

Lexical units as the centrepiece of a wordnet

Magdalena Derwojedowa¹, Stanisław Szpakowicz^{2,3}, Magdalena Zawisławska¹,
and Maciej Piasecki⁴

¹ Institute of Polish Language, Warsaw University, Warsaw, Poland

² School of Information Technology and Engineering, University of Ottawa, Ottawa,
Canada

³ Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

⁴ Wrocław University of Technology, Wrocław, Poland

Abstract

Semantic relations are the cornerstone of any wordnet. In most wordnets created thus far, however, the nature of the arguments of such relations is unclear. The synset, usually considered the basic unit of a wordnet, relies on the notion of *synonymy*, distinctly difficult to define, and on the even more vague notion of a *concept*. We argue that the synset should be consistently replaced by the lexical unit as the fundamental unit of a wordnet. A wordnet is a lexical database that operates as a dictionary or a thesaurus. Users expect *words* in such a resource, and they can understand words even without the benefits of syntax. Semantic or lexico-semantic relations, then, link meanings bound to words, or more generally to lexical units. The practical effect of this work is the adoption of lexical units as the basic building blocks of a Polish WordNet named *plWordNet*.

Keywords: wordnets, semantic relations, synonymy, synset, lexical units

1 Introduction

The synset, the core data structure in the original Princeton WordNet (Miller *et al.*, 1990), is a set of words — or, to be more precise, *lexical units* (LUs) — which express the same *concept*, that is to say, a set of synonymous LUs or, simply put, a set of synonyms. Any semantic relation in WordNet should thus be a relation between two such concepts or, in practice, sets of synonymous LUs that represent the concept. On the other hand, some relations (such as *derived_from* or *antonymy*) are defined as holding between lexical units rather than synsets (Miller *et al.*, 1990; Fellbaum, 1998). We believe that such discrepancy reveals two fundamental weaknesses of the notion of a synset: the unwarranted and unnecessary domination of synonymy over other lexico-semantic relations, and the vagueness of the basic entry in the lexical knowledge base. We will argue that a different decision on the basic unit in a wordnet makes the resource linguistically more sound and lexicographically more convenient to create and work with.

Our argument unfolds as follows: first we look more closely at the lexical unit and at synonymy, next we propose our perspective on the synset. In the remainder

of the paper we briefly examine the consequences of the new approach, in particular its effect on the size of a typical synset.

2 Definitions

A *lexeme* is a set of inflectional forms that share certain meaning. In text, lexemes are represented by words, or rather wordforms with specific values of inflectional categories. Most occurrences of wordforms are strings delimited by punctuation, but occasionally punctuation (in particular, spaces) is part of a wordform too. Examples include English phrasal verbs and reflexive pronouns in many languages (e.g., German *sich treffen* ‘meet’, French *se reposer* ‘rest’ or Polish *uśmiechać się* ‘smile’).

A *collocation* is a semantically compositional fixed multiword expression that originated in a syntactic construction, but over time became lexicalised. Examples: *gazeta codzienna* ‘daily newspaper’, *bilet miesięczny* ‘monthly (transportation) pass’, *kawa z mlekiem* ‘coffee with milk’ *maszyna do szycia* ‘sewing machine’. An *idiom* can be regarded as a special case of a fixed multiword expression whose meaning is not compositional, e.g., *kopnął w kalendarz* ‘kicked the bucket’ or *leje jak z cebra* ‘comes down in buckets’ (Lyons, 1989; Firth, 1957).¹

3 Synonymy

There is ample linguistic evidence that *strict* synonymy does not exist at all (Bloomfield 1933, 145, Hockett 1964, 15.1, Lyons 1989, 9.4, Apresjan 2000, 207, Edmonds and Hirst 2002—to quote a few). The idea of near-synonymy, however, is natural and non-controversial. On the other hand, at least one large-scale linguistic enterprise favours the ubiquity of synonymy (Apresjan, 2000). What is more, both viewpoints are not just defensible; they are linguistically sound. The co-existence of such contradictions is possible because, although synonymy is in both cases the same relation, it is applied to different linguistic objects.

Probably the simplest definition of synonymy says that it is mutual hyponymy (Lyons, 1989, 9.4). In other words, if A is a synonym of B, then “A is a kind of B” and “B is a kind of A”. For example, *ascending* is a kind of *going up* and the other way around, and so are *animal* and *beast* as well as *hot*, but though *girl* is a kind of a *woman*, not all women are girls. Another definition relies on mutual substitution in a context. A and B are synonyms if in a given context A can be substituted for B and B for A. Let us look at an example:

- (1) a. They ascended pretty fast.
 b. They went up pretty fast.
 c. They moved pretty fast.

The example shows the drawback of this definition of synonymy: it is quite likely that a context allows hypernyms as well. We believe, however, that the mu-

¹This distinction is intuitive enough for any lexicographer engaged to write wordnet entries. A detailed discussion of the differences between idioms and collocations, including recognition tests and diagnostic contexts, would be beyond the scope of this paper (cf. Broda et al. 2007).

tual substitution approach enables more subtle and less arbitrary discrimination of (near-)synonyms. For example, we can recognize synonymy between *chick* and *skirt* ‘young woman seen in the context of sexual attractiveness’, but not between any of these and *girl*. We feel that the inevitable drawbacks of exact synonymy are a myth (albeit a very persistent myth).

Here are a few motivating examples: *credible* and *believable*, *toilet* and *lavatory*, *aerial* and *antenna*. The pairs are synonymous apart from the register of the speech or dialect.

Near-synonyms are an imperfect, but quite effective, replacement for strict synonyms, but they bring a slightly different hazard into the picture. The term “synonym” comes from Greek *syn-ónymos* ‘having the same name’. If we took this intention literally, strings such as *orphan*, *parentless child* and *a child whose both parents died* would naturally be considered synonyms. Put this way, the whole derivation from a deep structure to the surface can be seen as selecting some synonymous means of expression from a set of possible expressions (cf. Apresjan 2000, 49), not only lexical but also syntactic. This is the danger of any approach to synonymy: a LU can in principle be replaced by its gloss. They can be substituted for each other, they can be used in the same context, they represent the same concept. We will now examine ways of avoiding such “proliferation” of synonymy.

4 Lexical units

Synonymy in the widest sense suggested at the end of the previous section would make any dictionary of synonyms a grammar (cf. Mel’čuk 1988). That is why, for lexicographic purposes, it is reasonable to assume that the basic entry is a word (lexeme) or at least a word-like linguistic entity (so, for example, reflexive pronouns are acceptable, and so are prepositions or adverbs in phrasal verbs, but few phrases or clauses would be). A lexical unit (LU) is, therefore, a word *sensu largo*, that is to say, it may be an idiom or even a collocation, but not a productive syntactic structure.

To make it more formal: a lexical unit is a string that has its morphosyntactic characteristics and a meaning as a whole. So *be on the mend*, *put the wind up*, *white collar*, (German) *spazieren gehen* ‘go for a walk’, *die Rede haben* ‘give a speech’, *ein Loch in den Bauch fragen* ‘badger’, (French) *donner un coup de doigt / pied* ‘punch / kick’, *courir à perdre haleine* ‘run very fast’ or Polish *być wysokiego / niskiego / średniego wzrostu* ‘be tall / short / of average height’ or *o mały włos* ‘by the skin of one’s teeth’ are all lexical units, but *go to a barber* or *buy a newspaper* are not.

In consequence, substrings within a LU have no meaning or inflection of their own, so they can be treated just as morphemes are inside a morphological structure (cf. Derwojedowa and Rudolf 2003). In other words, a LU is syntactically non-compositional (a terminal), but not necessarily semantically non-compositional.

5 Synsets, concepts and wordnets

In the Princeton WordNet, a synset is a pair {concept, words expressing this concept}. But what is a word? It is a pair {signifier, signified} or a vehicle (*signifiant* in de Saussurian terms) and a meaning (*signifié*). A *concept* should therefore be an extralinguistic, mental entity (cf. Miller *et al.* 1990). Let us look at the process of identifying a LU to be a representation of a concept: a concept (for example, *hand*) is recognized by the lexicographer, then words such as *hand*, *mitt*, *manus*, *paw* are recognized as having the common meaning *hand*. Consider the examples:

- (2) a. his hands of the pianist / surgeon
b. ?his paws of the pianist / surgeon
- (3) a. her manicured hands with long, slim fingers
b. ?her manicured mitts with long, slim fingers
c. ?her manicured paws with long, slim fingers

If examples 2b, 3b, 3c were acceptable (rather than, as we marked them, questionable), then a *paw* or a *mitt* would be a kind of *hand*, and at the same time a *hand* would be a kind of *paw* or *mitt*. The shared contexts would suggest that the meaning of all these words is the same. In fact it is not true, because the obvious intuition is that a *hand* is *not* a kind of *paw*, despite WordNet 3.0's verdict that {*hand*, *manus*, *mitt*, *paw* => *extremity*}.² Also, *mitt* and *paw* would be co-hyponyms rather than synonyms, because their meanings differ in other contexts.

It is a little unfortunate that WordNet's fundamental idea of a *concept* is so hard to pin down. In Miller *et al.* (1990) the terms *concept*, *lexicalised concept* and *meaning* are used interchangeably. Although none of them has a clear definition, it appears that they are also treated as equivalent, thus making for more confusion. In addition, various languages lexicalise concepts variously. For example, chairs, sofas, armchairs etc. have no lexicalised generalisation in Polish. In French, there is a single word: *siège*. The obvious gap in the Polish hypernym hierarchy must be filled by an artificial lexical unit *meble do siedzenia* 'furniture for sitting'.

The relationship between concepts and real objects is another source of worry for wordnet creators. An old scholastic maxim proclaims that *vox significat rem mediantibus conceptibus* [a word (*lit. voice*) signifies a thing through the medium of concepts] (Lyons, 1989, 4.1). Let us illustrate the problem with the various takes on the nature of the *dandelion*. Depending on the circumstances, it can be seen from various angles. All statements in the example below are true:

- (4) a. Dandelions are weeds.
b. Dandelions are flowers.
c. Dandelions are herbs.

We will not, however, normally recognize the words *weed*, *flower* and *herb* as synonyms. That is because word meaning consists of two elements—a reference to an object and a connotation (that is to say, objective features and features

²A successful test for synonymy would conclude that a paw is a kind of hand and a hand is a kind of paw.

attributed to the object by the speaker). Thus, *weed* is *an undesirable, wild plant, herb* is *a medicinal plant* and *flower* is *an ornamental plant*; *undesirable*, *medicinal* and *ornamental* are connotations. In effect, a wordnet should place the word *dandelion* in three different places in the hierarchy, to signal the differences in hypernyms and co-hyponyms.

The main point of the theory of a “concept behind synset” seems to be misguided. In a language the tendency toward precision is balanced by the economy. Pure synonyms break this rule — they are redundant (uneconomical) — while near-synonyms differ (precision), so do not share the meaning (Buttler *et al.*, 1986; Apresjan, 2000). A hypernym expresses a more general meaning. What is presented as a concept in WordNet is just the meaning of this hypernym, or it must be this part of the meaning of all words in a synset that is common — and this is probably the nearest to the more general LU, thus a hypernym (cf. Lyons 1989). We see no reason to think about synonymy in a special way and make it a fundamental relation in a wordnet.³ Instead we propose a purely linguistic basis for the construction of a wordnet: we take it to be a network of LUs connected by lexico-semantic relations. A synset is in this case just a “shortcut” for two or more LUs sharing the same set of relations.

Unless we have glosses at our disposal, the only means of discriminating homonymous LUs is to compare the networks of their relations. For example, H_2O and *water*⁴ are synonymous in the sense that water is a common name of the chemical compound; *sparkling water* and *still water* are hyponyms of *water* (but not of H_2O). On the other hand, *distilled water* is not a kind of drinking water, but rather a chemical compound of certain properties, thus a kind of H_2O . In effect, there are two LUs in WordNet: *water*₁ that is a synonym of a H_2O and a chemical compound, and *water*₂, a drink. If we had to lump those two LUs together, we would have to accept that, improbably, {**water**, H_2O } is a kind of {**drink**, **beverage**, ...}.

In a wide sense of the term, a *lexical relation* or *lexico-semantic relation* is any semantic relation between two LUs. It follows that antonymy, synonymy, hypernymy and lexical derivational relations (such as *derived_from* or *relational_adjective*) are lexical relations. In the Princeton WordNet few lexical derivational relations are present, but they are quite numerous in EuroWordNet (Vossen, 2002). In the Polish WordNet named *plWordNet* (Derwojedowa *et al.*, 2008), there are no separate classes of such relations. Instead, selected types of derivation are assembled into *pertainymy* and *relatedness* (Derwojedowa *et al.*, 2007). In the narrow interpretation of the term, a lexical relation is a relation of the sense (cf. Lyons 1989, 9.1 ff.) — that is to say, hyponymy, meronymy, synonymy etc. are lexical relations, but derived forms are mutually linked in some other way (they usually belong to the domain of grammar).

Miller *et al.* 1990 distinguish between lexical relations and semantic relations. The former link lexemes, the latter – meanings. Thus for example synonymy and antonymy are lexical, but hypernymy or meronymy are semantic. We do not follow

³What makes synonymy special is its rareness.

⁴A synset in WordNet 3.0.

this distinction. We believe that in a vocabulary⁵ lexical units are bound by relations.⁶ We assume that a *lexico-semantic relation* is (basically) non-derivational and is a kind of operation on senses of LUs. Technically, it is captured by diagnostic tests such as “A is kind of B (and there are other B)” (“a tulip is a kind of flower, and there are other flowers”, an instance of hyponymy) or “A is a kind of B and B is a kind of A” (“a spud a kind of a tater and a tater it is a kind of spud”, an instance of synonymy; cf. Vossen 2002; Derwojedowa et al. 2007).⁷ The number and types of such operations depend on the underlying theory: besides commonly accepted relations such as synonymy or antonymy there are elaborate systems like Mel’čuk’s lexical functions (Mel’čuk, 1988).

6 Consequences and a brief conclusion

We examined the notions of synonymy, synsets and lexical units. In contrast with the practice in the Princeton WordNet, we propose that a wordnet should be a network of *lexical units* linked by lexico-semantic relations. A lexical unit is a linguistic entity with certain semantics and certain morpho-syntactic characteristics. The semantics of a LU is determined by the network of relations in which it participates.

The view of the lexico-semantic relations presented in this paper has significant consequences for a wordnet that adopts this view; plWordNet (Derwojedowa et al., 2008) did. Crucially, most synsets are small or very small (in return, there are quite numerous homonymous synsets, e.g. *palący1*, *pikantny* ‘hot, spicy’, *palący2* ‘burning (sun)’, *palący3*, *pilny* ‘burning (problem), urgent’. On the other hand, often many co-hyponyms are dominated by one hypernym. Table 1 shows a snapshot of plWordNet (as of spring 2008): synsets larger than three units are rare.⁸

	1	2	3	4	5	6	7	8	9	>= 10
all synsets	53.27	25.02	12.42	5.25	2.11	0.91	0.43	0.18	0.17	0.22
nouns	65.92	19.48	7.92	3.83	1.38	0.62	0.36	0.13	0.17	0.19
verbs	10.63	43.88	25.64	10.83	5.18	2.09	0.61	0.40	0.27	0.48
adjectives	51.49	25.67	14.55	4.85	1.77	0.81	0.51	0.15	0.10	0.10

TABLE 1: LUs in a synset (%)

It is too early to assess fully the effect of adopting a consistently lexical design principle in our wordnet. A very preliminary comparison of the present average size of a plWordNet synset with the average size in the Princeton WordNet 3.0 is given in Table 2, but for more reliable evaluation we have to wait until the current phase of development will be concluded.

⁵Any wordnet, dictionary or thesaurus is a way of putting the vocabulary, that is to say, a set of LUs of a language, in order.

⁶Here we follow the old tradition of Trier and Porzig, cf. Lyons 1989.

⁷In plWordNet some derivational relations are present, partially for consistency with other wordnets, partially because “pure” lexical relations are far too few to weave a network for the whole vocabulary.

⁸In most cases the latter are expressive LUs, e.g. *beznadziejny*, *kiepski*, *marny*, *fatalny*, *lichy*, *podły*, *denny*, *nędzny*... ‘hopeless, lousy, meagre, fatal, miserable, rotten, crummy, poor...’.

	plWordNet	Princeton WordNet
nouns	1.618	1.782
verbs	2.722	1.819
adjectives	1.863	1.652

TABLE 2: Average number of LUs in a synset — on the basis of Miller *et al.* (2007)

References

- J. D. APRESJAN (2000), *Semantyka leksykalna. Synonimiczne środki języka [Lexical semantics]*, Warszawa, przeł. Z. Kozłowska and A. Markowski.
- L. BLOOMFIELD (1933), *Language*, New York.
- B. BRODA, M. DERWOJEDOWA, and M. PIASECKI (2007), Recognition of Structured Collocations in an Inflective Language, in *Proceedings of IMCSIT 2007*, <http://2007.imcsit.org/>.
- D. BUTTLER, H. KURKOWSKA, and H. SATKIEWICZ (1986), *Kultura języka polskiego [The culture of Polish language]*, volume I-II, Warszawa.
- M. DERWOJEDOWA, M. PIASECKI, S. SZPAKOWICZ, and M. ZAWISŁAWSKA (2007), Relacje w polskim WordNecie [Relations in Polish WordNet], Technical Reports 1, Politechnika Wrocławska, Instytut Informatyki Stosowanej.
- M. DERWOJEDOWA, M. PIASECKI, S. SZPAKOWICZ, M. ZAWISŁAWSKA, and B. BRODA (2008), Words, Concepts and Relations in the Construction of Polish WordNet, in *Proceedings of GWC 2008*, pp. 162–177.
- M. DERWOJEDOWA and M. RUDOLF (2003), Czy Burkina to dziewczyna i co o tym sądzą ich królewskie mości, czyli o jednostkach leksykalnych pewnego typu [... On certain type of lexical units], *Poradnik Językowy*, (5).
- P. EDMONDS and G. HIRST (2002), Near-Synonymy and Lexical Choice, *Computational Linguistics*, (28(2)):105–144.
- Ch. FELLBAUM, editor (1998), *WordNet. An Electronic Lexical Database*, MIT Press.
- J. R. FIRTH (1957), *Papers in Linguistics 1934-51*, chapter Modes of Meaning (1951), pp. 190–215, London.
- B. HAMP and H. FELDWEG (1997), GermaNet – a Lexical-Semantic Net for German, in *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid.
- C.F. HOCKETT (1964), *A course in modern linguistics*, Macmillan, New York.
- J. LYONS (1989), *Semantyka [Semantics]*, PWN, przeł. A. Weinsberg.
- I.A. MEL'ČUK (1988), *Dependency Syntax: Theory and Practice*, State University of New York Press.
- G. A. MILLER, R. BECKWITH, Ch. FELLBAUM, D. GROSS, and K. MILLER (1990), Introduction to WordNet: An On-line Lexical Database, *International Journal of Lexicography*, 3(4):235–244.
- George A. MILLER, Christiane FELLBAUM, Randee TENGI, Susanne WOLFF, Pamela WAKEFIELD, Helen LANGONE, and Benjamin HASKELL (2007), WordNet — a lexical database for the English language, URL <http://wordnet.princeton.edu/>, homepage of the project.
- P. VOSSEN (2002), EuroWordNet General Document Version 3, Technical report, University of Amsterdam.

