Word Association Thesaurus As a Resource for Building WordNet

Anna Sinopalnikova

Masaryk University, Brno, Czech Republic
Saint-Petersburg State University, Russia
Email: anna@fi.muni.cz

Abstract. The goal of the present paper is to report on the on-going research for applying psycholinguistic resources to building a WordNet-like lexicon of the Russian language. We are to survey different kinds of the linguistic data that can be extracted from a Word Association Thesaurus, a resource representing the results of a large-scaled free association test. In addition, we will give a comparison of Word Association Thesaurus and other language resources applied to wordnet constructing (e.g. text corpora, explanatory dictionaries) from the viewpoint of the quality and quantity of information they supply the researcher with.

1 Introduction

Since 1985 methodology of wordnet building has undergone significant changes. Starting with the primarily psycholinguistic techniques adopted in the Princeton WordNet (PWN), it switched to the entirely different methodology of the EuroWordNet (EWN) project based on the usage of existing resources, either the PWN itself within the expand model, or available national language resources within the merge model.

In this article we will introduce a connecting link between those two methodologies and present a resource, which, on the one hand, contains psycholinguistic data, but on the other hand, in a well-structured form that makes it computer-processable and, thus susceptible of both PWN and EWN methods.

In the second part of the paper we define some basic notions of psycholinguistics, necessary for the further discussion. Section 3 is dedicated to observation of different types of the empirical linguistic data derived from WAT and applied to wordnet constructing. In the last section we will compare the results of WAT usage with that of text corpora from the viewpoint of their coverage.

2 Basic Concepts

Originally the term '**association**' was used in psycholinguistics to refer to the connection or relation between ideas, concepts, or words, which exists in the human mind and manifests in a following way: an appearance of one entity entails the appearance of the other in the mind; thus '**word association**' being an association between words. In modern studies this term is often expanded to the scope of corpus linguistics and lexicography, but we will use it in its traditional sense.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 199–205.

The simplest experimental technique to reveal the association mechanism is a '**free association test**' (FAT). Generally, a list of words (**stimuli**) is presented to subjects (either in writing or orally), which are asked to respond with the first word that comes into their mind (**responses**). As opposed to other, more sophisticated forms of association experiments (e.g. controlled association test, priming etc.), FAT gives the broadest information on the way knowledge is structured in the human mind.

The results of FAT series carried out with several hundreds stimuli and a few thousand subjects, reported in a form of tables, were given the name '**Word Association Norms**' (WAN). The body of WAN constitutes the list of stimuli, lists of responses with their absolute frequencies for each stimulus word. Along with the response distribution, frequency of response is considered to be an essential index, reflecting the strength of semantic relations between words.

The first WAN were collected by Kent and Rosanoff [1] on the base of the list of 100 stimulus words including common nouns and adjectives, and 1000 subjects being involved. Since then, numerous WAN for many European and Asian languages (monolingual, as well as bilingual and trilingual) were published using mostly Kent and Rosanoff list of stimuli and expanding their experience to other languages, e.g. [2,3,4].

Word Association Thesaurus (WAT) is quite similar to WAN, but it excels significantly in size (it includes several thousands of stimuli). Also the procedure of data collection is much more complicated: a small set of stimuli is used as a starting point of the experiment, responses obtained for them are used as stimuli in the next stage, the cycle being repeated at least 3 times. In so doing, WAT is expected to be a 'thesaurus', i.e. to cover 'all' the vocabulary and reflect the basic structure of a particular language. As opposed to WAN, so far WATs are available for two languages only: English (by [5, Kiss et al]): 8400 stimuli – 54000 words – 1000 subjects, (by [6, Nelson et al]): 5000 stimuli – 75000 responses – 6000 subjects; and Russian (by [7, Karaulov et al]): about 8000 stimuli – 23000 words – 1000 subjects.

3 What Kind of Linguistic Information Could Be Extracted from WAT

It is usually questioned what FATs actually show? They do indicate that certain words are related in some way, but do not specify how. Although full of valuable information, the results of word association tests should be interpreted with great care [8].

The first who made an attempt of linguistic interpretation of word associations was Deese [9] who applied word associations to measure a semantic similarity of different words. His main assumption was that similar words must evoke similar responses. Thus, counting the stimulus word itself as a response by each subject, he computed the index of correlation between pairs of words as the intersection of the two distributions of responses and interpreted it as a measure of semantic similarity.

In the following subsections we demonstrate how WATs could help to solve the problems of the wordnet coverage and its appropriate structuring.

3.1 The Core Concepts of the Language

Experiments [10] show that in every language there is a limited number of words those appear as responses in WAT more frequently than other words. Such a set of words has much in

common with frequency lists (according to corpora-driven data) – they are among the most frequently used ones, and sets of top concepts (according to existing ontologies) – they have above-average number of relations to other words. This set is quite stable:

- it does not change much with time;
- it does not depends on the starting circumstances, e.g. on words that were chosen as the starting set of stimuli, or the number of subjects.

E.g., the Russian WAT [7] contains 295 words with more then 100 relations, among them are человек ('man'), дом ('house'), любовь ('love'), жизнь ('life'), есть ('be/eat'), думать ('think'), жить ('live'), идти ('go'), большой ('big/large'), хорошо ('good'), плохо ('bad'), нет (не) ('no/not') ..., while Edinburgh WAT [5] includes 586 such words: *man, sex, no (not), love, house; work, eat, think, go, live; good, old, small...*

These words determine the fundamental concepts of a particular language, and thus should be incorporated into lexical database as its core components (e.g., EWN Base Concepts [11]). Representing the most general concepts, these words are associated to most other (more specific) words by means of hyponymy relations. Extracting this set of basic concepts we are to tackle the problem of wordnet structuring.

3.2 Syntagmatic Relations

According to the law of contiguity, through life we learn "what goes together" and reproduce it together. Therefore, if a stimulus word is a verb, responses are expected to be all its co-occurring words: its right and left micro-contexts; nouns, adjectives and adverbs that could function in a sentence as its arguments.

This data could be incorporated into a wordnet both as surface context patterns for words (e.g. selectional restrictions/preferences, valency frames for verbs, etc.), and as deep semantic relations between words (e.g. ROLE/INVOLVED relations). Moreover, each pattern may be accompanied by the probabilistic index reflecting frequency of its occurrence in WAT (and, as a hypothesis, its probability in texts).

Also this data is useful for performing other tasks of wordnet constructing. It provides an empirical basis for distinguishing different senses of a word, establishing relations of synonymy, hyponymy, and antonymy.

3.3 Paradigmatic Relations

The law of contiguity may also explain the co-occurrence of paradigmatically related words in WAT. As synonyms, hyponyms/hyperonyms, meronyms/holonyms, or antonyms regularly go together in macro-contexts, they often appear together as pairs 'stimulus – response' in WAT.

Explicitly presented paradigmatic relations are a distinctive feature of WAT that differs it from other language resources (there is no such explicit information in explanatory dictionaries, and to extract it from corpora one needs to apply some sophisticated techniques).

This information may be included directly in terms of semantic relations between wordnet entries; also it helps us to enrich and to check out the set of relations encoded earlier.

3.4 Domain Information

Apart from the data on conventional set of semantic relations such as synonymy, hyponymy, meronymy etc., WAT provides more subtle information concerning domain structuring of knowledge. E.g., *hospital* → nurse, doctor, pain, ill, injury, load... This type of data is not so easy to extract from corpora, in explanatory dictionaries it is presented partly (generally covers special terminology only) and mostly based on the lexicographers' intuitions. E.g., *Syringe* – (medicine) *a tube with a nozzle and piston or bulb for sucking in and ejecting liquid in a thin stream*[1]. As opposed to conventional language resources (LRs), WAT explicitly presents the way common words are grouped together according to the fragments of reality they describe.

Domain relations may be attributed to each word in a wordnet; that give us broader (in comparison with context patterns, see 'Syntagmatic relations') knowledge of the possible contexts for each wordnet entry. The necessity of such an expansion becomes obvious if we take into account that domain information becomes crucial while we approach wordnet usage in IR systems.

3.5 Relevance of Word Senses for Native Speakers

The fact is that about 80% of associations of a word in WAT [12], as well as 90% of occurrences of a word in a corpus [13], are related to 1–3 of its senses. That allows us to measure the relevance of a particular word sense for native speakers, and, hence, to find an appropriate place for it in the hierarchy of senses. E.g., if we consider the word *lap* and its associations, we could find that 3 senses (lap_1 – 'the flat area between the waist and the knee of a seated person', lap_2 – 'one circuit of a track or racetrack' and lap_3 – 'take up with the tongue in order to drink') account for 61% of its word associations (cf. lap_1 → *knee, sit, sit on*, etc. lap_2 → *circuit, race, run*, etc. lap_3 → *cat, milk, pap* etc.). Those could be regarded as the most important from the viewpoint of native speakers. Other senses, such as 'polish (a gem, or metal or a glass surface)' obviously constitute the periphery (∼2%). And there is no hint of the sense 'a part of an item of clothing' while it is presented in the explanatory dictionaries (cf. [13]).

These empirical evidences also help us to define the necessary level of sense granularity: to include into the wordnet no more and no less senses of each word than native speakers do differentiate. Thus, the problem of unnecessarily over-multiplying of sense entries (usually mentioned regarding PWN 1.5.) could be avoided.

3.6 Relevance of Relations for Native Speakers

It is clear that in a WN words must have at least a hyperonym and desirably a synonym. But what concerns relations other than Hyponymy and Synonymy, how could we ensure that we include all the necessary relations, and that what we include is necessary? Relations are not the same for different PoS, but also they are not the same for different words within the same PoS. E.g., according to [5] for English native speakers the most relevant relation of *buy* is that to its conversive *sell,* while for *cry* the most important relation would be INVOLVED_AGENT *baby*.

[1] This definition as well as the ones below was taken from New Oxford Dictionary of English. Oxford University Press (1998).

3.7 Semantic Classification of Words Obtained by Using Formal Criteria Only

Within the same PoS the proportion of syntagmatic and paradigmatic associations varies considerably. E.g. for Russian verbs the number of syntagmatic associations can vary from 35% to 90%. This ratio correlates with syntagmatic features of verbs, such as a number of valencies, strength of valencies, and their character (obligatory/optional), which in turn correlate with semantic features of the verb. This hypothesis is proved while building semantic classifications of verbs on the basis of formal criteria (e.g. the number of syntagmatic associations). The resulted classes turned to have much in common with semantic classes acquired by means of logic or componential analyses (cf. [14,15]).

This data supply us with empirical basis for appropriate structuring of lexical database: grouping the words into semantic classes, etc.

4 WAT vs. Corpus

It is unanimously recognized that to build an adequate and reliable lexical database (e.g. wordnet), reflecting all the potentialities of a language, it is not enough to rely upon information produced by 'experts' (i.e. linguists, lexicographers) and stored in conventional LRs, whatever advantages for machine usage they offer [16]. One should rather explore the raw data, and extract information from language in its actual (i.e. written and spoken texts), and its potential use (i.e. native speakers' knowledge of language), that could be examine by means of psycholinguistic techniques.

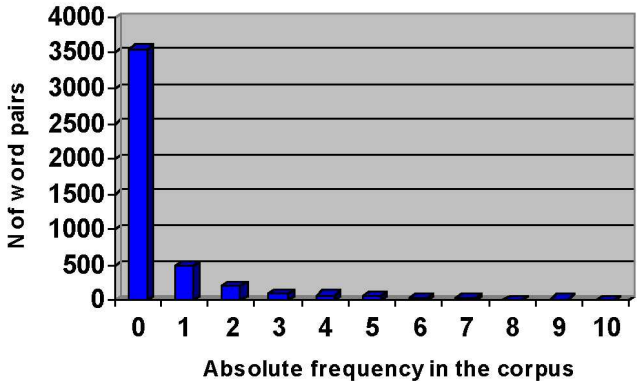

Fig. 1. Overlap between RWAT and the corpus.

Several researchers [17,18,19] performed statistical analysis and comparison of such 'raw' LRs, namely, text corpora and word associations, in order to confirm the correlation between frequency of XY co-occurrence in a corpus and the strength of association X-Y in WAN. Those experiments successfully demonstrated that corpora could be used to obtain the same measures of association strength as WAN, at least for the most frequent words. In our research we made a comparison in the opposite direction, and were to show that a

WAT covers more language phenomena than a corpus. For that purpose the Russian WAT [7] and a balanced text corpus of about 16 mln words were used. 6000 'stimulus-response' pairs e.g. бояться – темноты ('be afraid of – darkness') were extracted from RWAT in random order, and then searched in the corpus. The window span was fixed to $-10;+10$ words.

The most interesting result of our experiment was that about 64% word pairs obtained from subjects do not occur in the corpus (see the first column on Figure 1).

By excluding all unique associations (that with absolute frequency = 1) from the query list, the proportion of absent pairs may be reduced to 42%, which is still higher than expected. The distribution of the non-unique associations that were not found in the corpus could be seen in Table 1.

Table 1. Distribution of word associations that do not occur in the corpus.

N of occurrences in the corpus	N of occurrences in RWAT	% of all word pairs missed
0	2	48
0	3	22
0	4	14
0	5	8
0	6–10	5
0	11–15	<1
0	15–20	<1
0	>20	0

Looking for explanation we assumed that paradigmatically related words frequently appear as 'stimulus-respond' and less frequently co-occur in texts. But more detailed observation of the word pairs chosen revealed unexpectedly high ratio of syntagmatic word pairs to be absent. For verbs this number was about 84% of total amount of absent pairs. Whereas paradigmatically related words were regularly presented in the corpus.

Thus, we are to conclude that the experiment performed proves the value of WAT as a LR, which could supply the researcher with data otherwise inaccessible.

5 Conclusion

The advantages of using WAT in wordnet constructing may be stated as follows:

1. **Simplicity** of data acquisition.
2. Great **variety** of semantic information extracted.
   As it was shown in Sections 3 and 4, WAT is equal to or excels other LRs in several respects.
3. **Empirical nature** of data extracted (as opposed to theoretical one, cf. conventional dictionaries, that supposes the researcher's introspection and intuition to be involved, and hence, leads to over- and under-estimation of the language phenomena).
   As it was shown in Section 4, WAT may function as a source of 'raw' linguistic data, comparable to a balanced text corpus, and could supply all the necessary empirical information in case of absence of the latter.

4. **Probabilistic** nature of data presented (data reflects the relative rather then absolute relevance of language phenomena).

To sum up we may add, that the parallel usage of WAT and other LR is an efficient way of conducting constant checking-out of wordnet construction, its refining and expanding. Thus, we believe the high consistency and coverage of wordnets could be achieved.

References

1. Kent, G. H., Rosanoff, A. J.: A Study of Association in Insanity. American Journal of Insanity, 67 (1910) 37–96.
2. Kurcz, I.: Polskie normy powszechnosci skojarzen swobodnych na 100 slow z listy Kent-Rosanoffa. Studia psychologiczne, tom 8. Warszawa (1967).
3. Novák, Z.: Volné slovní párové asociace v češtině. Praha (1988).
4. Rosenzweig, M. R.: Etudes sur l'association des mots. Année psychol. (1957).
5. Kiss, G. R., Armstrong, G., Milroy, R.: The Associative Thesaurus of English. Edinburgh (1972).
6. Nelson, D. L., McEvoy, C. L., Schreiber, T. A.: The University of South Florida word association, rhyme, and word fragment norms (1998) `http://www.usf.edu/FreeAssociation/`.
7. Karaulov, Ju. N. et al.: Russian Associative Thesaurus. Moscow (1994, 1996, 1998).
8. Clark, H. H.: Word associations and linguistic theory. In: J. Lyons (ed.). New horizons in linguistics. Harmondsworth: Penguin (1970) 271–286.
9. Deese, J.: The Structure of Associations in Language and Thought. Baltimore (1965).
10. Ufimtseva, N. V.: The core of the Russian mental lexicon (on the basis of large-scaled association tests). In: Proceeding of the Conference on Corpus Linguistics and Linguistic Databases. St-Petersburg, (2002) (in Russian).
11. Vossen, P. (ed.): EuroWordNet: A Multilingual Database with Lexical Semantic Network. Dodrecht, Kluwer (1998).
12. Ovchinnikova, I. G., Shtern, A. S.: Associative strength of the Russian words. In: Psycholinguistic problems of phonetics and semantics. Kalinin (1989) (in Russian).
13. Hanks, P.: Immediate context analysis: distinguishing meanings by studying usage. In: Heffer, Ch., Sauntson, H. (eds.) Words in Context. CD. Birmingham (2000).
14. Sinopalnikova, A. A. Classifying Russian Verbs according to their syntagmatic word associations. Diploma thesis, Saint-Petersburg State University (2000) (in Russian).
15. Ushakova, A. A. Classifying Russian Verbs: Componential and Definition analysis. Diploma thesis, Saint-Petersburg State University (2000) (in Russian).
16. Calzolari, N.: Lexicons and Corpora: between Theory and Practice. In: Proceedings of the 8þ International Symposium on Social Communication. Santiago de Cuba, (2003). 461–469.
17. Church, K. W., Hanks, P.: Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics 16(1) (1990) 22–29.
18. Wettler, M., Rapp R.: Computation of Word Associations Based on the Co-Occurrences of Words in Large Corpora. In Proceedings of the 1st Workshop on Very Large Corpora: Academic and Industrial Perspectives. Columbus, Ohio (1993) 84–93.
19. Willners, C.: Antonyms in context: A corpus-based semantic analysis of Swedish descriptive adjectives. PhD thesis. Lund University Press (2001).