Word Association Norms as a Unique Supplement of Traditional Language Resources

Anna Sinopalnikova[1, 2], **Pavel Smrz**[1]

[1]Faculty of Informatics, Masaryk University
Botanicka 68a, 602 00 Brno, Czech Republic
[2]Saint-Petersburg State University
Universitetskaya 11, Saint-Petersburg, Russia
{anna, smrz}@fi.muni.cz

Abstract

The paper deals with reuse and testing of psycholinguistic resources in linguistic studies. We report on the experimental work aiming at comparing and evaluation of linguistic information provided by Word Association Norms (WAN) with that derived from other language resources (LRs), namely corpora and wordnets. Results for 3 European languages: Russian, English and Czech are presented. Quantitative and qualitative outputs of the research show that WAN can be employed to enrich the traditional LRs and serve as a tool for their consistency checking.

Introduction

It is generally accepted that most today's NLP applications need quality language resources (LRs), especially those based on real data. The well-established methods gain from large corpora, various dictionaries, lexical semantic networks, databases etc. This paper deals with another kind of LR – Word Association Norms (WAN) that has been neglected, or at least not regularly used in linguistics studies till now. Results of our research clearly show that WAN can be employed to enrich the traditional LRs and serve as a tool for their consistency checking.

Basic Notions

Word Association Norms (**WAN**) are a collection of empirical data obtained through large-scaled psycholinguistic experiments known as free association tests. The standard technique of the experiments is as follows: words (**stimuli**) are presented to subjects, who are asked to respond with the first word that comes into their mind (**responses**). The list of stimuli and lists of responses ranged according to their frequency in the answers constitute the body of WAN. For example, the distribution of responses for the stimulus *needle* is as follows:

> *Thread: 41, pin: 13, sharp: 6, sew: 5, cotton: 2, dressmaker: 1, fix: 1, prick: 1, sewing: 1, sow: 1, spring; 1, stitch: 1,* etc.

The first WAN were collected by Kent and Rosanoff (1910) 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; Minnesota WAN by Palermo and Jenkins (1964) still being the most famous and influential one.

In its most sophisticated form WAN are expanded to associative network for several thousands words. The cycle of data collection is repeated several times: a small set of stimuli is used as a starting point of experiment, responses obtained for them are used as stimuli in the next run, and so on. The complicated procedure of data collection is applied to assure WAN to become a 'thesaurus', i.e. to cover all the vocabulary and map the basic structure of a particular language. So far, large WAN, the so-called **Word Association Thesauri (WAT)** are available for two languages only: English (Kiss et al, 1972) and (Nelson et al, 1992), and Russian (Karaulov et al, 1994-1998). For other languages only small WAN including 100-200 stimuli are available.

Motivation

Primarily designed as a psychological tool for revealing and measuring insanity, nowadays WAN are mostly used in sociological, particularly gender, or ethnolinguistic studies (to measure the differences between ages, sexes, cultures or nations). Their role as a source of linguistic information is generally neglected. However, the first attempt to make a linguistic interpretation of WAN data was made by Deese in 1965. He applied WAN to measure a semantic similarity of different words, using as a base his assumption that similar words must evoke similar responses.

Thus, our **first** reason was to avail of and extend the existing tradition of using psycholinguistic resources in linguistic studies. WAN are available for many languages, and their electronic form make them applicable to regular tasks of computational linguistics, for example, to wordnet construction.

The **second** argument was inspired by the general view on corpora as the only source of raw data that allow them to take a unique position of the "sacred cow" of modern linguistics. We are convinced that in many respects WAN

present an alternative to corpora, and thus, could be used for testing corpora, and checking their consistency, coverage and representativeness.

Research procedure

Two sets of experiments have been performed to compare WAN to other types of LRs, namely corpora and lexical databases. The goal was to test how WAN could help to enrich the available LRs and to check their consistency. The experiments were based on LRs, which differ in volume and coverage. That allowed measuring the dependence of the LR quality on its size, and also gave an impulse to further speculation on the principal differences between languages.

In all experiments described in the section below the following WAN have been used:

- RAT - Russian WAT by Karaulov et al (1994-1998): 8000 stimuli - 23000 words covered – 1000 subjects,
- EAT - Edinburgh WAT by Kiss et al (1972): 8400 stimuli – 54000 words covered - 1000 subjects,
- Czech WAN (Novak et al, 1996): 150 stimuli - 4000 words covered – 250 subjects.

WAN vs. Corpus

The first series of experiments aimed at comparison of WAN to corpora. Although several researchers have already proved that corpora are comparable to WAN in that they provide the same measures of association strength between words (Church & Hanks, 1990; Wettler & Rapp, 1993; Willners, 2001), we made a comparison in the opposite direction, and were to show that WAN cover more language phenomena than a corpus.

For that purpose

- Bokrjonok 3.0. - balanced corpus for Russian of about 16 mln words,
- BNC - British National Corpus (112 mln),
- CNC - Czech National Corpus (160 mln) and its unbalanced version (630 mln words)

have been used.

5000 word associations, such as e.g. *mouse – cheese, dark - alley* have been extracted from each WAN in random order, and then searched in the corpora. The window span was fixed to -10; +10 words. A word association X-Y observed in WAN was treated to be equal to a co-occurence of words X and Y in a text, whether in the immediate context or expanded one.

WAN vs. Corpus: Russian

The most interesting result of the experiment was that about 64% word associations obtained from Russian subjects in experiment do not occur in the corpus. By excluding all unique associations (that with absolute frequency = 1) from the query list, the proportion of absent pairs could be reduced to 49%, which was still higher than expected. Looking for the explanation we assumed that paradigmatically related words (*e.g. hate – feel*) appear more frequently as 'stimulus-respond' in WAN and less frequently co-occur in texts. But more detailed observation of the given word associations revealed unexpectedly high ratio of syntagmatic associations to be absent. For verbs this number was up to 84% of total amount of absent pairs. On the other hand, paradigmatically related word associations were usually presented in the corpus. The qualitative analysis of the non-unique associations, which were absent in the corpus, demonstrated the following distribution over semantic relations (see Table 1).

Relation	% of associations missing
PARADIGMATIC:	**21,4**
antonymy	1,5
cause	1,6
co-hyponymy	4,9
has_subevent	0,8
hyponymy	2,5
is_subevent	2,9
meronymy	0,5
synonymy	2,9
xpos_near_synonymy	3,6
others	0,2
SYNTAGMATIC:	**48,7**
Adj+N	7,9
N+Adj	4,8
V+Adv	9,1
V+N (agent)	3,5
V+N (instrument)	1,4
V+N (location)	1,5
V+N (object)	8,3
V+N (patient)	9,8
V+V	1,1
others	1,3
DOMAIN	**13,4**
OTHER	**16,5**

Table 1: Distribution of word associations across semantic relations: Bokrjonok.

As even syntagmatic relations, which were expected to be present first and foremost, were not extracted from the corpus, the main conclusion drawn is that the corpus of 16 mln words is still not enough to cover the core of the Russian vocabulary and present its structure.

WAN vs. Corpus: English

While applying much larger corpus in case of EAT and BNC, we have found that, firstly, the total number of missing word associations was much smaller (31 %) and, secondly, the proportion of absent syntagmatic and

paradigmatic associations was very different (see Table 2).

Relation	% of associations missing	
	Eng	Cze
PARADIGMATIC	57,1	61,2
SYNTAGMATIC	8,4	10,9
DOMAIN	21,7	12,1
OTHER	12,8	15,8

Table 2: Distribution of word associations across semantic relations: BNC and CNC.

The obtained results are in agreement with the general view about critical role of the corpus size on its coverage. Together with the data for Russian, this information allows measuring the extent of corpus coverage.

The detailed observation of the data missing in the BNC gave us evidences for the following statements:

- As for the **paradigmatic** relations acquisition, even a large corpus could not compare with WAN. This particularly holds for such relations as synonymy and hyponymy, when the difference in register, style, or genre prevents co-occurrence of neutral words with 'coloured' ones, e.g. *sex – fornicate* (archaic or humorous), *ire* (poetic) – *anger, cowardly – yellow* (slang). Moreover, in this particular experiment it turned to be valid also for some pairs when both synonyms/words were neutral terms e.g. *astonish – surprise, inanimate – dead, malady – illness.*

We are aware that our simple methods are hardly enough to extract such information from a corpus. Thus, more sophisticated technique of lexicosyntactic patterns proposed by Hearst (1998) were tested on the given material, yet with little success. Employment of complex methods of statistical analysis, such as (Lin, 1998; Kilgariff et al, 2004) could probably succeed, but they would require amounts of text even larger then BNC could now provide.

- WAN are indispensable source of information about **low frequent** words, otherwise inaccessible. Relations of such words as *perambulate* (N_{BNC} = 3), *fornicate* (N_{BNC} = 6) are presented in their range and variety in WAN: e.g. *perambulate - walk: 30, pram: 17, baby: 9, push: 8, about: 1, dawdle: 1,move: 1, promenade: 1, slowly: 1, stroll:1, through:1, wander:1*, etc.
- WAN turned to be useful for acquiring **Domain** relations; absent portion of them was surprisingly large for such corpus as BNC e.g. *ink-pot – pen: 24, non-violence – peace 29, offside – soccer 2.*

WAN vs. Corpus: Czech

The last case differs significantly from the previous two. We have at our disposal a large corpus and a small WAN, covering only 4000 words.

We have found that the total number of missing word associations was only 514 (10,28%), and the proportion of the syntagmatic and paradigmatic associations among them was similar to that for English. Thus, our aim was to prove that even though Czech WAN could not compare with CNC in coverage of relations between words, it is still useful what concerns those specific types listed in the previous section. The number of absent associations did not allow us to make any general statement, however, some important features of particular words were extracted: e.g. synonymy relation *polámaný – rozlámaný* ('damaged'), *osladit - pocukrovat* ('to sweeten') or subevent relation as in *šetřit* ('to economize') – *mamonit* ('be tightwad') obtained from associations. Furthermore, our conclusion about the less frequently used words could be supported by Czech examples: e.g. *pocukrovat* (N_{CNC} = 12), *mamonit* (N_{CNC} = 6).

WAN vs. Wordnet

The second series of experiments aimed at comparison of WAN to wordnet-like lexical databases.

For that purpose

- Princeton WordNet 2.0 (115 000 synsets),
- Czech Wordnet 1.8 (28 000 synsets),
- RussNet 0.2 - semantic network for Russian linking lexical semantics to derivational morphology (5500 synsets)

have been used. The same set of 5000 word associations has been searched in the wordnets. The following requirements were established: a word association X-Y observed in WAN was treated to be equal to a direct relation between words X and Y (both directions X→Y or Y→X), or inherited transitive relation $X \rightarrow x_1 \rightarrow \ldots \rightarrow x_n \rightarrow Y$ in the wordnet.

The qualitative results of the second series of experiments were predictable: almost all of the associations searched were not found in wordnets (91% for Russian, 74% for English, and 89% for Czech). The possible explanation concerns the difference in the very nature of these LRs: WAN being a primary resource provide the raw empirical data, while WN is a derived resource and present the interpretation of the data, thus deals mostly with types of relations, not instances.

Relation	% of associations missing		
	Eng	Cze	Rus
PARADIGMATIC	14,6	28,7	41,3
SYNTAGMATIC	30	33,7	26,7
DOMAIN	43,1	31,4	33,1
OTHER	12,3	6,2	9,9

Table 3: Distribution of word associations missing in wordnets.

Different size of WAN for each language forced us to make different conclusions in each case and interpret figures differently.

The detailed analysis of association distribution for Russian and Czech wordnets gave us evidences for their inconsistency wrt paradigmatic relation presentation. In case of RussNet the insufficient number of relations per word complemented it. The main role played the absence of synonymy, involved/role, and xpos_near_synonymy

links. Thus, we should admit that the research results clearly indicate further steps and necessary directions of RussNet expansion.

In case of Czech the small size of WAN and the average size of the WN enable us to come to conclusion about the insufficient coverage of the letter. Czech WAN contain only the most frequent words, but not all of them were found in the wordnet and not all their relations were presented.

The WordNet coverage was not a crucial factor for the English. But still some inconsistencies were found, and evidences of the necessity of further expansion due to the new types of relation were drawn.

We suppose that a direct mapping between WAN and a wordnet is impossible, and direct incorporation of the word associations into a wordnet is unreasonable in most cases. There should be a preliminary step of manual analysis and generalization of the data using the hyponymy hierarchy of concepts. E.g. all associations of the same type e.g. *drink – water, beer, milk, ale, Coca-cola, coffee, juice,* etc. found in WAN should be generalized as *drink* ROLE_OBJECT *beverage* relation.

Yet several types of WAN-driven data are open for direct implementation to wordnet:

- As we have already mentioned, the high ratio of absent associations was partly due to certain **semantic relations missing** in Princeton WN. It seems to be reasonable to introduce in PWN such relations as e.g. *seek – find: 56,* (CAUSE), *moo – cow: 70, neigh – horse: 57* (INVOLVED_AGENT), *pale – pallor* (XPOS_NEAR_SYNONYMY).
- In many cases WordNet **Domains** do not involve relations between very common and frequently used words, e.g. *needle – thread: 41*. Often, when domain relations could be inferred from the hyponymy hierarchy, explicit domain link may be unnecessary (e.g. kinship terms). But *needle* and *thread* are linked through *artifact, artefact* concept only, and the distance between them according to PWN is 6 steps.

Among the most frequent types of domain relations, that were extracted from WAN and were not found in WN we should mention:

- name of domain (situation) – domain member e.g. *hospital – nurse:8, finance – money: 61, chiropody – feet:57;*
- participant – participant e.g. *pepper – salt: 58, tamer – lion: 69, mouse – cat: 22; needle – thread: 41*
- participant – circumstance e.g. *umbrella – rain: 58;*
- participant – pointer to its action/function/role e.g. *larder – food: 58, envelope – letter: 60,* etc.

Our findings allow us to make some interesting conclusions with respect to wordnet consistency-checking:

- **Possible wrong location** in the hierarchy: e.g. *kitten – cat: 54, pup – dog: 63. Kitten, pup* are treated in PWN as hyponyms of *young mammal,* and have no direct hyperonymy link to *cat, dog,* although the discovered information is included into their definitions.

Conclusions

We are to conclude that the performed experiments show that in several respects WAN are equal to or excels other LRs. Also we proved that WAN is a kind of LR, which supply the researcher with data otherwise inaccessible (that concerns, for example, automatic acquisition of paradigmatic or domain relations). WAN may function as a source of 'raw' linguistic data, comparable to a large text corpus, and could supply all the necessary empirical information in case of absence of the latter.

As for the consistency checking, we may add that the parallel usage of WAN and other LR is an efficient way of conducting regular checking of wordnet construction, its refining and expanding. We believe that suggested expansions of wordnets by means of WAN-driven data could make them more useful and applicable to the current NLP tasks.

References

Church, K. W., Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics 16 (1). MIT Press. (pp.22-29).

Deese, J. (1965). The Structure of Associations in Language and Thought. Baltimore.

Hearst, M. A. (1998). Automatic Discovery of WordNet Relations. In: Fellbaum, C. (ed.) WordNet: an Electronic Lexical Database. MIT Press, Cambridge, MA.

Karaulov, Ju.N., Cherkasova, G. A., Ufimtseva, N.V., Sorokin, Ju. A., Tarasov, E.F. (1994, 1996, 1998) Russian Associative Thesaurus. Moscow.

Kent, G.H., Rosanoff, A.J. (1910). A Study of Association in Insanity. American Journal of Insanity, 67 (pp. 37-96).

Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D. (2004). The Sketch Engine. In: Proceeding of EURALEX' 2004 (to be published).

Kiss, G.R., Armstrong, G., Milroy, R. (1972). The Associative Thesaurus of English. Edinburgh.

Lin, D. (1998). Automatic Retrieval and Clustering of Similar Words. In: Proceedings of the COLING/ACL'98. Montreal, Canada.

Nelson, D.L., McEvoy, C.L., Schreiber, T.A. (1998). The University of South Florida Word Association, Rhyme, and Word fragment norms. http://www.usf.edu/FreeAssociation/.

Novák, Z. (1988). Volné slovní párové asociace v češtině. Praha

Palermo, D., Jenkins, J. (1964). Word Association Norms. University of Minnesota Press, Minneapolis.

Wettler, M., Rapp R. (1993) . 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 (pp. 84-93).

Willners, C. (2001). Antonyms in context: A corpus-based semantic analysis of Swedish descriptive adjectives. PhD thesis: Lund University Press.