UDK 025.4:81'373
81'373.46:025.4
Pregledni rad
Prihvaćen za tisak: 20. listopada 2025.
https://doi.org/10.22210/suvlin.2025.100.03

Beata Jarosz–Mazur,
Maria Curie–Skłodowska University, Lublin
beata.jarosz–mazur@mail.umcs.pl

Domain–specific Wordnets: Projects and Achievements Overview

The subject of this article are domain–specific wordnets containing information on specialised vocabulary in specific fields of knowledge. The wordnets are understood, on the one hand, as lexical databases and, on the other, as relational onomasiological dictionaries. The aim is to review what has been achieved in the construction of such lexicons, i.e. to determine which specialised wordnets were planned, which have already been built and whether the results of this work are publicly available. Based on an analysis of English–language publications and on the results of Internet resource searches, it can be said that the achievements in building domain–specific wordnets are not spectacular. About a dozen of them have been developed worldwide so far, but it is difficult to assess the progress of the work, as most of them are not available to users. Several projects have instead been abandoned at an early stage of development. For these reasons, the methodology for constructing domain–specific wordnets is still experimental and constitutes a current research problem.

1. Introduction

Wordnets are huge lexical databases of various languages that have wide applications in computational linguistics, in natural language processing research, and in the development of tools designed to automatically perform tasks, "such as information retrieval, machine translation, question–answer systems, and text summarization" (Smith and Fellbaum 2004: 31). This is because wordnets, "even quite incomplete, [...] are the only machine–tractable lexico–semantic resources for those languages" (Piasecki et al. 2009: 11). From the linguistic point of view, a wordnet is perceived as "an onomasiological dictionary where the main goal is to link words together in semantic fields based on semantic relations" (Fjeld and Nygaard 2009: 13) and as "a formal lexical semantics model that reflects the organization of the mental lexicon" (Barbero and Amaro 2024: 3). It resembles a thesaurus, because "its building block is a synset consisting of all the words that express a

given concept" (Fellbaum 1998: 8) and "all words within a synset mean exactly the same in a certain context" (Poprat et al. 2008: 33).

Over more than three decades, a large number of specialised lexical units from different fields of knowledge have been added to the pioneering Princeton WordNet, whose construction was initiated in the mid–1980s by a team led by George Miller (Miller et al. 1990). In order to distinguish them and highlight their specific character, they are marked with appropriate labels (e.g., MEDICINE {*zymosis*}, SPORTS {*lead*}, ARCHITECTURE {*attic*}, LAW {*trial*}). However, both processes are not carried out systematically and consistently (cf. Bentivogli et al. 2004b). This is because, on the one hand, in generic wordnets information on primarily general language is gathered and, on the other hand, many words belonging to terminological systems and industry nomenclatures are not marked with appropriate identifiers (e.g., in the English Princeton WordNet 3.1. nouns *columnist*, *editorial*, *gatefold*, *reportage*, *scoop*, *teaser* are not tagged with the label MEDIA OR JOURNALISM[1]). Moreover, researchers emphasise that specialised knowledge is presented in generic wordnets in a selective and simplified manner (Fellbaum 1998: 6; Bodenreider and Burgun 2002; cf. Dobrowolska and Szpakowicz 2014). This manifests itself in inadequate (from expert's point of view) simplifications of semantic and hierarchical relationships (cf. Piasecki et al. 2009: 7) and in incomplete definitions, in which usually only information needed by non–experts is included[2] (e.g., Magnini and Speranza 2001: 150; Fellbaum et al. 2006: 323).

For these reasons, the need for the creation of specialised wordnets, in which special lexica from particular fields of knowledge, technology or industry would be collected, was quickly apparent. The ideas of specific projects began to be developed as early as the turn of the 21st century, but despite the passage of time "the creation of a domain–specific wordnet is [still] a more recent phenomenon, of which there are relatively few examples" (Arıcan et al. 2021: 244). The methodology for their creation, drawing to varying degrees on the assumptions and the model of the pioneering Princeton WordNet, is not fully established and rarely becomes the subject of scholarly publications (Jarosz 2024; cf. Di Felippo 2010: 93). A literature review additionally shows that no paper that takes a comprehensive look at this issue has yet been written. Researchers usually discuss the assumptions of particular projects and experiences in their development (e.g., Sagri et al. 2004; Poprat et al. 2008) or address related topics, e.g., techniques for linking domain–specific wordnets with generic wordnets (e.g., Magnini and Speranza 2001; Amaro and Mendes, 2012; Barbero and Amaro 2024), functioning of specialised lexis in generic wordnets and the necessity of extending these lexicons with vocabulary from different

1 http://wordnetweb.princeton.edu/perl/webwn [accessed: 2.11.2024].

2 However, this is not a rule. Researchers also draw attention to the following WordNet's issues: (1) misleading definitions edited by lexicographers who are not specialists in a given field, (2) definitions extracted from specialised thesauri, which may not be understood by lay users (Smith and Fellbaum 2004: 38; Fellbaum et al. 2006: 323).

specific domains (e.g., Burgun and Bodenreider 2001; Buitelaar and Sacaleanu 2002; Zanotti et al. 2012).

In this perspective, it is worthwhile to take a close look at domain–specific wordnets, which are modern thesauri functioning online. This task first requires establishing which specialised wordnets have been developed to date, which ones were intended to be built and to what extent these plans have been realised. Then, on the basis of a comparative analysis of the project ideas and results, it will be possible to characterise the models and solutions implemented and thus determine to what degree the methodology for creating domain–specific wordnets is stabilised and which areas require clarification. Research designed in this way (carried out from a linguist's perspective) will highlight not only the problems and challenges of constructing such thesauri, but also the shortcomings of the theoretical and methodological assumptions. Due to the complexity and breadth of this issue, in this article the results of the first stage of research, aimed at reviewing the achievements in constructing specialised wordnets, will be presented. The outcome of the aforementioned comparative analysis will be published on another occasion.

2. Methodology

The subject of this text is domain–specific wordnets, which are lexical databases and modern thesauri containing information on specialised vocabulary in specific fields of knowledge, technology or industry as used in oral, written and computer–mediated professional communication (cf. Magnini and Speranza 2001; Barbero and Amaro 2024). The aim is to overview the achievements in constructing such resources, i.e. to identify which domain–dependent wordnets have been planned so far, which ones have already been built, and whether the results of this work are publicly available. It should be emphasized that the paper has focused on wordnets containing specialised lexicons in a specific domain, thus omitting multidisciplinary (multiterminological) wordnets[3]. This decision was motivated by a desire to gather information on domain wordnets *sensu stricto*.

The data necessary for this article were obtained primarily from the English–language scientific literature. Firstly, in October 2024, a thematic search of the material available in the ResearchGate database referring to wordnets in a broad sense was carried out. The queries 'specialised wordnet' [p1], 'domain–specific wordnet' [p2], 'terminological wordnet' [p3] were used in the website's search engine. With

3 One would think, for example, of the TermiNet wordnet, which was supposed to contain Brazilian Portuguese terminology from many domains (e.g., politics, economy) and was developed between 2009 and 2011 by the Research Group of Terminology in Federal University of São Carlos in cooperation with the Interinstitutional Center for Computational Linguistics in Brazil (Di Felippo 2010). Another example is the Portuguese LexTec database containing in 2012 over 8 000 lexical units from ten domain–specific wordnets, such as banking, telecommunications, tourism (Amaro and Mendes 2012: 148–149). Unfortunately, the link provided by the authors of the text on this database (Marrafa et al. 2014: 1048) leads nowhere (see: http://www.instituto–camoes.pt/lextec [accessed: 7.4.2025]).

the 'Only full–texts' filter enabled, the following number of records was retrieved: p1 – 18 370, p2 – 10 700, p3 – 12 220. From these results, a set of about 40 publications (mainly written in the last 20 years) that mentioned or described domain–specific wordnets or their drafts was then extracted[4] (most of them can be found in *References* section). In these materials, we primarily searched for information on the lexical resources that were intended to be collected in the wordnet, the degree to which the project work had been completed, and its availability[5]. Additionally, an analogous search was conducted in other databases (ACL Anthology, Academia.edu and Google Scholar) in order to make the collection resource as large as possible. However, it was not possible to find further publications on domain–specific wordnets. It should also be noted that the limitation of the search to full–text versions was due to the desire to find sources with adequate accessibility. The publication title itself would have been insufficient in such designed research.

As the literature analysed did not actually provide website addresses where particular domain–specific wordnets could be found, an additional search of online resources was conducted. Names of the specialised wordnets enumerated in the collected publications (mentioned below, in section 3) were typed into Google as queries and then all the records obtained (both thematic and academic institution websites, researchers' e–profiles, and data deposition platforms) were reviewed to determine whether the thesauri of interest were available online.

Publications on individual wordnets usually contain general information on selected aspects, so in order to clarify relevant details, we attempted to obtain additional feedback from the researchers involved in the development of those thesauri. Unfortunately, most of the emails sent to a dozen people mainly in October 2024 (and repeated a month later) remained unanswered. Three researchers provided explanations and important details that made it possible to improve the characterisations in this paper: Prof. Christiane Fellbaum of Princeton University in USA, engaged in the Medical Wordnet project, Prof. Udo Hahn (emeritus) from Friedrich–Schiller–Universität Jena in Germany, involved in the work on BioWordNet, and Prof. Olcay Taner Yıldız of Özyeğin University in Istanbul (Turkey), contributing to Tourism WordNet and Estate WordNet.

4 English–language texts predominate among these publications, but thanks to the functionality of searching English abstracts of non–English publications, two Italian texts (Bodrato 2006; Bocco et al. 2008) and two Turkish texts (Parlar et al. 2019; Arıcan et al. 2021) were also found, all of which also concerned domain–specific wordnets.

5 It should be emphasized that the research is conducted from linguist's perspective. Such a caveat is necessary because the creation of any e–lexicon requires the collaboration between linguists and computer scientists (e.g., Bentivogli et al. 2004a: 40) and therefore the characterisation of such creations may consider different aspects and viewpoints – linguistic or technical.

3. Results – projects and development degree

On the basis of our research out, it can be concluded that over a period of more than 25 years, a dozen domain–specific wordnets containing professional lexis in the different fields of knowledge and industry have been developed in various (mainly Italian) science centres. The projects, which have been characterised quite thoroughly in the analysed literature, are listed in Table 1 and described below according to specific criteria. A few other specialised wordnets were mentioned in English–language publications, but as they were not discussed in greater details, they are only briefly mentioned in the following overview.

Table 1. Domain–specific wordnets described in the analysed English–language literature.

Name of specialised wordnet	**Field of knowledge**	**Lan–guage(s)**	**Entities involved♦**	**Project duration♦♦**	**Project status**	**Development degree**	**Accessibility♦♦♦**
ArchiWordNet [ArchiWN]	Architecture and construction	Italian and English	Istituto per la Ricerca Scientifica e Tecnologica in Trento restructured into Bruno Kessler Foundation, Italy	1999–[2004]–[2012]	Unknown [in 2012, declaration of further development]	Partly completed	Unavailable
BioWordNet [BioWN]	Biomedicine	English	* Jena University Language & Information Engineering Lab (JULIE Lab) at Friedrich–Schiller–Universität Jena, Germany	*?–[2008]*	Abandoned at the beginning	Unrealised	–
Economic–WordNet [EcoWN]	Economics and finance	Italian	* Institute of Computational Linguistics in Pisa – National Research Council of Italy; Istituto per la Ricerca Scientifica e Tecnologica in Trento; Consorzio Pisa Ricerche, Italy	The end of the 20th century–[2000]–[2001]	Unknown	Partly completed	Unavailable
Estate WordNet [EstateWN]	Real estate	Turkish	* Starlang Yazılım Danışmanlık in Istanbul; Departments of Computer Engineering at Istanbul University and Boğaziçi University; Zingat Real Estate Information Systems in Istanbul; Turkey	?–[2019]	Suspended/discontinued	Partly completed	Available upon request

JurWordNet [JurWN]	Law	Italian	Institute of Legal Information Theory and Techniques in Florence; Laboratory for Applied Ontology at the Institute of Cognitive Sciences and Technologies – National Research Council of Italy	1999–[2003]–[2004]	Unknown [in 2004, still under development]	Partly completed	Available
Maritime WordNet [MarWN]	Navigation and maritime transport	Italian	Institute of Computational Linguistics in Pisa – National Research Council of Italy	2003–[2004]–[2010]	Unknown [in 2010, still under development]	Partly completed	Unavailable
Medical WordNet [MedWN]	Medicine	English	* Princeton University, State University of New York at Buffalo, USA; Institute for Formal Ontology and Medical Information Science at Saarland University, Germany; Berlin–Brandenburg Academy of Sciences Berlin, Jena University Language & Information Engineering Lab (JULIE Lab) at Friedrich–Schiller–Universität Jena, Germany	*[2004]–[2006]*	Abandoned at the beginning	Unrealised	–
Tourism WordNet [TourWN]	Tourism	Turkish	Starlang Yazılım Danışmanlık, Istanbul, Turkey	?–[2021]	Suspended/ discontinued	Partly completed	Available upon request

♦ In the case of BioWordNet, EcoWordNet, Estate Wordnet and Medical WordNet, it was not possible to find precise information on the entities involved in the studies, therefore the centers represented by the researchers participating in the development of these thesauri and describing the process in scientific articles are included in the table.

♦♦ Dates without parentheses are indicated in publications as the starting point of the project. Dates in parentheses refer to the year of the publication describing a given domain–specific wordnet or the principles of its construction. Dates written in italics refer to wordnets that were planned but ultimately abandoned.

♦♦♦ Accessibility was verified three times – in October and November 2024, and in February 2025.

Source: own study based on information in: [ArchiWN] Bentivogli et al. 2004a; Bertorello 2012; [BioWN] Poprat et al. 2008; [EcoWN] Roventini et al. 2000; Magnini and Speranza 2001; [EstateWN] Parlar et al. 2019 (and data from Prof. O.T. Yıldız); [JurWN] Gangemi 2003; Sagri et al. 2004; [MarWN] Marinelli et al. 2004; 2010; Roventini and Marinelli 2004; [MedWN] Smith and Fellbaum 2004; Fellbaum et al. 2006; [TourWN] Arıcan et al. 2021 (and data from Prof. O.T. Yıldız).

From the data provided in Table 1, it is clear that the majority of domain–specific wordnets contained or were intended to contain specialised lexis from a specific language: Italian [EcoWN, JurWN, MarWN], English [BioWN, MedWN] or Turkish [EstateWN, TourWN]. In contrast, only one bilingual English–Italian thesaurus [ArchWN] was developed. Interestingly, only two projects dealt with medical terminology [BioWN, MedWN] and in each of the remaining wordnets researchers focused on nomenclature used in a different field [ArchWN, EcoWN, EstateWN, JurWN, MarWN, TourWN].

With regard to all of the specialised wordnets listed, it is difficult to determine precisely the status of the realisation of the original design, but based on the information contained in the publications, it is clear that even in the case of projects that were abandoned at an early stage [BioWN, MedWN] some of the planned work was carried out. The creators of Medical WordNet considered their project as "to some degree a visionary enterprise" (Fellbaum et al. 2006: 331), but they performed the initial activities. From the information they provided, we know that in order to make an initial estimate of the volume of the wordnet under development, they extracted "a test lexicon of 2 838 single–word medical terms" from various resources (Smith and Fellbaum 2004: 38) and discovered that only 11 terms from this collection (mainly compounds, such as *breastfed*, *coldsore*) were absent from the then English WordNet 2.0. They also verified the adequacy of the definitions contained in this general wordnet, prepared a test corpus of 1 644 sentences and attempted to do preliminary work on it. However, lack of funding derailed the progress of the project (Fellbaum et al. 2006: 326, 331).

On the other hand, German researchers intending to build BioWordNet, containing English biomedical terminology, encountered some serious problems at the very beginning of their work. This is because their aim was to extract the necessary data from 60 publicly accessible biomedical ontologies collected in the Open Biomedical Ontologies database.[6] Ontologies are databases that organize not words, but concepts labeled with appropriate words and linked by ontological relations (e.g., Piasecki et al. 2009) and are therefore relevant sources of data needed to build a wordnet. Unfortunately, the experimental procedure of converting the ontologies to a wordnet that was developed by the BioWordNet's creators appeared to be imperfect and generated various errors (for a detailed description see Poprat et al. 2008). An additional complication was the technical limitations of the

6 There are currently 1506 ontologies on this platform: https://www.bioontology.org [accessed: 2.11.2024].

software used to construct wordnets in the first decade of the 21st century. Among other things, it was impossible to assign more than 15 meanings to a single word and to introduce new types of semantic relations between synsets specific to biomedicine. Solving these problems required a major modification of the established methodology and a huge amount of time and effort, so the BioWordNet project was abandoned.

In contrast, significant work has been completed on the bilingual ArchiWordNet containing the Italian and the English terminology in the fields of architecture and construction. The history of this specialised wordnet dates back to the end of the 20th century (cf. Bodrato 2006: 199) and it is known that, in the first phase of the project 3 800 synsets for several domains were created: <BUILDINGS AND BUILDINGS COMPLEXES>, <BUILDING ELEMENT, CONSTRUCTION ENTITY PART>, <MATERIAL> (Bertorello 2012: 40). This work presumably continued, as in the second decade of the 21st century it was reported that the original concept had to be revised, but since there have been no new publications for more than 10 years, it is impossible to estimate the growth of the database and to assess the progress of the project. In addition, as one can read in the publications, this domain–dependent wordnet was initially available only on the intranet at the Italian Politecnico di Torino[7] (Bentivogli et al. 2004a: 39; cf. Bocco et al. 2008) and later the collected data were integrated into MultiWordNet. This is also evidenced by an illustration in one of the articles showing an example of an ArchiWordNet entry displayed from within MultiWordNet (Bertorello 2012: 40). Unfortunately, this resource could not be accessed as of October and November 2024, when the research was conducted. An attempt at accessing the website where MultiWordNet should be located[8] and which was listed in the Global WordNet Association's online catalogue,[9] generated a "403: Forbidden" error. Another attempt was made in February 2025, but again no publicly available dataset of ArchiWordNet could be found.

The work on JurWordNet built as "an extension for legal domain of the Italian ItalWordNet (IWN)" was also quite advanced (Sagri et al. 2004: 305). The collected data was planned to be included in the aforementioned Italian general wordnet, but – as with MultiWordNet – despite several attempts, it was not possible to access it in either October–November 2024 or February 2025[10]. However, as a result of a painstaking online search, a dataset entitled "JurWordNet" was found on the Lynx project platform (Legal Knowledge Graph for Multilingual Compliance Services) that has become available – according to the metadata – under a CC–BY–4.0 licence

7 Efforts to contact the creators of ArchiWN and JurWN were unsuccessful.

8 See: https://multiwordnet.fbk.eu/english/home.php [accessed: 2.11.2024].

9 See: http://globalwordnet.org/resources/wordnets–in–the–world/ [accessed: 2.11.2024].

10 It should be noted that the list of wordnets provided by the Global WordNet Association states that the access to ItalWordNet is "restricted". However, on the website of the Institute of Computational Linguistics in Pisa, where this wordnet was constructed, there is no information on the exact place where this dataset was deposited. See: http://globalwordnet.org/resources/wordnets–in–the–world/ [accessed: 2.11.2024]; https://www.ilc.cnr.it/en/progetti/italwordnet/ [accessed: 27.2.2025].

in March 2019. This resource can also be accessed on the EuroTermBank platform co–funded by the European Union, containing 16 million terminology units from 45 languages[11]. The entries belonging to JurWordNet are highlighted in this database with an appropriate label and hence it is evident that this domain–dependent wordnet has certainly been developed to a certain extent. The self–calculation has revealed that the resource provided on both platforms counts 4901 synsets and 5169 lemmas. It is not known, however, whether this is a complete dataset[12] and whether the researchers have fulfilled their initial intentions.

To unfortunately unknown extent, a lexical dataset referred to in sources as EcoWordNet and described as a specialised wordnet containing Italian vocabulary in economics and finance was also constructed[13] (e.g., Gangemi et al. 2003: 745; Roventini and Marinelli 2004: 194). The thesaurus started to be developed at the end of the 20th century (Roventini et al. 2000) and, as the publications state, counted at the beginning of the 21st century "about 5000 lemmas distributed in about 4700 synsets" (Magnini and Speranza 2001: 151). A few years later, it was reported that this terminological dataset was publicly available (Sagri et al. 2004: 306), since it was – like JurWordNet – incorporated into ItalWordNet. Regrettably, as already mentioned, the Italian general wordnet is currently inaccessible and it is impossible to verify information about EcoWordNet.

The lack of access to ItalWordNet makes it also impossible to assess the progress of work on Maritime WordNet (also called "Maritime Domain Lexicon" and "MariTerm"), which is a terminological lexicon containing Italian specialised vocabulary in the field of navigation and maritime transport[14] and which "has been structured according to the design principles of the generic wordnet" (Marinelli et al. 2004: 465). A part of the project must have been completed, however, since the publications provide accurate information on the volume of data collected and declare the work is in progress. The advancement of Maritime WordNet, on the other hand, is evidenced by comparing the quantitative indicators of the set of lemmas included in this lexicon: in 2004 there were 2256 lemmas (grouped in 1736 synsets; Marinelli et al. 2004: 466), and in 2010 the number rose to 4000 (Marinelli et al. 2010: 2288).

One article also mentions two other projects of Italian thesauri containing special vocabulary on: (1) taxation law as well as (2) labour law and union labour rules

11 https://eurotermbank.com/collections/664 [accessed: 15.10.2024].

12 There are indications suggesting that this may be a draft version. One piece of evidence is the inconsistent spelling of lemmas, i.e. unjustified capital letters in expressions such as *Data documento* ['document date'] and *Lavoro economia* ['labour economics'], as well as lower–case proper names, e.g., **banca centrale europea* ['European Central Bank']. It is possible that these terms were present in such a form in the automatically processed source materials, extracted with erroneous spelling, and uncorrected later on.

13 In one publication it was stated, apparently mistakenly, that the wordnet contained ecology–related lexis (Barbero and Amaro 2024: 14). This information is not confirmed by other sources.

14 The authors emphasised that this wordnet "involves many other fields of knowledge ranging from geography and meteorology to cartography, from astronomy and law to maritime contracts and transport technology" (Marinelli et al. 2004: 466).

(Marinelli et al. 2010: 2288). From the descriptions it appears that they were built analogously to Maritime WordNet, text corpora were created for them, and they collected 1600 and 1500 lemmas respectively. However, these thesauri have not been characterised in detail, and it has not been reported whether they have been made available. Unfortunately, there was no success in finding them, so it is not possible to determine whether they are indeed structured as wordnets (and therefore they have not been included in Table 1).

As far as the Turkish Tourism WordNet described a few years ago is concerned, it is evident that it was built to some extent, according to an article by a group of researchers (Arıcan et al. 2021) for the purpose of conducting precise domain–specific sentiment analysis and semantic annotation. From the information provided by Professor Olcay Taner Yıldız of Özyeğin University in Istanbul (Turkey), this lexicon contains 14 819 lemmas from the tourism domain (both commonly used and strictly domain–specific). They were organised within 13 355 synsets, which were then – as in any wordnet – linked by means of a network of semantic relations (cf. Arıcan et al. 2021: 244). In a similar way, the Estate WordNet (Bakay et al. 2021) characterised in a text written in Turkish (Parlar et al. 2019) was constructed. One can learn from the abstract, nevertheless, that this thesaurus counted 7000 words organised within 11 000 synsets. Professor Yıldız, who was also involved in this project, clarified that after the article was published, the built database was cleaned up and counts 6424 lemmas and 6298 synsets. He added, however, that due to lack of funding, both projects are not currently ongoing, but it is possible to access both resources upon request. It should be mentioned that publications mention that Turkish researchers have built several other domain–specific wordnets (Sanıyar et al. 2023: 87). Unfortunately, despite several attempts, no additional information could be found on this subject, and it is not even known which lexicons from which domains are (or were intended to be) collected in them.

As a final remark, domain–dependent wordnets are sometimes described as the abovementioned ontologies; strictly speaking, the latter organise domain–specific concepts rather than domain–specific lexical units (e.g., urbanism; see Lacasta et al. 2008). In recent articles, GeoWordNet (e.g., Tessarollo and Rademaker 2020) is sometimes mentioned among specialised wordnets and characterised as "a semantic and linguistic resource obtained from the integration of GeoNames with WordNet plus the Italian section of MultiWordNet" (Giunchiglia et al. 2010: 122). This dataset, however, is not a wordnet, but a multilingual descriptive ontology, which in 2011 contained "110 459 classes, 6 927 078 instances, 6 927 078 'instance of', 89 266 'is a' and 5325 transitive 'part of' relations, […] 98 907 associative relations" (Maltese and Farazi 2011: 10). Moreover, GeoWordNet includes proper names organised within a conceptual framework with additional geographic locative data, rather than specialised vocabulary.

4. Conclusion and discussion

With regard to the multiplicity of fields within which specialised lexicons have been developed over the centuries, it becomes obvious that the achievements to date in building domain–specific wordnets are not spectacular. It is, of course, difficult to estimate what percentage of all domain–dependent wordnets is accounted for by the aforementioned projects, which (according to bibliographic data, among other things) were largely initiated and implemented at the turn of the 21st century.[15] However, based on an analysis of the English–language literature and information available on the Internet, it can be assumed that this is a representative group. In publications addressing the broad topic of wordnets mainly the abovementioned thesauri are described (Tessarollo and Rademaker 2020; Arıcan et al. 2021; Barbero and Amaro 2024) and as far as newer projects are concerned, only two Turkish wordnets [EstateWN, TourWN] are detailed[16]. Surprisingly, however, among the specialised wordnets mentioned in the literature, only one is available online without any restrictions, albeit it is not known whether it is a complete dataset [JurWN]. Two other thesauri can be accessed on request [EstateWN; TourWN], another two cannot be reached at all [ArchiWN; MarWN], and the remaining projects listed have been abandoned at an early stage of development [BioWN; MedWN].

Researchers point out that there is a high demand for such lexical databases containing information on specialised vocabulary in various fields. This is due to, inter alia, the specificity of tools for automatic processing of domain–specific texts (cf. Smith and Fellbaum 2004: 31; Poprat et al. 2008: 31). Indeed, the effectiveness of such software depends on access to data on both non–expert and expert vocabulary (e.g., Buitelaar and Sacaleanu 2002; Di Felippo 2010; Zanotti et al. 2012), because in professional texts "the specialised lexicon, i.e. words that denote more specific concepts and knowledge, emerging from specific domains […], co–exist and co–occur with the common lexicon, i.e. the set of words that denote concepts and knowledge shared by average speakers" (Barbero and Amaro 2024: 1). The usefulness of domain–dependent wordnets is further recognised in a variety of professional spheres, in teaching activities "and in general whenever a reference to terms of this specific domain is needed" (Marinelli et al. 2004: 465).

15 The analyses presented in this text are based on a search of English–language publications available online. It is therefore possible that (1) other domain–specific wordnets are described in printed works that are not indexed on the internet, are written in other languages, and do not have abstracts in English; (2) other projects of this type have been completed or initiated, but for various reasons have not been discussed in research papers; (3) specialised wordnets are currently being constructed that have not yet been discussed or presented (see the next footnote).

16 However, Turkish scientists are working on further resources of this type, as are Polish researchers, who in 2025 began building a subwordnet containing professional music vocabulary, which will be an extension of the general wordnet PlWordNet (Alberski et al. [in print]).

Acknowledgments

The author would like to thank Professor Christiane Fellbaum from Princeton University (USA) for clarifying some issues regarding the Medical WordNet concept, Professor of Computational Linguistics Udo Hahn (emeritus) from Friedrich–Schiller–Universität Jena (Germany) for providing information on the BioWordNet project and Professor Olcay Taner Yıldız of Özyeğin University in Istanbul (Turkey) for sharing the information on Tourism WordNet and Estate WordNet. The author is also grateful to Bartłomiej Alberski, PhD, from the CLARIN–PL consortium established at the Wrocław University of Science and Technology (Poland) as part of the European research network CLARIN ERIH for his support in verifying the accessibility of wordnets.

References

Alberski, Bartłomiej, Alicja Helena Derych, Hubert Jankowski, Beata Jarosz–Mazur, Maciej Piasecki, and Pawel Dembowski (in print). Overview of Specialized Wordnets in the Context of Building a Subwordnet Describing Music Professiolect.

Amaro, Raquel, and Sara Mendes (2012). Towards merging common and technical lexicon wordnets. In: Michael Zock and Reinhard Rapp, eds. *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*. Mumbai: The COLING 2012 Organizing Committee, 147–160.

Arıcan, Bilge Nas, Özçelik Merve, Aslan Deniz Baran, Sarmış Elif, Parlar Selen, and Olcay Taner Yıldız (2021). Creating domain dependent Turkish WordNet and SentiNet. In: Sonja Bosch, Christiane Fellbaum, Marissa Griesel, Alexandre Rademaker and Piek Vossen, eds. *Proceedings of the 11th Global Wordnet Conference*. Potchefstroom: Global Wordnet Association, 243–251, https://doi.org/10.18653/v1/2021.gwc–1.28.

Bakay, Özge, Ergelen Özlem, Sarmış Elif, Yıldırım Selin, Arıcan Bilge Nas, Kocabalcıoğlu Atilla, Özçelik Merve, Sanıyar Ezgi, Kuyrukçu Oğuzhan, Avar Begüm, and Olcay Taner Yıldız (2021). Turkish WordNet KeNet. In: Sonja Bosch, Christiane Fellbaum, Marissa Griesel, Alexandre Rademaker and Piek Vossen, eds. *Proceedings of the 11th Global Wordnet Conference*. Potchefstroom: Global Wordnet Association, 166–174, https://doi.org/10.18653/v1/2021.gwc–1.19.

Barbero, Chiara, and Raquel Amaro (2024): Are We Talking about the Same Thing? Modeling Semantic Similarity between Common and Specialized Lexica in WordNet. *Languages* 9: 1–19, https://doi.org/10.3390/languages9030089.

Bentivogli, Luisa, Bocco Andrea, and Emanuele Pianta (2004a). ArchiWordNet: Integrating WordNet with Domain–Specific Knowledge. In: Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum and Piek Vossen, eds. *Proceedings of the 2nd International Global Wordnet Conference*. Brno: Masaryk University, 39–46.

Bentivogli, Luisa, Forner Pamela, Magnini Bernardo, and Emanuele Pianta (2004b). Revising the Wordnet Domains Hierarchy: semantics, coverage and balancing. In: Gilles Sérasset, Susan Armstrong, Christian Boitet, Andrei Popescu–Belis and Dan Tufis,

eds. *Proceedings of the Workshop on Multilingual Linguistic Resources MLR2004*. Geneva: COLING, 94–101.

Bertorello, Anna Rita (2012). A new hierarchy of ArchiWordNet (AWN): building parts implementation with image. In: Christiane Fellbaum and Piek Vossen, eds. *GWC 2012. 6th International Global Wordnet Conference. Proceedings*. Brno: Tribun EU, 40–44.

Bocco, Andrea, Bodrato Enrica, and Antonella Perin (2008). Archiwordnet, un thesaurus di settore integrato nel wordnet della lingua generica: compilazione e applicazioni. *AIDA informazioni* 1/2: 77–87.

Bodenreider, Olivier, and Anita Burgun (2002). Characterizing the definitions of anatomical concepts in WordNet and specialized sources. In: *Proceedings of the 1st Global WordNet Conference*. Mysore: Central Institute of Indian Languages, 223–230.

Bodrato, Enrica (2006). Il fondo fotografico Paolo Verzone: restauro e catalogazione. *Archivi* 1: 195–199.

Buitelaar, Paul, and Bogdan Sacaleanu (2002). Extending Synsets with Medical Terms. In: *Proceedings of the 1st Global WordNet Conference*. Mysore: Central Institute of Indian Languages, 216–222.

Burgun, Anita, and Olivier Bodenreider (2001). Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System. In: Dan Mlodovan, ed. *Proceedings of NAACL'2001 Workshop. WordNet and Other Lexical Resources: Applications, Extensions and Customizations*. Pittsburgh: Association for Computational Linguistics, 77–82.

Di Felippo, Ariani (2010). The TermiNet Project: an Overview. In: Thamar Solorio and Ted Perdersen, eds. *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*. Los Angeles: Association for Computational Linguistics, 92–99.

Dobrowolska, Marta, and Stan Szpakowicz (2014). Terminology in WordNet and in plWordNet. In Heili Orav, Christiane Fellbaum and Piek Vossen, eds. *Proceedings of the 7th Global Wordnet Conference, GWC 2014*. Tartu: University of Tartu, 299–303.

Fellbaum, Christiane (1998). Introduction. In: Christiane Fellbaum, ed. *WordNet – an electronic lexical database*. Massachusetts: The MIT Press, 1–20.

Fellbaum, Christiane, Hahn Udo and Barry Smith (2006). Towards new information resources for public health. From WordNet to Medical WordNet. *Journal of Biomedical Informatics* 39(3): 321–332, https://doi.org/10.1016/j.jbi.2005.09.004.

Fjeld, Ruth Vatvedt and Lars Nygaard (2009). NorNet – a monolingual wordnet of modern Norwegian. In Bolette Sandford Pedersen, Anna Braasch, Sanni Nimb and Ruth Vatvedt Fjeld, eds. *Proceedings of the NODALIDA 2009 workshop WordNets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*. Northern European Association for Language Technology, 13–16.

Gangemi, Aldo, Sagri Maria–Teresa, and Daniela Tiscornia (2003). Metadata for Content Description in Legal Information. In: Vladimir Marik, Werner Retschitzegger and Olga Stepankova, eds. *14th International Workshop on Database and Expert Systems Applications, 2003. Proceedings*. Berlin: Springer, 745–749, https://doi.org/10.1109/DEXA.2003.1232110.

Giunchiglia, Fausto, Maltese Vincenzo, Farazi Feroz and Biswanath Dutta (2010). GeoWordNet: A Resource for Geo–spatial Applications. In: Lora Aroyo, Grigoris Antoniou, Eero Hyvönen, Annette Teije, Heiner Stuckenschmidt, Liliana Cabral and Tania Tudorache, eds. *The Semantic Web: Research and Applications. 7th Extended Semantic Web Conference, ESWC 2010, Heraklion, Crete, Greece, May 30 June 2, 2010. Proceedings. Part I*. Berlin: Springer, 121–136.

Jarosz, Beata (2024). Internetowy słownik języka zawodowego polskich dziennikarzy prasowych. Koncepcja tezaurusa dziedzinowego typu wordnet – preliminaria [An online dictionary of the professional language of Polish press journalists. The concept of domain–specific wordnet–like thesaurus – preliminary comments]. *Prace Językoznawcze* 2: 203–219, https://doi.org/10.31648/pj.10146.

Lacasta, Javier, Nogueras–Isso Javier, Zarazaga–Soria Francisco Javier, and Pedro R. Muro–Medrano (2008). Generating an urban domain ontology through the merging of cross–domain lexical ontologies. In: Jacques Teller, Chris Tweed and Giovanni Rabino, eds. *Conceptual Models for Urban Practitioners*. Bologna: Società Editrice Esculapio, 69–84.

Magnini, Bernardo, and Manuela Speranza (2001). Integrating Generic and Specialized Wordnets. In *Proceedings of the 2nd Conference on Recent Advances in Natural Language Processing*. Tzigov Chark: RANLP–2001, 149–153.

Maltese, Vincenzo and Feroz Farazi (2011). Towards the Integration of Knowledge Organization Systems with the Linked Data Cloud, https://core.ac.uk/download/pdf/150083512.pdf [accessed: 20. 10. 2024].

Marinelli, Rita, Roventini Adriana, and Alessandro Enea (2004). Building a Maritime Domain Lexicon: a Few Considerations on the Database Structure and the Semantic Coding. In: Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa and Raquel Silva, eds. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon: European Language Resources Association (ELRA), 465–468.

Marinelli, Rita, Roventini Adriana, Spadoni Giovanni, and Sebastiana Cucurullo (2010). Lexical Semantic Resources in a Terminological Network. In: Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, eds. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10*). Valletta: European Language Resources Association (ELRA), 2288–2291.

Marrafa, Palmira, Amaro Raquel, and Sara Mendes (2014). LexTec – a rich language resource for technical domains in Portuguese. In: Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis, eds. *Proceedings of the 9th International Conference on Language Resources and Evaluation – LREC 2014*. Reykjavik: European Language Resources Association, 1044–1050.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller (1990), Introduction to WordNet: An on–line lexical database. *International Journal of Lexicography* 3: 235–244.

Parlar, Selen, Arıcan Bilge Nas, Erkek Mehmet, Çayırlı Kamil, and Taner Olcay Yıldız (2019). Emlak Alanına Özgü Kelime Ağı [Domain Dependent Wordnet for Real Estate]. In: *Proceedings of the 27th Signal Processing and Communication Applications Conference (SIU 2019)*. Sivas: IEEE, 404–406.

Piasecki, Maciej, Stanisław Szpakowicz, and Broda Bartosz (2009). *A Wordnet from the Ground Up*. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej.

Poprat, Michael, Elena Beisswanger, and Udo Hahn (2008). Building a BioWordNet by Using WordNet's Data Formats and WordNet's Software Infrastructure – A Failure Story. In: K. Bretonnel Cohen and Bob Carpenter, eds. *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*. Columbus: Association for Computational Linguistics, 31–39.

Roventini, Adriana, Antonietta Alonge, Nicoletta Calzolari, Bernardo Magnini, and Francesca Bertagna (2000). ItalWordNet: a Large Semantic Database for Italian. In: Maria Gavrilidou, George Carayannis, Stella Markantonatou, Stelios Piperidis and Gregory Stainhauer, eds. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. Athens: European Language Resources Association (ELRA),
http://www.lrec–conf.org/proceedings/lrec2000/pdf/129.pdf [accessed: 20.10.2024].

Roventini, Adriana, and Rita Marinelli (2004). Extending the Italian WordNet with the Specialized Language of the Maritime Domain. In: Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, and Piek Vossen, eds. *Proceedings of the 2nd International Global Wordnet Conference*. Brno: The Global Wordnet Association, 193–198.

Sagri, Maria–Teresa, Daniela Tiscornia, and Francesca Bertagna (2004). Jur–WordNet. In: Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, and Piek Vossen, eds. *Proceedings of the 2nd International Global Wordnet Conference*. Brno: The Global Wordnet Association, 305–310.

Sanıyar, Ezgi, Oguzhan Kuyrukçu, and Olcay Taner Yıldız (2023). StarNet: A WordNet Editor Interface. In German Rigau, Francis Bond, and Alexandre Rademaker, eds. *Proceedings of the 12th Global Wordnet Conference*. San Sebastián: Global Wordnet Association, 84–90, https://doi.org/10.18653/v1/2023.gwc–1.10.

Smith, Barry, and Christiane Fellbaum (2004). Medical WordNet: A New Methodology for the Construction and Validation of Information Resources for Consumer Health. In: E. Yuste, S. J. Jekat, A. K. Pantli, G. Massey, eds. *Proceedings of the 20th International Conference on Computational Linguistics, Geneva, 23–27 August 2004*. Geneva: Association for Computational Linguistics, 31–38.

Tessarollo, Alexandre, and Alexandre Rademaker (2020). Inclusion of Lithological terms (rocks and minerals) in The Open Wordnet for English. In: Thierry Declerck, Itziar Gonzalez–Dios, and German Rigau, eds. *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*. Marseille: The European Language Resources Association (ELRA), 33–38.

Zanotti, Cristian, Jorge Vivaldi, and Marcé Lorente (2012). Upgrading WordNet: a Terminological Point of View. In: Christiane Fellbaum, and Piek Vossen, eds. *GWC 2012. 6th International Global Wordnet Conference*. Brno: Tribun EU, 390–399.

Specijalizirani wordneti: pregled projekata i postignuća

Predmet su ovoga rada specijalizirani *wordneti*, računalne leksičke baze koje sadrže podatke o specijaliziranom vokabularu u pojedinim stručnim područjima. *Wordneti* se s jedne strane shvaćaju kao leksičke baze podataka, a s druge strane kao relacijski onomaziološki rječnici. Cilj je rada dati pregled postignuća u izgradnji takvih *wordneta*, tj. utvrditi koji su specijalizirani *wordneti* planirani, koji su već izgrađeni i jesu li rezultati tog rada javno dostupni. Na temelju analize publikacija na engleskom jeziku i rezultata pretraživanja internetskih izvora može se reći da postignuća u izgradnji *wordneta* nisu osobita. Do sada ih je diljem svijeta razvijeno desetak, ali teško je procijeniti napredak rada na njima jer većina nije javno dostupna korisnicima. Štoviše, nekoliko je projekata napušteno u ranoj fazi razvoja. Iz tih razloga metodologija za izgradnju *wordneta* još je uvijek eksperimentalna i predstavlja trenutni istraživački problem.

Ključne riječi: wordnets, semantički odnosi, specijalizirani tezaurusi
Key words: Wordnets, semantic relations, specialised thesaurus