# What WordNet does not know about selectional preferences[1]

Michal Boleslav Měchura
*Fiontar*, Dublin City University

*Selectional preferences are the tendencies of words to co-occur with other words that belong to certain semantic types. In this paper, I will investigate how closely these corpus-attested preferences correspond to WordNet. For example, for all possible direct objects of cancel, is there a single category (or a union of several categories) in WordNet that subsumes them, and only them? Selectional preferences manifest themselves in authentic texts and can be revealed through corpus analysis. I will introduce an experimental tool I have built which attempts to do this automatically by aligning corpus-extracted lists of collocates (for example a list of the direct objects of cancel) with WordNet. The strength of this method is that it can discover and name selectional preferences automatically, but its weakness is that it can only do so when WordNet contains a suitable category. We will see that WordNet often lacks a category (or even a union of several categories) that fully corresponds to an attested selectional preference – for example, there is no category in WordNet that includes all the kinds of events that can be direct objects of cancel (meeting, wedding, concert etc.) but excludes those that cannot (accident, sunset, invention etc.).*

## 1. Introduction

Selectional preferences, sometimes also called selectional restrictions, are a well-known phenomenon in linguistics. In this paper, I will work from the following definition: a selectional preference is the tendency of a *base word* to fill a particular slot in its valency pattern with a *collocate* of a particular semantic type. For example, the verb *eat* is transitive, it has a slot in its valency pattern for a direct object and has a tendency to fill this slot with nouns and noun phrases from the semantic type FOOD: the unmarked usage is to talk of eating things like *bread, lunch* and *jam* but not *scissors, money* or *weather* (unless one is speaking metaphorically or somehow extending the meanings of the words involved). Note that a selectional preference is specific to a syntactic role: *eat* has a preference for FOOD in its direct-object slot but not in its subject slot. A selectional preference can then be recorded formally as a triple consisting of a base word, a role and a semantic type: <*eat*, direct-object, FOOD>.

Selectional preferences are arbitrary. In some cases, it may be possible to explain a selectional preference away as a logical consequence of meaning, as one can in the trivial case of *eat*: we talk of eating food because we do eat food. In other cases the motivation for the selectional preference is less clear: we *join* things like *political parties* and *armies*, but we only *enlist* in *armies*: to say that one is enlisting in a political party is odd. So, the verbs *join* and *enlist* have different selectional preferences in spite of being (near-)synonyms. There does not necessary need to be an explanation for why this is so, it simply is so. The important thing is that knowledge about selectional preferences can help explain the difference between the usage of what seem like synonyms.

Looking at selectional preferences cross-linguistically, they can often help explain how a pair of putative translational equivalents differ from each other. The English verb *drink* has a preference for LIQUID but its Irish equivalent *ól* also has an additional preference for TOBACCO (*d'ól sé píopa*, literally 'he drank a pipe'). The English verb *subscribe* has a preference for PERIODIC PUBLICATIONS (*she has subscribed to the newsletter*) and IDEOLOGIES (*he subscribes to the notion that people should be free to choose*) but its Czech equivalent *předplatit* only has a preference for the former. Again, these differences can sometimes be explained away as a consequence of meaning (the Czech *předplatit* is transparently derived from *před* 'pre' +

[1]This paper is a summary of the author's M.Phil. dissertation (Měchura 2008) at Trinity College, University of Dublin.

*platit* ‘pay’) but in many other cases the difference must be taken as arbitrary. For example, the selectional preferences of the English verbs *drive* and *ride* do not overlap with those of their German equivalents *fahren* ‘drive’ and *reiten* ‘ride’: in English we *ride* motorcycles and bicycles, but in German people ‘drive’ (*fahren*) them, the verb *reiten* ‘ride’ is only for animals like horses (Soehn 2005).

## 2. Aligning selectional preferences to WordNet

It is reasonable to assume, in my opinion, that selectional preferences observed in authentic texts are an external manifestation of the internal organization of the mental lexicon. There is probably a category in the speaker’s mind for things that can be *eaten*, another category for things can be *cancelled*, another for things that can be *subscribed to*, *joined*, *enlisted in*, *driven* and so on. In other words, there is an ontology in the mind and selectional preferences are its observable traces. The purpose of this paper is to find out how this mental-internal ontology relates to WordNet (Fellbaum 1998). For example, for all nouns that typically occur as the direct objects of *join*, is there a single category in WordNet (or a union of several categories) that subsumes them all, and which at the same time excludes any nouns that cannot occur as its direct objects?

Technically, this is a fairly easy problem to solve. We need to extract from a corpus a list of the most frequently occurring collocates of a particular base word in a particular syntactic role, for example the direct objects of *cancel*. We can use a corpus query system such as the Sketch Engine (Kilgarriff et al. 2004) for that. Once we have such a list, we can try to see whether there is a category in WordNet that subsumes all or most of the items on that list. WordNet is a large-coverage network of words, of the concepts they denote, and of the relations between the concepts such as hyponymy and meronymy. So WordNet can be used to obtain a large number of sets containing words for *kinds of* other things (kinds of people, kinds of events), *parts of* things (parts of a house, body parts), things that *contain* certain parts (things that have blades, things that have wings) and a handful of other kinds of sets such as words that belong in a particular domain (words from law, biology), adjectives that represent values of a quality (values of size, speed) and so on. I have derived from WordNet thousands of lexical sets of this nature, ranging in size from just two words to thousands of words, where the largest set is “kinds of entity”, the set containing all nouns in WordNet. I have built a tool (called SenseMaker) which, when given a list of corpus-extracted collocates, finds the best matches for that list among its database of WordNet-derived lexical sets. I have used SenseMaker to conduct several experiments to find out how well selectional preferences align with WordNet.

Before we proceed to the results of the experiments, some technical aspects of this method need to be dealt with. There is an existing line of research in natural language processing on using corpora and WordNet to machine-learn selectional preferences, starting with (Resnik 1993). The method has some well-documented limitations, mostly related to noise of one kind or another. They are detailed in the dissertation on which this paper is based (Měchura 2008) and we will not deal with those here. The exact structure of SenseMaker’s database and its comparison algorithm is also documented fully in the dissertation and will not be dealt with here.

SenseMaker displays results as a listing of lexical sets that best match the collocate list, ordered by a score of “match quality” (see Figure 1). A link is available next to each lexical set that leads to a list of all words in that lexical set, including those that did not occur on the

collocate list. This can be used to make predictions: if nouns like *lunch* and *dinner* occur as objects of some verb, and if these are matched to the set 'kinds of meal' by SenseMaker, then SenseMaker is essentially making the prediction that other members of the set, such as *breakfast* and *snack*, can also appear as objects of the same verb. One can verify these predictions against intuition to determine whether SenseMaker has overgeneralized or not.

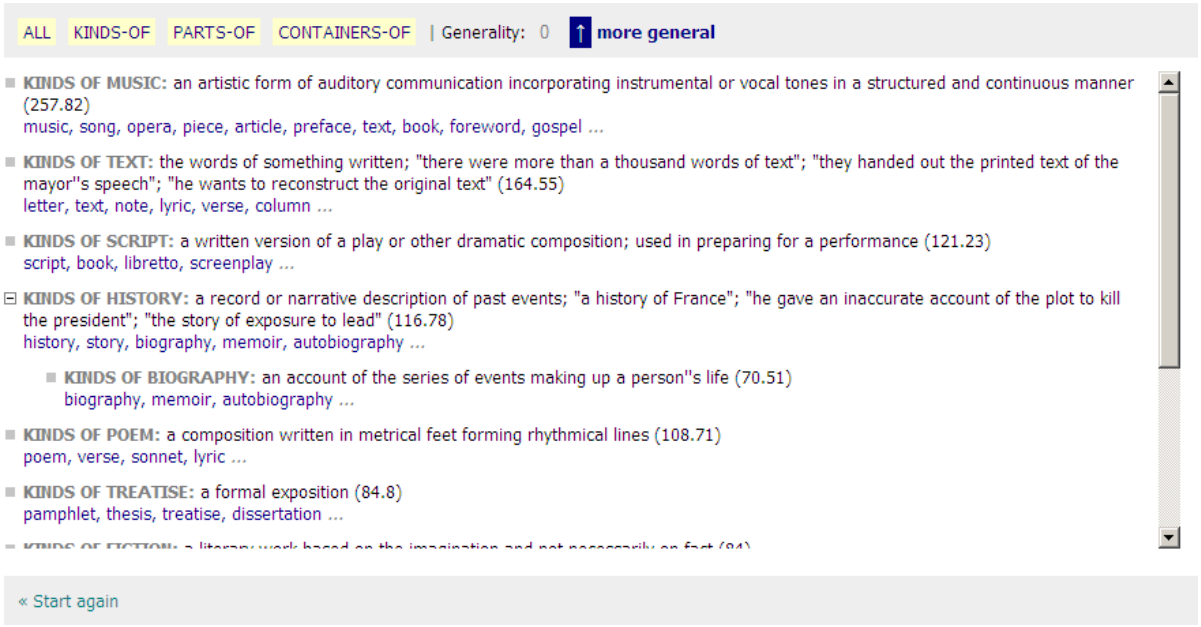


Fig. 1. SenseMaker's generalizations for the 50 most salient direct objects of the verb 'write'.

## 3. The results

I have conducted several experiments in which I have used SenseMaker to try and match corpus-extracted collocate lists against WordNet-derived lexical sets. The corpus used for these experiments was the British National Corpus and the corpus query system was the Sketch Engine. Three experiments will now be presented as examples that yield typical results.

### 3.1. The subjects of 'live'

The top 100 subjects of the verb *live* are straightforward as the large majority of them refer to living things, the most prominent subset being people. SenseMaker matches them correctly to the synset ORGANISM to its hyponym PERSON.

There are cases when the subject of *live* refers not to a person but to a group of people such as *family*, *population* and *generation*. The nearest match for these in WordNet is the synset SOCIAL GROUP but this does not include *population* and *generation*. The nearest hypernym of SOCIAL GROUP that does include *population* and *generation* is GROUP, but this is an overgeneralization as it subsumes groups of non-living things as well. Another candidate is the synset PEOPLE, glossed as '(plural) any group of human beings (men or women or children) collectively' – this does include *population* and *generation* but does not include many others such as *family* and *household*. Neither of the meronymy-derived lexical sets is of any help, either: the synset PERSON is a meronym of PEOPLE and its hyponyms are meronyms of a handful of other synsets (including PARENT, CHILD and SIBLING which are meronyms of FAMILY), but there is no way in WordNet to infer that, for example, HOUSEHOLD contains PERSON. It turns out that WordNet does not have a category that subsumes all – an only – kinds of groups of people.

Surprisingly, no metaphorical uses of *live* (such as *the legacy lives on*) are frequent enough in the British National Corpus to have made it to the list of the 100 most salient subjects of the verb.

### 3.2. Direct objects of 'cancel'

SenseMaker reveals that most of the things we can *cancel* are kinds of EVENT. This is intuitively valid, but how precise is it? Some of the predictions it makes include events like *occurrence*, *incident* and *crash*. It is odd to say that somebody *cancels the occurrence* of something or that someone *cancels an incident*. Upon reflection, it turns out that one can only cancel events that have been pre-planned, such as *meeting*, *wedding* and *concert*. Therefore, for an event to qualify as a direct object of *cancel*, it has to be capable of being pre-planned. No such category exists in WordNet but some prominent subsets do, such as MEETING, JOURNEY and SHOW.

There is often an implication that if one cancels an event, the event has not happened yet. For example, cancelling a wedding normally implies that the wedding hasn't happened yet. But you can also cancel things that are already in progress such as *subscription*, *contract* and *registration*. These are not EVENTS but STATES or more precisely STANDING ARRANGEMENTS. Subscriptions and contracts are arrangements which usually last a long time and can be made to cease to exist (that is, cancelled) while they are in progress. Again, there is no such category in WordNet but some of its prominent subsets are, including CONTRACT and ARRANGEMENT, which SenseMaker reveals.

There is some conceptual overlap between the two types of things one can cancel. For example, when you book a trip with a travel agent and then you cancel the booking, you are at the same time cancelling the standing arrangement that exists between you and the travel agent, as well as the pre-planned event of going on the trip. Additionally, there are cases of metonymic extension which confuse SenseMaker, the most striking of which is *milk*. In the phrase *cancel the milk*, the *milk* does not of course refer to the white liquid that comes out of a cow's udder, it refers to the standing arrangement of having one's milk delivered to the house every morning. Thus *cancel the milk* belongs in the same group as *cancel the subscription* and *cancel the contract* but SenseMaker does not group it as such because WordNet does not contain the sense of *milk* as 'arrangement to have milk delivered'.

### 3.3. Nouns modified by 'immediate'

The adjective *immediate* conveys roughly the idea that two things are adjacent, without anything intervening between them. This broadly defined meaning is exploited copiously in discourse to connect physical objects with their surroundings (*in the immediate vicinity of the airport*), to connect events to events that follow (*their immediate reaction was negative*) or to events that precede (*the immediate cause of death was in the stomach*), to connect people to other people (*he reversed the policies of his immediate predecessor*), and a host of other meanings which resemble one of the above to varying degrees.

If we wish to find and name *immediate*'s selectional preferences, we can start by grouping the corpus-extracted collocates manually into the three categories mentioned above (SURROUNDINGS, EVENTS and PEOPLE) and see if SenseMaker finds any matches for them in its database of lexical sets derived from WordNet. Let us start with SURROUNDINGS. This list includes nouns like *vicinity*, *surroundings*, *neighbourhood*, *environs* and *hinterland*. SenseMaker finds that the most likely candidates are the synsets GEOGRAPHICAL AREA and SECTION (the latter is glossed as 'a distinct region or subdivision of a territorial or political

area or community or group of people') but neither contains all the words (GEOGRAPHICAL AREA does not include *locality*, SECTION does not include *hinterland*) and both make some incorrect predictions: we do not normally say *the immediate town* or *the immediate meadow* even though they are both kinds of GEOGRAPHICAL AREA.

Very similar results obtain when we use SenseMaker to find matches for the other categories mentioned, that is, nouns that denote EVENTS and nouns that denote PEOPLE. We end up with synsets which appear intuitively correct but turn out to be too general because they make incorrect predictions.

Interestingly, SenseMaker matches some of the words from the collocational sets with the lexical set "terms from law". The words matched include *interest*, *relief*, *cause*, *effect*, *action*, *answer*, *use* and *release*. While not all of them are specifically law terms, not even when used in combination with *immediate*, this does correctly suggest that *immediate* belongs in a formal register.

## 4. Discussion

The single most important finding to conclude from the experiments is that selectional preferences are hard to "pin down". We have seen that what seems like a single concept, for instance GROUP OF PEOPLE, may in fact have several potential matches in WordNet, none of which is completely satisfactory. The reason why WordNet does not reflect selectional semantic types accurately is because it was never designed to. Essentially, WordNet was compiled by asking informants questions such as *do you agree that a car accident is an event? do you agree that to be snoring, you must also be sleeping?* and so on (Miller and Fellbaum, 2007: 270ff) (in the case of WordNet the informants are mostly the compilers themselves but that is beside the point). These questions make an appeal to the informant's introspection and the answers to them are a product of reasoning. Thus, it is hoped, the internal organization of the mental lexicon is revealed.

But there is another way to reveal the internal organization of the mental lexicon, and that is to observe selectional preferences in action. We observe that humans combine *immediate* with *surrender* in the same way that they combine *immediate* with *impression*, and thus we can conclude that *surrender* and *impression* have something in common. Unlike the question-and-answer approach, this bypasses the informant's explicit reasoning capacity and instead focuses on his or her instinctive language use. Thus, it is a more direct window onto the mental lexicon. It is, after all, a well-known fact in empirical linguistics that people do not always use language in the way they claim they use it. Therefore, it is not unexpected that there should be a discrepancy between what exists in the mental lexicon and what people claim exists in it when prompted.

## 5. Conclusion

The purpose of this paper has been to find out whether selectional preferences correlate with the semantic types inherent in WordNet. The answer to that question is "not completely". The same answer is likely to apply to other WordNet-like databases. Having recognized this, the next obvious question is to ask, what should an ontology actually look like if it were to reflect accurately the semantic types involved in selectional preferences – but answering that question is not the objective of the present paper.

## Bibliography

Fellbaum, C. (ed.). 1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Kilgarriff, A.; Rychlý, P.; Smrž, P.; Tugwell, D. (2004). 'The Sketch Engine'. In *Proceedings of the 11th Euralex International Congress*. Lorient: Université de Bretagne Sud.

Měchura, M. (2008) *Selectional Preferences, Corpora and Ontologies*. M.Phil. dissertation, Trinity College, University of Dublin.

Miller, G. A.; Fellbaum, C. (2007). 'Semantic networks of English'. In Hanks, P. (ed.). *Lexicology: Critical Concepts in Linguistics*, volume 6. Oxford: Routledge.

Resnik, P. S. (1993). *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania.

Soehn, J.-P. (2005). 'Selectional restrictions in HPSG: I'll eat my hat!'. In Müller, S. (ed.). *Proceedings of the 12th International Conference on Head-Driven Phrase Structure Grammar*. University of Lisbon.

British National Corpus. http://www.natcorp.ox.ac.uk/.

The Sketch Engine. http://www.sketchengine.co.uk/.