā/ ‘walked’ may add the aspect of continuation. The examples (5) and (6) are illustrative:
5.4. Outline of the current work on application of lexicon-grammar approach to Indian languages

The current work on the application of the lexicon-grammar approach to Indian Languages is still in its initial phase and focusses on the identification of crucial linguistic phenomena on the syntactic and semantic levels for Hindi, Odia and Bengali. However, already at this stage, we identify several similarities with respect to Slavic languages, in particular Polish.

The following two tasks are our priority now.

1. An exhaustive analysis of the verb-based collocations (conjunct verbs) in the selected Indian languages like Hindi, Odia and Bengali is intended to give a clear description of what kind of nouns and adjectives can be combined with support verbs (vector verbs) to convey the intended meanings. This study will also address the synonymy necessary to correctly extend Indian wordnets to cover collocations.

2. Some compound verbs can take even four or five roots. A detailed analysis of the compound verbs in Hindi, Odia and Bengali will bring out the order in which different vectors and polar verbs are used in such multiword constructions and what grammatical functions the latter perform in various lexical environments.

6. Conclusions

In this paper, we presented some of our achievements in the field of Language Engineering for Polish over more than 30 years (cf. (Vetulani & Vetulani, 2019) and other references below), and we present the recently launched project to apply the methodology presented here to the selection of Less/Medium Resourced Indian languages.

These achievements, in parallel with those of the other Polish academic centres, mean that Polish is no longer considered as belonging to the class of Less-Resourced Languages as it was in the 1990s. Not being exhaustive, this presentation aims at showing the nature of the challenges to be faced when developing sophisticated systems for language processing for Less-Resourced Languages. It is fundamental to possess text (or and speech where necessary) corpora to serve as an empirical basis for the computer modelling of language competence, and to create tools to build and process these corpora (such as lemmatizers, taggers, concordance creators, electronic dictionaries and lexicons, grammars, parsers and text (voice) generators, etc.). The acquisition of such an instrumentarium is a conditio sine qua non for leaving the group of Less-Resourced Languages, and thereby acquiring the necessary potential for creating advanced applications on an industrial scale. To start working to join the still small group of High-Resourced Languages, it is essential to dispose of initial manpower composed of researchers and engineers with a
linguistic or/and IT background. In our case, this pre-condition was satisfied, although initially, only very few were experts in both domains. The need to acquire a solid empirical basis for computer processible resources and related tools for creating non-trivial systems with human language competence was obvious for us from the beginning, but, nevertheless, only work on real-size systems with language competence (such as POLINT-112-SMS) would appear to be a key milestone in the understanding of the scale of the problem. An important side-effect of this project was the creation of a strong multidisciplinary group of well-trained experts in language-related technologies capable of facing operational challenges. These practical works confirmed our understanding of the problems of the scarcity of tools, but first of all permitted us to identify our priorities on the way to enriching our instrumentarium. Among these priorities, there is a further development of the Polish lexicon-grammar integrated with the PolNet lexical basis.

Our developments were presented recently at the large impact conferences as WILDRE 5 workshop (affiliated with LREC 2020) and ICCCI (Viet-Nam, 2020). The intention behind these presentations was to encourage colleagues from India and other countries to take on the exciting challenge to push forwards the development of language resources for the Less- or Medium-Resourced but prominent Indo-Aryan languages.

Annex – Selected Examples (by Panchanan Mohanty)30

A.1 Difference between clause-initial and clause-final complementizer

1.1. Hindi has only the clause-initial complementizer /ki/.
   rawi jântâ hai [ki râm ghar-par hai]
   Ravi knows Comp. Ram home-Loc. be-3Sg.Prs.
   *Ravi knows that Ram is at home.*

1.2. Odia and Bengali have both the clause-initial complementizer /je/ and the clause-final complementizer /boli/ and /bole/ respectively.

1.2.1. Odia

1.2.1.1. Clause-final.
   [râma ghar-e ach-i boli] rabi jâN-e
   Ram home-Loc. be-3Prs. Comp. Rabi know-3Prs.
   *Rabi knows that Ram is at home.*

1.2.1.2. Clause initial.
   rabi jâN-e [je râma ghar-e ach-i]
   Rabi know-3Prs. Comp. Ram home-Loc. be-3Prs.
   *Rabi knows that Ram is at home.*

1.2.2. Bengali

1.2.2.1. Clause final.
   [râm bâRi-te âch-e bole] robi jâN-e
   Ram home-Loc. be-3Prs. Comp Rabi know-3Prs.
   *Rabi knows that Ram is at home.*
1.2.2.2. **Clause initial.**

robi jān-e [je rām bāRi -te âch-e]
Rabi know-3Prs. Comp. Ram home-Loc. be-3Prs.
*Rabi knows that Ram is at home.*

### A.2 Agreement

#### 2.1. Verb and modifier

In Hindi the verb and the modifier agree with the subject in terms of gender and number. However, if the subject is in the ergative case the finite verb agrees with the object.

**Hindi:**

2.1.1.1
rām Ghar jā-e-gā
Ram home go-Fut.3Sg.M.
*Ram will go home.*

2.1.1.2
sītā ghar jā-e-gī
tā home go-Fut.3Sg.F.
*Sita will go home.*

2.1.1.3
rām-ne Cāval khā-yā
Ram-Erg. rice-M. eat-Perf.3Sg..M.Sg.
*Ram will eat rice.*

2.1.1.4
rām-ne rotī khā-ī
Ram-Erg. bread-F. eat-Perf.3Sg.F.Sg.
*Ram ate bread.*

2.1.1.5
sītā-ne cāwal khā-yā
tā-Erg. rice-M eat-Perf.M.3Sg.
*Sita ate rice.*

2.1.1.6
sītā-ne rotī khā-ī
tā-Erg. bread-F. eat-Perf.F.3Sg.
*Sita ate bread.*

#### 2.2. Person and number

Odia has person and number agreement between the animate subject and the finite verb, but not between the inanimate subject and the verb. However Bengali has only person agreement and no number agreement. Odia and Bengali do not have grammatical gender and, therefore, no gender agreement between the subject and the verb. But feminine nouns are derived from most masculine human nouns by adding the feminine
gender suffixes whereas sex-indicating adjectives are used with the non-human nouns to indicate whether they are male or female.

2.2.1. Number agreement with animate subjects:

**Odia**

se ga-lā
she/he/it go-Past.3Sg.
She/he/it went.
se-māne ga-le
they go-Past.3Pl.

They went.

2.2.2. No number agreement with inanimate subjects:

go-Te kār ga-lā
one-Cl. car go-Past.3Sg.

One car went.

dui-Tā kār ga-lā
two-Cl. car go-Past.3Sg.

Two cars went.

A.3. Gender

Non-human animate nouns in Odia and Bengali cannot take a gender marker unlike Hindi; instead a sex-indicating word is used.

**Hindi:**

billā (male cat), billi (female cat)

**Odia:**

aNDirā bilei (male cat), māi bilei (female cat).

**Bengali:**

hulo beRāl (male cat), meni beRāl (female cat)

A.4. Singular and plural markers

Unlike Hindi and Polish, Odia and Bengali nouns do not take the plural markers if they are preceded by a plurality indicating numeral or a quantifier.

4.1. Odia

cari-jaNa pilā ās-i-th-il-e
four-Cl. child come-Perf-Aux-Past-3Pl.

Four children had come.
4.2. Bengali

car-jon chele esh-e-ch-il-o
four-Cl. boy come-Pref.-Aux-Past.
Four boys had come.

A.5. Compound verbs

Hindi, Odia, Bengali use compound verbs frequently and regularly while the compounds of the form verb + verb do not exist in Polish, except when the polar verb is an infinitive. On the other hand, compounds with the verb + verb (Inf.) structure are frequent in Polish.

Hindi:

ritā so-ga-i
Rita sleep-go-Perf-F-3Sg.
Rita slept.

Odia:

ritā soi-ga-l-ā
Rita sleep-go-Past-3Sg.
Rita slept.

Polish:

Ram poszedł spać /subject + vector verb (finite) + polar verb (infinitive)/
Ram go-Past-3Sg (to) sleep Inf /a one-word synonym for the predicative phrase poszedł spać does not exist in Polish/
Ram went to sleep. /as in several Indian languages the vector verb may modify aspect/.

But:

Ram śpi
Ram sleep-Pres-3Sg
Ram sleeps.

Notes

1. See, for example, (Vetulani Z., 1988) reporting our results obtained already in 1984 at GIA, Marseille. Development of Real Size IT Systems.
2. This system is an outcome of the POLEX Polish Lexicon Project (‘POLEX – Polska Leksykalna Baza Danych No KBN8550301007’) realised by Z. Vetulani, B. Walczak, T. Obrębski, G. Vetulani and other team members during 1994–1996 (Vetulani et al., 1998a).
3. The resource is distributed through ELRA. ISLNR: 147-211-031-223-4; ID: ELRA-L0047.
4. http://www.cairn.info/revue-travaux-de-linguistique-2003-1-page-145.htm (consulted on 31 March 2021).
5. GENELEX was continued by LE-PAROLE (1996–1998), LE-SIMPLE (1998–2000) and GRAAL (1992–1996) projects.
6. See, for example, Maurice Gross (Gross, 1975) and Kazimierz Polański (Polański, 1992).
7. CEGLEX Consortium: AMU, Poznań/Poland/, Charles University, Prague/Czech Rep./, GSIERLI, Charenton/France/(coordinator), Lingware, Szeged/Hungary/.
8. From the GRAMLEX project description by Eric Laporte, project coordinator. Quoted by (Vetulani et al., 1998b).
9. GRAMLEX Consortium: AMU, Poznań/Poland/, ASSTRIL, Marne la Vallée/France/ (coordinator), CLR, Salerno/Italy/, Hungarian Academy of Science, Budapest/Hungary/.
10. These tools and applications were: (i) a lemmatizer/tagger (LEXAN) (Vetulani et al., 1997, 1998b), (ii) a generator of inflected forms for simple and compound lexemes (Vetulani et al., 1998b), (iii) a syntactic concordance generator (SCON) (Vetulani et al., 1998b), (iv) a tool for the extraction of compound terms and terminology from texts (EXTRACT) (Vetulani et al., 1998b), (v) an application for the structure analysis of dictionary entries (VERBAN) (Vetulani et al., 1998b), (vi) an application for acquisition of the lexicon from dictionary definitions (NOUNAN) (Vetulani et al., 1998b), and (vii) an application for interactive analysis of dictionary definitions (NOUNDAN) (Vetulani et al., 1998b).
11. Grant of Ministry of Science and Higher Education (MNiSZW) Nr R0002802 (2006–2010).
12. Emergency telephone service maintained by the Polish Police (equivalent to tel. 112).
13. See (Vetulani2014) for the core PolNet bibliography until 2014.
14. At about the same time, Wrocław Technical University started another successful wordnet project plWordnet (pl. Słowosieć) (Piasecki et al., 2009) based on a different methodology.
15. In the Princeton WordNet the basic entities are synsets, that is, classes of synonyms related by semantic relations of which the most important are hyponymy and hyperonymy.
16. See (Vetulani et al., 2009) for PolNet development algorithm.
17. ORBIS, an interface to a database on planets, was implemented in PROLOG by Alain Colmerauer and Richard Kittredge (around 1982) to show the qualities of the declarative programming paradigm. The initially bilingual system was extended with a module for Polish by Zygmunt Vetulani while his research fellowship at GIA (Groupe Intelligence Artificielle, University Aix-Marseille II) in 1984 (Vetulani, 1988).
18. In Vetulani (1994), we described the switch technique which combined with appropriate heuristics permits, on the ground of morphological and the valency information, the reduction of the complexity of parsing sometimes down to linear.
19. This dictionary is described in two monographs. The first one (Vetulani G., 2000) describes the initial phase of work on a dictionary of verb–noun collocations together their usage in sentences as predicates (2862 predicative nouns). This work was done manually. The extension of the resource to 14.600 collocations was described in the second book (2012) reporting further, computer-assisted work. A part of this resource was integrated with PolNet.
20. In (Vetulani et al., 2010) we described a computer-assisted algorithm to extract collocations directly from text corpora. Still, involvement of qualified lexicographers is necessary.
21. The term support verb was first introduced to linguistics by Zellig Harris (Harris, 1964), on the occasion of his research on nominalisation. Maurice Gross (Gross, 1975) used the term verb support (French) in his work on lexicon-grammars, while Ch. Fillmore (Fillmore et al., 2002) preferred to use the word light verb in the project VerbNet. G. Vetulani (Vetulani, 2000) uses the term czasownik podporowy (Polish). The term support verb is close to that of vector verb widely used in Indian linguistics (see Chapter 5 in this paper).
22. Here support verb + predicative noun, but more generally any other predicative word (such as an adjective or adverb) instead of a noun.
23. For example in order to seed-up parsing (Vetulani, Z., 1991).
24. In Polish, we observe the phenomenon of syntactic synonymy (Jędrzejko, 1993) where for some predicative verbs their morpho-syntactic structure varies from the morpho-syntactic structure of their semantic synonyms in the form of verb–noun collocation (e.g. for the direct complement). For consistency with our methodological assumptions, we range these synonymous forms in distinct synsets interconnected by a special semantic similarity relation.
25. From 14.400 in PolNet 2.0 to 17.564 in PolNet 3.0.
26. In Kraków, Poznań, Rzeszów, Śląsk (Silesia), Warszawa, Wrocław, and other places.
27. The Eighth Schedule of the Indian Constitution lists 22 languages. These are: Assamese, Bengali, Bodo, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Meitei (Manipuri), Marathi, Nepali, Odia, Punjabi, Sanskrit, Santhali, Sindhi, Tamil, Telugu,
and Urdu. 12 of these languages are Indo-European (Assamese, Bengali, Gujarati, Hindi, Kashmiri, Maithili, Marathi, Odia, Punjabi, Sanskrit, Sindhi, and Urdu) and are spoken by over 75% of the Indian population.

28. In this paper, we use the term compound verb to refer to a multi-verb collocation. However, in our other papers on lexicon-grammar and verbnets, we usually use this term in a broader sense that covers the totality of collocations playing a predicative function (as e.g. verb–noun or verb–adj collocations, here called conjunct verbs).

29. Such a multi-verb construction is called a conjunctive participle.

30. Terminology in this appendix is compatible with the Indian linguistic tradition.

Acknowledgements

The presented achievements were large due to the financing of works from public funds, for which we thank the grant(s) donors. The authors thank their colleagues who have made a creative contribution to the achievements over the last 30 years and presented here. The authors expressed their gratitude to Dr. Abhijit Debnath for the Bengali data, to Prof. Krzysztof Stroński for consultations, as well as to Dr. Jolanta Bachan and Prof. Belinda Maia for critical comments of the draft version and the final proofreading of the manuscript of this paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Zygmunt Vetulani, studied mathematics (MSc 1973, AMU), romance philology (MA 1989, AMU), foundations of mathematics (PhD 1977, Warsaw University). Alexander von Humboldt Foundation Fellowships: Uni. Bielefeld (1987–1989), and DFKI Saarbrücken (2018). Habilitation in computer linguistics (1990, AMU, Faculty of Neophilology). Full professorship in technical sciences, 2006. Main fields of interests: computational linguistics, artificial intelligence (IT systems with natural language competence), electronic resources for language engineering (electronic dictionaries, grammars and text corpora), Less-Resourced Languages, linguistic algorithms (parsing), formal description of natural languages, applications in literary studies, medicine and public security. Over 120 publications in computer science, linguistics, mathematics and logic. Since 1973 various research positions (AMU and TU Koszalin). Initiator and head of the Department of Computer Linguistics and Artificial Intelligence (1993–2020). Manager or principal investigator in national and international projects. EC expert on Human Language Technologies (1994–2006). Initiator and principal organizer of biannual Language and Technology Conferences (LTC) from 1995 until now.

Grażyna Vetulani, Master studies in Romance Philology, Institute of Romance Studies, Faculty of Modern Languages and Literatures, Adam Mickiewicz University in Poznań (M.A., 1979). PhD in linguistics at the Adam Mickiewicz University (1988). Since 2000 Associate professor at the Adam Mickiewicz University: Head of Department of Romance Linguistics and Comparative Studies (currently). Full Professor (2013). Principal Investigator and Project Manager in national and European research projects. Main research interests: French linguistics, comparative studies (Polish-French), logical and formal aspects of language, computer linguistics (computer-oriented morphology of Polish, predicative nouns in French and Polish), history of French. Author of research papers and books in comparative and French linguistics.

Panchanan Mohanty, has written and edited 30 books in English and Odia, translated 4 books to Odia, and published more than 160 papers in journals of Linguistics and Translation Studies published in India and abroad. He was President of the Linguistic Society of India, Chief Editor of Indian Linguistics, Deputy Editor of International Journal of Dravidian Linguistics, and President of the Dravidian Linguistics Association. He was Principal Investigator of a number of Language Technology projects funded by the University Grants Commission and the Ministry of Electronics and
Information Technology, Government of India. He Chaired the Peer Review and Steering Group of the Ministry of Electronics and Information Technology Project ‘Development of Dependency Tree Banks for Indian Languages’. He has discovered two tribal languages named Walmiki and Malhar and prepared a dictionary of Gorum, an endangered Munda language. Now he is a Professor of English Linguistics at GLA University, Mathura, India.

References
Antoni-Lay, M.-H., Francopoulo, G., & Zaysser, L. (1994). A generic model for reusable lexicons: The genelex project. Literary and linguistic computing, vol. 9, no 1, pp. 47–54. University Press.

Fillmore, C. J., Baker, C. F., & Sato, H. (2002). Seeing arguments through transparent structures. In: Proceedings of Third International Conference on Language Resources and Evaluation, Proceedings, vol. III, Las Palmas, pp. 787–791.

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

Harris, Z. S. (1964). The elementary transformations: Transformations and discourse analysis papers No. 54. University of Pennsylvania.

Jedrzejko, E. (1993). Nominalizacje w systemie i w tekstach współczesnej polszczyzny /in Polish/. Wyd. Uniwersytet Śląski. Katowice.

Jedrzejko, E. (1998). Słownik zwrotów werbo-nominalnych. Zeszyt próbny. /in Polish/. Energeia. Warszawa.

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

Lamiroy, B. (2003). Maurice Gross (1934-2001). In Travaux de linguistique 1/2003 (n°46) (pp. 145–158).

Miller, G. A., Beckwith, R., Fellbaum, C. D., Gross, D., & Miller, K. (1990). Introduction to WordNet: An on-line lexical database*. International Journal of Lexicography, 3(4), 235–244. https://doi.org/10.1093/ijl/3.4.235

Mohanty, P. (1989). Gender-agreement in Oriya. Proceedings of the 15th All India Conference of Linguists. Thiruvananthapuram: International School of Dravidian Linguistics.

Mohanty, P. (2008). Dravidian substratum and Indo-Aryan languages. International Journal of Dravidian Linguistics, XXXVII(1), 1–20.

Mohanty, P. (2010). Wordnets for Indian languages: Some issues. In P. Bhattacharyya, C. Fellbaum, & P. Vossen (Eds.), Principles, construction and application of multilingual wordnets. Proceedings of the 5th Global Wordnet conference (pp. 57–64). Narosa Publishing House.

Mohanty, P. (2011). Oriya: A confluence of aryan, dravidian and munda. In O. N. Koul (Ed.), Indo-Aryan linguistics (pp. 201–212). Central Institute of Indian Languages.

Mohanty, P. (2016). Brutt e mo poshe kuTumba: oDia: bha:sha:baigyanika carca:ra nutana diganta (in odia, linguistic studies on new horizons of Oriya grammar). 2nd ed. Friends’ Publishers.

Mohanty, P., Malik, R. C., Bhol, B., et al. (2017). Issues in the creation of synsets in Odia WordNet. In N. S. Dash (Ed.), The WordNet in Indian languages (pp. 175–200). Springer.

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

Piasecki, M., Szpakowicz, S., & Broda, B. (2009). A wordnet from the ground up. Oficyna Wydawnicza Politechniki Wrocławskiej.

Polański, K. (1992). Słownik syntaktyczno-generatywny czasowników polskich. vol. I–IV, Ossolineum, Wrocław, 1980–1990; vol. V, IJP PAN, Kraków.

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

Szymczak, M. (1995). Słownik języka Polskiego (Polish dictionary). Państwowe Wydawnictwo Naukowe.

Vetulani, G. (2000). Rzeczowniki predykatywne języka polskiego. W kierunku syntaktycznego słownika rzeczowników predykatywnych na tle porównawczym.(in Polish) (Predicate nouns of Polish. Towards a syntactic dictionary of predicate nouns). AMU Press, Poznań.
Vetulani, G. (2012). Kolokacje werbo-nominalne jako samodzielne jednostki języka. Syntaktyczny słownik kolokacji werbo-nominalnych języka polskiego na potrzeby zastosowań informatycznych. Część I (Verb-noun collocations as language units. Syntactic dictionary of Polish verb-noun collocations for NLP applications. Part I) (in Polish). AMU Press, Poznań.

Vetulani, Z. (1988). PROLOG implementation of an access in Polish to a data base. In Studia z automatyki, XI, Państwowe Wydawnictwo Naukowe, Warszawa (pp. 5–23).

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

Vetulani, Z. (1990). Corpus of consultative dialogs. Experimentally collected source data for AI applications. Adam Mickiewicz University Press.

Vetulani, Z. (1991). Lexical preanalysis in a DCG parser of POLISH. In E. Klein (Ed.), Betriebslinguistik und Linguistikbetrieb. Akten des 24 Linguistischen Kolloquiums, Bremen 1989, (Ling. Arbeiten 260/261) (pp. 389–395). Max Niemeyer.

Vetulani, Z. (2004). Komunikacja człowieka z maszyną. Komputerowe modelowanie kompetencji językowej (Man-machine communication. Computer modelling of language competence) (in Polish). Akademicka Oficyna Wydawnicza EXIT.

Vetulani, Z., Martinek, J., Obrębski, T., & Vetulani, G. (1995). The CEGLEX dictionary model for Polish. In: Bazylewicz, R., Kossak, O. (eds.) Proceedings of the 4th and 5th International Conferences UKR SofT (Lviv, 1994, 1995), SP «BaK», Lviv, 1995, pp. 144–150.

Vetulani, Z., & Obrebski, T. (2010). Resources for extending the PolNet – Polish WordNet with a verbal component. In P. Bhattacharyya, C. Fellbaum, & P. Vossen (Eds.), Principles, construction and application of multilingual Wordnets. Proceedings of the 5th Global Wordnet conference (pp. 325–330). Narosa Publishing House.

Vetulani, Z., & Osinski, J. (2017). Intelligent information bypass for more efficient management. Comp. Methods in Science and Technology, 23(2), 105–123.

Vetulani, Z., & Vetulani, G. (2014a). Through wordnet to lexicon grammar. In F. Kakoyianni Doa (Ed.), Penser le lexique grammaire: Perspectives actuelles (pp. 531–543). Editions Honoré Champion.

Vetulani, Z., & Vetulani, G. (2014b). Verb-Noun collocations in PolNet 2.0.. In V. Henrich & E. Hinrichs (Eds.), Proceedings of the Workshop on computational cognitive and linguistic approaches to the analysis of complex words and collocations (CCLCC 2014) (pp. 73–77). ESSLLI.

Vetulani, Z., & Vetulani, G. (2015). Synonymie et granularité dans les bases lexicales du type WordNet. Studia Romanica Posnaniensia, vol. 42/1, WN UAM, Poznań, pp. 113–127.

Vetulani, Z., Vetulani, G., Kochanowski, B. (2016). Recent advances in development of a lexicogrammar of Polish: PolNet 3.0. In N. Calzolari (Ed.), The 10th conference on language resources and evaluation (pp. 2851–2854). ELRA.
Vetulani, Z., & Vetulani, G. (2019). The case of Polish on its way to become a well-resourced language. In A. Gilles (Ed.), *International conference on language technologies for all: enabling linguistic diversity and multilingualism worldwide. Proceedings of LT4All* (pp. 388–392). European Language Resources Association.

Vetulani, Z., & Vetulani, G. (2020). Polish lexicon-grammar development methodology as an example for application to other languages, in: Girish, N. J. Proceedings of the WILDRES – 5th Workshop on Indian Language Data: Resources and Evaluation, pages 51–59.

Vetulani, Z., Walczak, B., Obrebski, T., & Vetulani, G. (1998a). *Unambiguous coding of the inflection of Polish nouns and its application in the electronic dictionaries – format POLEX*. AdamMickiewicz University Press.

Vetulani, Z., Walkowska, J., Obrebski, T., Marciniak, J., Konieczka, P., & Rzepecki, P. (2009). An algorithm for building lexical semantic network and its application to PolNet – PolishWordNet project. In Z. Vetulani & H. Uszkoreit (Eds.), *LTC 2007. LNCS (LNAI)* (Vol. 5603, pp. 369–381). Springer. https://doi.org/10.1007/978-3-642-04235-5_32

Vossen, P., Bloksma, L., Rodriguez, H., Climent, S., Calzolari, N., & Peters, W. (1998). *The euroWordNet base concepts and top ontology, final version*. https://www.researchgate.net/publication/228594694_The_EuroWordNet_Base_Concepts_and_Top_Ontology

Walkowska, J. (2009). Gathering and analysis of a corpus of Polish SMS dialogs. In M. A. Klopotek (Ed.), *Challenging problems of science* (pp. 145–157). Publishing House EXIT.