Terminology in WordNet and in plWordNet

Marta Dobrowolska
Institute of Informatics
Wrocław University of Technology
Wrocław, Poland
martadobr@gmail.com

Stan Szpakowicz
Institute of Computer Science
Polish Academy of Sciences
Warsaw, Poland
&
School of Electrical Engineering
and Computer Science
University of Ottawa
Ottawa, Ontario, Canada
szpak@eecs.uottawa.ca

Abstract

We examine the strategies of organizing terminological information in WordNet, and describe an analogous strategy of adding terminological senses of lexical units to plWordNet, a large Polish wordnet. Wordnet builders must cope with differences in lexical and terminological definitions of a term, and with the boundaries between terminological and lexical information. A somewhat adjusted strategy is required for Polish, though both WordNet and plWordNet rely mainly on semantic relations in organizing the terminological and general-language units. The proposed guidelines for plWordNet, built on several distinct combinations of denotation and connotation, have a solid theoretical underpinning but will require a large-scale verification of their effectiveness in practice.

1 Introduction

The study of lexicography invokes three types of definition: lexicographic, encyclopedic and terminological.

> The object of a lexicographic definition is [...] the verbal representation, the word itself; the object of the terminological definition is the concept, the abstract representation of the entity existing in the real world (Hudon, 1998, pp. 80-81).

The third definition type describes the real-world object itself, recalling everything that is known about it (Hudon, 1998, p. 81). The three types broadly correspond to linguistic dictionaries, encyclopaedias and specialised dictionaries respectively. It would be very unlikely, however, to find a purely linguistic dictionary with entries devoid of encyclopaedic or terminological elements. Section 2 describes how different kinds of information can be included in a dictionary entry, how they are combined or separated. We first compare lexicographic and encyclopaedic aspects of definitions, and then consider how the lexicographic and terminological aspects are related. Section 3 presents data from Princeton WordNet[1] (Fellbaum, 1998) and demonstrates the strategies used by its authors to solve the problem of the kinds-of- information diversity. Section 4 proposes guidelines for adding terminology to plWordNet, a large Polish wordnet. The problem of the diversity of kinds of information can be framed as three questions:

1. What relations should link units differing in the kind of knowledge to which they refer?
2. Are the relations sufficient to pinpoint the differences between stylistic registers?
3. How do glosses help diversify PWN units by the kind of knowledge they represent?

2 Lexicography versus terminology

Svensén (2009, p. 289) holds that most lexicographers consider boundaries between linguistic and encyclopaedic information to be fluid, and often

[1] abbreviated as PWN throughout this paper

find it hard to define what to regard as linguistic or encyclopaedic. An encyclopaedic definition may be included in typologies of lexical definitions:

> maximally rich definition, reflecting world knowledge rather than merely knowledge of the language, contains all kinds of highly specific information and a lot of practical which is not universally invalid (Geeraerts, 2003, p. 90).

It is, however, impossible to distinguish between lexical and encyclopaedic information:

> in the final account, the lexical information is determined by the encyclopaedic fact of particular real-world features. Thus the lexical information is derivable from and not independent of the encyclopaedic information [...] (Bauer, 2005, p. 127).

Some words, mainly nouns but also verbs and adjectives, should have a considerably stronger direct connection with the world than function words such as pronouns, conjunctions and prepositions (Svensén, 2009, p. 292). Dictionaries, depending on the amount and organization of the encyclopaedic element, occupy different positions on the scale of encyclopaedicity. Differences occur at the level of entries as well:

- an encyclopaedic entry is mainly headed by a common noun or a proper name, a linguistic entry – by any type of word;
- a linguistic entry is attached to the item serving as a lemma, whereas an encyclopaedic entry dealing with a certain subject could have another lemma without having to change the content of the entry (Svensén, 2009, p. 290);
- a linguistic entry may contain information about lexical and grammatical collocations of lexical units, their pragmatic functions, syntactic behaviour and so on (Fuertes-Olivera and Arribas-Baño, 2008, p. 2).

Definitions of technical and other specialized terms, like encyclopaedic definitions, do not relate to linguistic units with their universally understandable and accepted meanings, but to specific concepts established in their areas of knowledge. Traditional terminology also declares the independence of linguistics and follows its own rigorous principles:

- the onomasiological perspective (how to express a given concept);
- univocity (one term should only refer to one, clear-cut concept);
- synchrony (focus on the present meaning of terms);
- compliance of the definitions with ISO standards (Temmerman, 2000).

This may suggest a vast distance between terminography and lexicography (which treats those principles as options), but many lexicographers note that these two sciences should meet on several planes. One of the mentioned fields is Language for Specific Purposes (LSP), a term currently used to refer to specialized communication. The methodological confluence between terminography and lexicography is driven by a move away from the concept as the centre of attention.

> This change of emphasis has deep methodological repercussion, which imply the abandoning of the traditional method of onomasiological work in favor of semasiological approach which has a great deal in common with lexicography (Fuertes-Olivera and Arribas-Baño, 2008, p. 8).

Confluences between terminology and lexicology were the focus of the experiment which was designed to check whether the application of the terminological definitions will streamline the process of human-based subject indexing. Terms and descriptors (basic thesaurus units, selected to represent a specific concept in a thesaurus and in indexed documents) share these essential properties (Hudon, 1998, pp. 72-73):

- they represent single concepts in a domain,
- they are signs founded in natural language,
- they reflect language patterns established in a field of specialty.

Whereas definitions are the key component of a term bank, the heart of a thesaurus has traditionally been its relational structure. Hudon concludes, however, that definitions and relationships are assigned complementary roles in a terminological thesaurus. Definitions precisely characterize the meaning of the descriptor, while relationships pinpoint the place of the unit in the lexical hierarchy (Hudon, 1998, p. 78).

We will tackle the questions posed in Section 1, given that PWN is (among other things) a kind of thesaurus which contains both definitions and relations, and that lexical, encyclopaedic and terminological information is interrelated.

3 Terminology in PWN

In its role as a thesaurus, PWN brings together, often in the same synsets, elements of general language usage and LSP units absent from general-purpose dictionaries. Experts and laymen sometimes react differently to the same word,[2] and some words are not even in a typical layman's idiolect.[3] Svensén (2009, p. 243) notes: "To the expert, the extension of a technical terms is often small [...] whereas the intension is large".

Terminological definitions tend to refer to other terms. In a wordnet, therefore, the presence of a terminological synset requires the presence of its hypernyms, hyponyms, meronyms and so on. It makes little practical sense to put specialist language in a separate network. The difference between the professional and lay point of view is seldom clear, and even if it were, it might be too subtle to be captured by semantic relations alone.

We see two methods of putting terminology in a wordnet when the same denotation corresponds to a terminological and a general connotation:

- create two lexical units and differentiate them by the hypernyms of their two synsets;
- create one lexical unit and define it by two or more hypernyms of the synset to which it belongs (or by one hypernym if both meanings have the same *genus proximum*).

Lay and specialist meanings may also differ both in connotation and denotation. For example, the PWN 3.0 synsets ***star 1*** and ***star 3*** refer to different concepts but have the same hyponym ***celestial body, heavenly body***. The difference is signalled by glosses, by other relations (the scientific term ***star 1*** has two holonyms and several instances), and by domain (***star 1*** is linked to astronomy).

Another strategy is needed when two or more different lemmas have the same denotation, but different connotations and probably different stylistic registers. Should units belonging to general language and LSP be placed in one synset, or should they be linked by the another semantic relation, such as hyponymy (the general meaning as a hypernym of the specialist sense) or some form of relatedness?

We examined a sample of 200 nouns drawn independently and at random from a homogeneous population of PWN nouns. There are 94 common nouns belonging to general language (including those both in general and specialist registers, *e.g.,* western hemisphere), 20 proper names and 86 terms. Interestingly, most of those 200 nouns have glosses without a usage example. Only 23 synsets have usage examples and just one of them is a terminological synset. One can observe a tendency: the less encyclopaedic the noun, the more likely its usage is to be noted. Grammatical and lexical collocations are typically the kind of linguistic information not necessary in an encyclopaedic definition (Fuertes-Olivera and Arribas-Baño, 2008, p. 2).

Coming back to the role of glosses (question 3 in Section 1): they may contain information which is distinctly lexical (*e.g.,* usage examples) or encyclopaedic (*e.g.,* dates of birth and death), and so signal the character of the concept. They do not, however, pinpoint all the necessary features of terms, because they are not terminological definitions as discussed by Hudon (1998, p. 81).

A significant part of the sample, 32 lexical units, belongs to the biological taxonomy. Synsets referring to taxonomic definitions often contain several lexical units: purely scientific terms, such as Latin names of the taxa, as well as names in the vernacular. In this case, the strategy is to join in one synset all units, no matter to what register of language they belong. That is the case of the synset ***oxeye daisy 2, ox-eyed daisy 1, marguerite 1, moon daisy 1, white daisy 1, Leucanthemum vulgare 1, Chrysanthemum leucanthemum 1***.

Merging different kinds of knowledge and thus registers of language is also noticeable outside synsets, in relations between them. For example, another taxon name from our sample, ***genus Colaptes 1***, is a holonym of ***flicker 2***. It is not a species but a general name of certain woodpeckers, and its hyponyms are the names of the species of such

[2] Such a word has the same denotation (literal meaning), but different connotations (intepretations).

[3] Try *penicillamine*, *enterotoxin* and *modiolus* without peeking in a dictionary!

woodpeckers. This is one of many examples of the impossibility of organising lexical and terminological synsets in independent networks. On the other hand, to distinguish terminology from general language, terms are often linked by several *domain* relations to synsets which refer to certain domains. Such relations signal that some synsets (*e.g.,* atom) are members of the domains named by other synsets (*e.g.,* physics, chemistry). This relation has three types: topic, region and usage.

As it happens, our sample does not contain units which have counterparts with the same lemma and denotation, but different connotation, which would be signalled by double hypernymy. This may suggest that the strategy adopted by PWN authors is to not single out senses on the grounds of subtle differences of lay and specialist knowledge, but to concentrate on the same denotation.

4 A design for plWordNet

The basic element of plWordNet is not a synset, but a lexical unit (Maziarz et al., 2013), which can be assigned its own register/stylistic label and a gloss containing a usage example. It is, then, appropriate to consider distinct meanings of the same unit in different language registers. Taking into account the connection between a lemma, denotation and connotation, we propose a strategy of putting terminology into plWordNet, which considers three cases. The guidelines have already been put to a practical test: they inform the work of a team charged with adding terminology to plWordNet.

Case 1

There are two different words: a (technical) term and a word from general language, with the same denotation but different connotations, *e.g., kot domowy* 'domestic cat' and *kot* 'cat'. When two words denote the same object, their register determines whether they land in one synset or in two synsets. In plWordNet, certain pairs of registers are considered close, others – distant (Maziarz et al., 2014). The specialist and general registers are close, so we put *kot domowy* and *kot* in the same synset. Substantially different registers, *e.g.,* specialist and obsolete, are distant, so we put *pies* 'domestic dog' and *sobaka* 'dog (a borrowing from Russian, obsolete and stylistically marked in contemporary Polish)' in different synsets and link those synsets by relatedness.

Case 2

There is one word with two connotations but one denotation, *e.g., krew* 'blood'.[4] The boundary between specialist and general knowledge is not sharp: elements of specialist knowledge can enter the general vocabulary. So, we create one unit but describe it in two ways: it should have both terminological and lexical hyponyms.

Case 3

There is one word with two connotations, as well as two denotation, *e.g., para 1* 'a substance in the gas phase at a temperature lower than its critical point' (vapour) and *para 2* 'the hot mist that appears when water is boiled' (steam). We insert two lexical units and describe them differently. Different meanings of one word can be closely related. Consider, *e.g.,* the word *jeżyna*: ***jeżyna 1*** '*Rubus* L.' is a hypernym of ***jeżyna 3*** 'blackberry bush'. As this example shows, general words can be defined, via hyponymy and hypernymy, by specialist terms.

The meaning of lexical units and synsets in plWordNet – as in any wordnet – is defined principally by semantic relations. Whatever defining phrases appear in glosses have an auxiliary character. On the other hand, the role of stylistic register is noteworthy: they allow the reconstruction of specialist definition paths and distinguish them from general-language paths. We note that labels play the same role as *domain* relations in PWN.

The distinction between general and specialist registers may sometimes lead to an excessive specialisation of the meanings. This effect can be significantly reduced if encyclopaedic and lexical information is placed in the same synset.

5 Conclusions

This study has proposed a precise strategy of adding terminological senses of lexical units to plWordNet. We began by investigating the strategies adopted by the authors of PWN. While discussing the choices, we considered three aspects of a lexical unit: its lemma, denotation and connotation. There are, naturally, differences between PWN and plWordNet, due to the typological differences between the languages and the to the model adopted for plWordNet (Maziarz et al., 2013). Some choices made in the two wordnets

[4] The terminological definition is "a connective tissue composed of blood cells suspended in blood plasma", and the general-language definition is "a red fluid in animals circulating in veins and arteries".

are quite dissimilar: plWordNet avoids, for example, putting in one synset units with the same denotation, but with distant stylistic registers. It appears that register values will play a more significant role in plWordNet than in PWN, and more emphasis will be placed on differences in connotation. We observed, however, two similarities between the English and Polish wordnet. They both reflect the fluidity of the boundaries between specialist and general knowledge, and in both of them semantic relations remain the principal tool for defining senses.

The accuracy and effectiveness of the strategy we have proposed in this paper must be verified in practice by a large team of plWordNet builders. The observations thus gathered may also lead to a refinement of the strategy. The ultimate test of plWordNet with terminology in place will be its successful applications, but that is quite beyond the scope of this paper.

Acknowledgment

Financed by the Polish National Centre for Research and Development project SyNaT.

References

Laurie Bauer. 2005. The Illusory Distinction between Lexical and Encyclopedic Information. In Arne Zettersten Henrik Gottlieb, Jens Erik Mogensen, editor, *Proc. Eleventh International Symposium on Lexicography, May 2002, University of Copenhagen*, pages 111–116.

Christiane Fellbaum, editor. 1998. *WordNet – An Electronic Lexical Database*. The MIT Press.

Pedro A. Fuertes-Olivera and Ascensión Arribas-Baño. 2008. *Pedagogical Specialised Lexicography: The representation of meaning in English and Spanish business dictionaries*, volume 11 of *Terminology and Lexicography Research and Practice*. John Benjamins.

Dirk Geeraerts. 2003. Meaning and definition. In Piet van Sterkenburg, editor, *A practical guide to lexicography*, pages 83–93. John Benjamins.

Michèle Hudon. 1998. *An assessment of the usefulness of standardized definitions in a thesaurus through interindexer terminological consistency measurements*. PhD thesis, University of Toronto.

Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz. 2013. The chicken-and-egg problem in wordnet design: synonymy, synsets and constitutive relations. *Language Resources and Evaluation*, 47(3):769–796.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2014. Registers in the System of Semantic Relations in plWordNet. In *Proc. Global WordNet Conference*, Tartu, Estonia.

Bo Svensén. 2009. *A handbook of lexicography: the theory and practice of dictionary-making*. Cambridge University Press.

Rita Temmerman. 2000. *Towards New Ways of Terminology Description: The sociocognitive approach*, volume 3 of *Terminology and Lexicography Research and Practice*. John Benjamins.