Semantic Relations in Cognitive eLexicography

Gerhard Kremer and Andrea Abel
EURAC Research, Bolzano

Whereas dictionary design has traditionally been guided by the results of dictionary use research, recent approaches in lexicographic research are strictly user-centred. We support the idea of integrating empirical cognitive evidence into this type of research, thus fruitfully exploiting it for both, the selection (and subsequently presentation) of lexical data and the acquisition of such data from corpora. Focusing on the extraction of semantic relations to be illustrated in electronic learners' dictionaries, we analyse the results of two behavioural experiments on the production as well as the perception of semantic relations. The main goal of the experiments was to determine which relations are cognitively salient in speakers' minds. With the objective of developing a method to automatically extract cognitively salient semantic relations from corpora, we describe and discuss findings of the first analyses conducted on composite part relations. In future this might serve as a basis for the elaboration of new strategies aimed at enriching lexical databases and dictionaries.

1. Introduction

The design and conception of new dictionaries as well as new editions of existing dictionaries has traditionally been guided by the results of dictionary use research (see, e.g., Wiegand 1987, 1998), which focuses on the analysis of dictionary use situations (e.g., by means of questionnaires, tests, think-aloud protocols, or log-files, etc.). Furthermore, more recent approaches in lexicographic research are strictly user-centered and based on a detailed description of the users' problems and needs in different communication-oriented situations (see, e.g., Bergenholtz/Tarp 2002, 2003; Abel 2003). Evidence concerning users' problems and needs can be gained from needs analyses, from research on first and second language acquisition, from learner error analyses, etc.

In a user-centered approach, empirical cognitive evidence should be taken into consideration, too, as it can play an important role for both, (a) the selection (and in a further step the presentation) of lexical data in a cognitively motivated way and (b) the acquisition of these data from corpora.

Accordingly, in this paper we focus on the very specific field of pedagogical as well as electronic lexicography, in the following called eLexicography[1], and concentrate on one particular aspect, namely semantic relations, i.e. the field of paradigmatic relations or relations of concepts (seen from a linguistic-structuralist and a cognitive-oriented perspective, respectively). The ultimate goal of our research is the development of new strategies for the enrichment of electronic learners dictionaries, such as – in the present case – the ELDIT dictionary (Electronic Learners' Dictionary German – Italian; see, e.g., Abel/Weber 2000, Knapp 2004), on the basis of cognitively motivated decisions.

Semantic relations are relevant for pedagogical lexicography from a variety of perspectives. An extensive presentation of semantic relations within a word entry is important in productive L2-tasks, when a language learner has to choose the semantically correct word out of many possibilities. Furthermore, as the co-occurrence of semantically related items within a text section is probable, in writing tasks the learner should have easy access to those items. More generally, the presentation of semantic relations offers useful access paths to lexical

[1] Corresponding to the title of the conference "eLexicography in the 21st century: new challenges, new applications", held in Louvain-la-Neuve (Belgium) from 22 to 24 October 2009 (http://www.uclouvain.be/en-cecl-elexicography.html)

data in addition to access paths based on merely formal criteria. Semantic relations are important with respect to meaning description, too, as the meaning of a word is composed of a number of features such as the reference to an entity in the world, connotational aspects, etc., but also by sense relations to other words. Conveying the meaning of words is one of the core functions of dictionaries, and multifaceted meaning descriptions can improve the usability of a dictionary. Lastly, the organization of the native speaker's mental map for each word or concept is characterised by the existence of a complex network of multidimensional information and relations, including semantic ones. Thus, a dictionary should include such relation networks (which exist in the learner's mental lexicon only fragmentarily), as they are important for language learning and processing (see, e.g., Casiddu 1996a/b, Jackson 2003, Abel/Campogianni 2005). Before dealing with aspects of adequate data presentation, the tasks regarding data selection and acquisition have to be tackled, which is accordingly the central issue of this paper.

2. Related Work

Semantic relations are often available in lexical resources. Different approaches are used for analysing and selecting candidates to be included. This section discusses some of them.

The ELDIT dictionary offers the possibility to explore the semantic neighbourhood of a word meaning through 'word fields', i.e., a set of closely related words, such as hyponyms, cohyponyms, (quasi-)synonyms, etc., which are presented as interactive graphics in the user interface (see, e.g., Abel/Campogianni/Reichert 2004). The relations that define 'word fields' in ELDIT have until now been chosen on didactic and theoretical lexico-semantic grounds (among others, structural semantics and word field theory - see, e.g., Geckeler 2002 and Hoberg 1970) rather than being based on experimental data determining which relations are more salient for native speakers. However, 'word field' input in ELDIT has been manually carried out by lexicographers by using data sources such as online lexical resources (e.g., the 'Wortschatz-Portal' – Universität Leipzig, products and projects related to WordNet, such as the Visual Thesaurus and the Italian WordNet) and synonym dictionaries (e.g., 'Duden. Die sinn- und sachverwandten Wörter' 1997, 'Dizionario sinonimi e contrari' 1999), resulting in a rather small set of entries (currently, a few hundred).

WordNet is an electronic lexical database, where synonymous words are combined into semantically related synsets which are linked to each other. The choice for the types of relations used is not based on studies showing their cognitive salience. Furthermore, WordNet comprises only taxonomy-related semantic relations, and the same set of relations is used for all words with the same part of speech (e.g., for nouns: synonymy, antonymy, hypernymy, hyponymy, and meronymy). Finding semantically related synsets has been done manually in WordNet (cf. Fellbaum 1998).

Other lexical resources including semantic relations exist (such as 'KirrKirr' – see, e.g., Manning/Jansz/Indurkhya 2001, 'Alexia' – see, e.g., Chanier/Selva 1998, the 'Longman Language Activator', etc.), but their entries were either linked to each other manually, or the method has not been made transparent.

Methods for the automatic extraction of semantic relations from corpora have been developed, many based on one of the first approaches proposed by Hearst (1992): hypernym-hyponym relations were collected using a simple lexico-syntactic pattern ('[noun], such as [noun],...'). In a similar approach, Almuhareb and Poesio (2004) used pure word-based

patterns, whereas Pantel and Pennacchiotti (2006) used seed instances to discover more lexico-syntactic patterns for a specific relation type.

To our knowledge, no lexical resource exists where the choice for the included semantic relation types is based on an empirical cognitive study, and for which relations were extensively extracted via an automatic method from text corpora.

3. Experiments

To investigate which semantic relations are cognitively salient to native speakers, and to eventually extract appropriate relation candidates for dictionary entries, two behavioural experiments were conducted. Both production and perception of semantic relations were tested. In the experiments the two target languages German and Italian were used in order to detect differences and similarities. To facilitate a systematic method of semantic relation acquisition in the next phase, differences between *classes* of concepts were analysed, considering *types* of semantic relations.

In the production experiment (following the line of experiments for the acquisition of semantic norms, e.g., as conducted by McRae/Cree/Seidenberg/McNorgan 2005), participants were handed out sheets with words representing concepts which were chosen from 10 concept classes (mammals, birds, fruit, vegetables, body parts, clothing, implements, vehicles, furniture, buildings). The task was to write down short descriptive phrases for each concept, given the time limit of 1 minute. The German and Italian phrases produced were annotated with semantic relation types which were taken mainly from the set of those used by Wu/Barsalou (2004). For example, when a participant described the concept ‘dog’ with ‘has four legs’, this phrase was annotated with the semantic relation type ‘ece’ (i.e., external component of an entity, or part).

For the analysis the number of relations produced for each semantic relation type was counted. Overall distributions were found to be similar between the two target languages. Next, distributions of semantic relation types within each concept class were analysed. Figure 1 illustrates the deviations from the overall distributions. The German and Italian data have very similar deviation patterns of over-/underrepresentation of semantic relation types, showing again their language-independent distribution. This suggests a language-independent acquisition of semantic relations. Similar distribution patterns for the three broad classes of animals, plants, and man-made objects (plus body parts) are evident. This supports the idea that for different concept classes different sets of semantic relation types are cognitively more salient. An unsupervised cluster analysis (given the type counts for each concept) further supported this idea. Details of the production experiment and the analysis are described in Kremer/Abel/Baroni (2008).

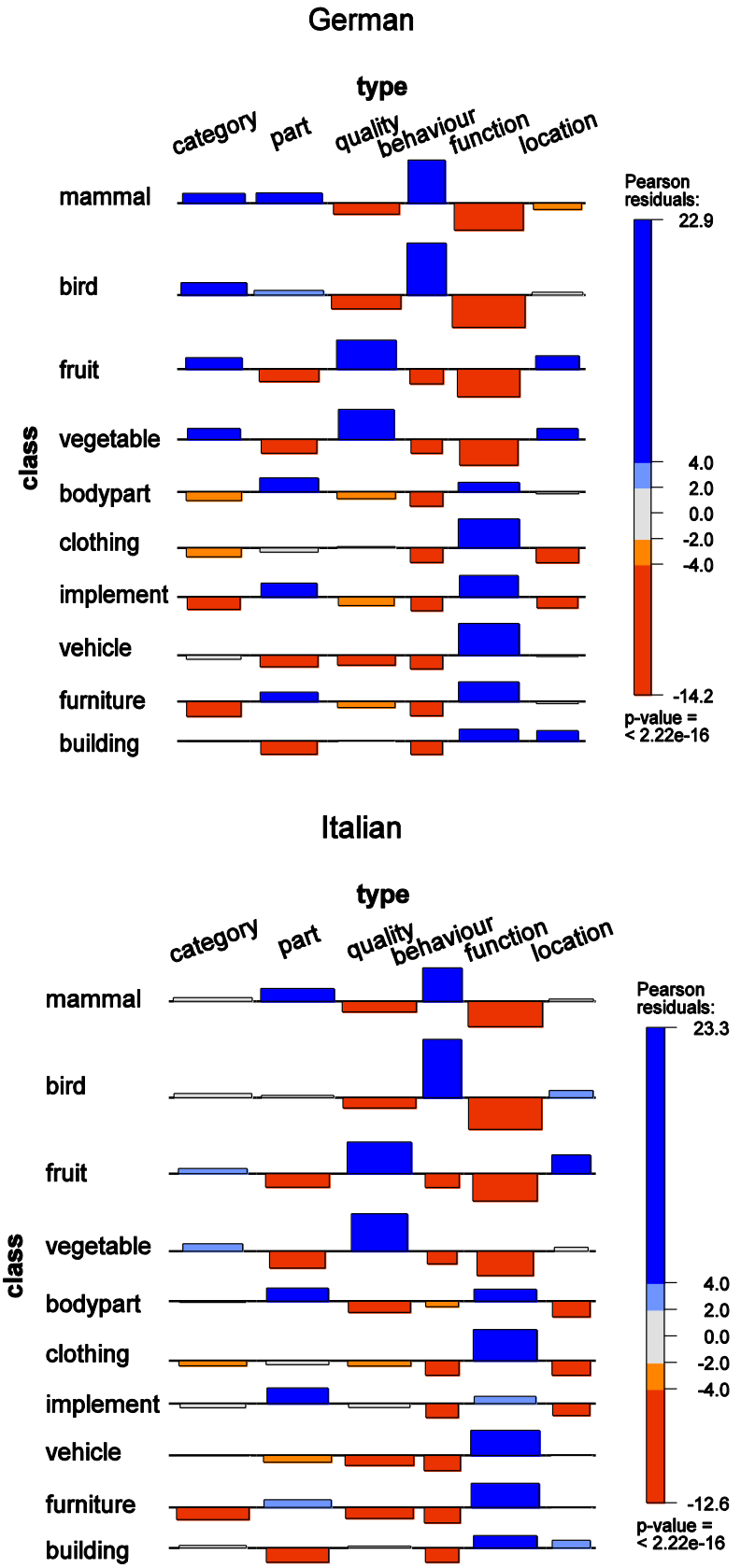

Figure 1. Deviations from overall relation type distributions. Positive/negative deviations are depicted by blue/red vertical bars above/below the baseline. Degree of significance of deviation (according to a Pearson residual test) is represented by saturation of colour.

The results indicate which semantic relation types are cognitively more salient when participants produce semantic relations, but it could be different from what they perceive. Thus, a follow-up experiment was conducted, during which the participant's task was to decide if a concept word and a second word were a valid pair in the sense that the second word could be used for describing the concept word. Valid word pairs were taken from the production experiment data. Words for non-valid word pairs were selected manually from those words in a large German WaCky corpus (a web text corpus developed within the WaCky initiative) that appeared in the context of the concept words and had high association measures. This was done to avoid the possibly confounding factor of non-valid word pairs with low association measures. To provide the Italian participants with the same word meanings, the German stimuli were translated into Italian (being aware that these might have different association measures). Concept word and paired word were presented subsequently on a monitor with a maximum response time limit of 2 seconds. Response times and responses were recorded.

The response time data and the response errors were analysed using a mixed effects model. As influencing factors of the response time, concept class, relation type, and their interaction were given to the model. The factors chosen to have possible random effects were subject, concept word, and length of second word (including the factor language did not have significant differences). Neither the statistical nor the visual analysis of the reaction time distributions and response errors confirmed or contradicted the results from the production experiment: Significantly low and high reaction times (and error rates) did not consistently match with over- and underrepresented relation types. Nevertheless, the existing significant differences in reaction time distributions indicate differing degrees of cognitive salience, which supports in general the results from the production experiment and the underlying idea for this project.

4. Towards the Acquisition of Semantic Relations from Corpora

Applying the experiment results, the idea of how to populate a lexical database is to determine a concept’s class membership and subsequently choose the cognitively salient semantic relation types depending on that class. Now, the next goal is to automatically extract semantically related words for each of those relation types from text corpora. For this purpose, choosing a representative text corpus is critical, as it ideally should include enough information about the target concepts. This first study focuses on the German language, investigating the occurrence of the production data in the large WaCky corpus. A difficulty is finding for a given concept word those semantically related words which belong to the target relation type.

As a first target relation type to extract, part relations are focussed on, in particular including those composed of a noun and preceding modifiers (e.g., ‘big ears’), which, to our knowledge, has not been addressed in other projects, yet. For developing an automatic acquisition method, the concept-part pairs as they were produced in the production experiment are examined and compared with their occurrences in the corpus: All contexts were extracted that included the pairs of concept word and part relation (looking only for the noun in composite relations) produced in a 20 sentence window. Regarding the composed part relations, all possible modifiers of the noun were collected that appeared within a 4 word window preceding the noun.

Only about 6% of the 376 concept-part pairs (leaving out the modifiers) were not found in the corpus (including dialect specific words); the others occurred at widespread frequency numbers (from 64 to 86201).

From the 336 pairs of concepts and composite part relations, about 38% were not found in the corpus[2]. Looking at which other modifiers occurred with the pairs of concepts and part relations, a list of frequency ranked modifiers for the pairs of concepts and relations produced with a modifier was generated. Examining this list, a majority of modifiers would not be desirable for extraction, but a few were synonymous to those produced in the experiment or at least they belonged to the same attribute type (colour, size, surface property, etc.).

To see how many relations were preferably produced with or without modifier, Figure 2 is provided. It shows for each concept class the absolute counts of those part relations which were produced exclusively without modifier, exclusively with modifier, or both with and without modifier. Most concept classes differ a lot from each other, comparing the relation of the three modifier status. Nevertheless, within the broader concept class of animals and plants and the broader concept class of man-made objects and body parts, a general similarity is evident: For animals and plants, less (or about equally many) part relations were produced exclusively without modifiers, in contrast to the other concept classes, where much more were produced without modifiers – the implement class being the only exception.

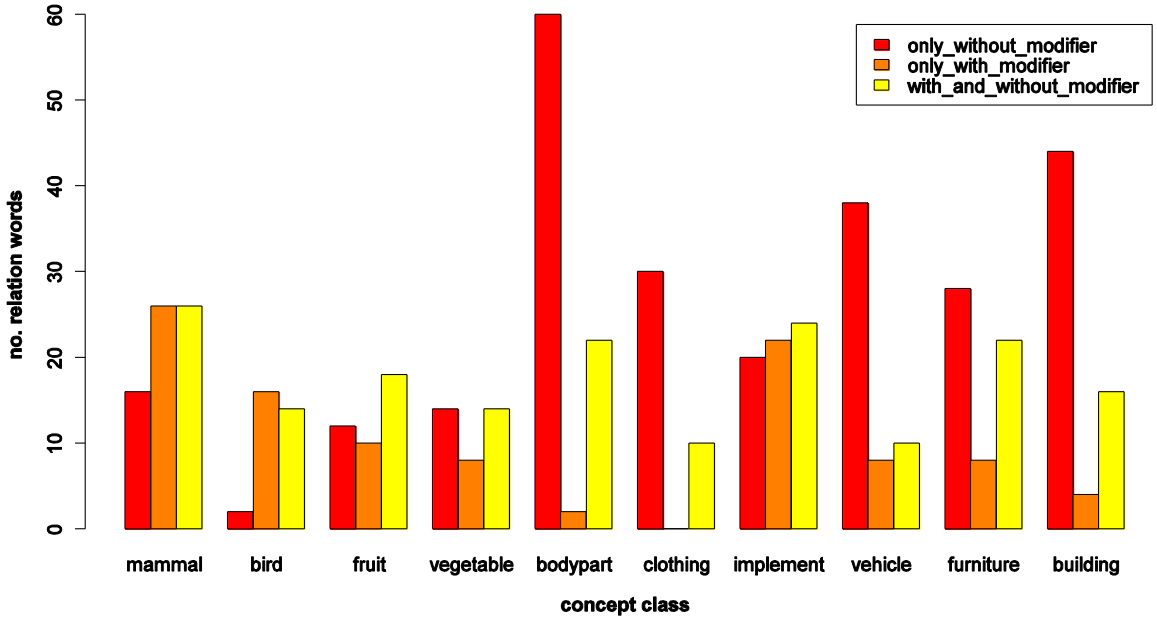

Figure 2. Numbers of part relation words produced - for 3 different modifier status.

The above findings suggest a more elaborate extraction method for composite part relations, taking into account the differences in modifier status between concept classes which were found in the production data. Apart from that, a large web corpus seems to be representative enough for the production experiment data in the aspects investigated.

[2] 28% when leaving out those pairs where the composite relation was not at all in the corpus

5. Conclusion

The relevance of our approach for lexicography with special focus on electronic pedagogical lexicography has been shown and its importance within the context of a nascent cognitive eLexicography has been discussed. The findings of our ongoing research support the consideration of empirical evidence from cognitive experiment results for the enrichment of lexical databases and dictionaries with semantic relations. After the preliminary investigations on (composite) part relations, the development of methods for the automatic extraction of cognitively salient semantic relations is the next step in this project. However, questions regarding the integration as well as the presentation of semantic relations in lexical databases or dictionaries is an open task for subsequent research.

References

Abel, A. (2003). *Alte und neue Problematiken der Lernerlexikographie in Theorie und Praxis*. Innsbruck. Dissertation (unpublished).

Abel, A., Campogianni, St.; Reichert, J. (2004). 'Wortfelder in einem zweisprachigen elektronischen Lernerwörterbuch: Darstellung der paradigmatischen Bedeutungsbeziehungen in der pädagogischen Lexikographie am Beispiel von ELDIT. In Williams, G.; Vessier, S. (eds.). *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004*. Lorient: UBS. Vol. II. 437-442.

Abel, A.; Campogianni, St. (2005). 'Facetten der Bedeutungsbeschreibung – ein integrativer Ansatz in der elektronischen Lernerlexikographie (aufgezeigt am Beispiel von ELDIT)'. In Mård-Miettinen, K.; Niemilä, N. (eds.). *Fachsprachen und Übersetzungstheorie. Vakki-Symposium XXV., Vörå 12.-13.02.2005*. Publikationen der Studiengruppe für Fachsprachenforschung. Vaasa: Universität Vaasa. 62-72.

Abel, A.; Weber, V. (2000). 'ELDIT – A Prototype of an Innovative Dictionary'. In Heid, U.; Evert, St.; Lehmann, Egbert et al. (eds.). *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000*. Stuttgart: Universität Stuttgart. Vol. II. 807 - 818.

Almuhareb, A.; Poesio, M. (2004). 'Attribute–Based and Value–Based Clustering: An Evaluation'. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Morristown (NJ, USA): Association for Computational Linguistics. 158-165.

Bergenholtz, H.; Tarp, S. (2002). 'Die moderne lexikographische Funktionslehre. Diskussionsbeitrag zu neuen und alten Paradigmen, die Wörterbücher als Gebrauchsgegenstände verstehen.' In *Lexicographica* 18. 253-263.

Bergenholtz, H.; Tarp, S. (2003). 'Two opposing theories: On H.E. Wiegand's recent discovery of lexicographic functions'. In *Hermes* 31. 171-196.

Casiddu, M. B. (1996a). 'Lessico mentale e produzione verbale. Modelli psicolinguistici e applicazioni didattiche'. In *Lingua e Nuova Didattica* 96 (2). 47-58.

Casiddu, M. B. (1996b). 'Lessico mentale e produzione verbale. Modelli psicolinguistici e applicazioni didattiche'. In *Lingua e Nuova Didattica* 96 (3). 23-36.

Chanier, T.; Selva, T. (1998). 'The alexia system: The Use of Visual Representations to Enhance Vocabulary Learning'. In *Computer Assisted Language Learning* 11 (5). 489-522.

Dizionario sinonimi e contrari (1999). Stopelli, P. (ed.). Milano: Garzanti.

Duden. Die sinn- und sachverwandten Wörter. Synonymwörterbuch der deutschen Sprache (1997). Müller, W. (ed.). Mannheim: Dudenverlag.

ELDIT (Electronic Learners' Dictionary German – Italian) [on line] http://www.eurac.edu/eldit [Access date: 22 Feb. 2010]

Fellbaum, C. (ed.). (1998). *WordNet: An Electronic Lexical Database. Language, Speech, and Communication*. Cambridge: MIT Press.

Geckeler, H. (2002). 'Anfänge und Ausbau des Wortfeldgedankens'. In Cruse, D. Alan; Hundsnurscher, F.; Job, M. et al. (eds.). *Lexikologie. Ein internationales Handbuch zur Natur und Struktur von Wörtern und Wortschätzen*. (Handbücher zur Sprach- und Kommunikations-wissenschaft; Bd. 21). Berlin – New York: de Gruyter. 1. Teilband. 713-728.

Hearst, M. A. (1992). 'Automatic Acquisition of Hyponyms From Large Text Corpora'. In *Proceedings of the Fiftheenth International Conference on Computational Linguistics* (Coling 1992). Morristown (NJ, USA): Association for Computational Linguistics. 539-545.

Hoberg, R. (1970). *Die Lehre vom sprachlichen Feld. Ein Beitrag zu ihrer Geschichte, Methodik und Anwendung*. Düsseldorf (= Sprache der Gegenwart 11): Schwann.

Italian WordNet [on line] http://multiwordnet.fbk.eu/online/multiwordnet.php [Access date: 22 Feb. 2010]

Jackson, H. (2003). *Lexicography. An Introduction*. London – New York: Routledge.

KirrKirr [on line] http://nlp.stanford.edu/kirrkirr/ [Access date: 22 Feb. 2010]

Knapp, J. (2004). *A new approach to CALL content authoring*. Hannover (Dissertation, http://www.eurac.edu/NR/rdonlyres/83E0D545-899B-45A7-B144-281140FB9B9E/0/knappPhD.pdf)

Kremer, G.; Abel, A.; Baroni, M. (2008). 'Cognitively Salient Relations for Multilingual Lexicography'. In Zock, M.; Huang, Ch.-R. (eds.). *Proceedings of the Workshop on Cognitive Aspects of the Lexicon* (COGALEX 2008), Manchester, UK. Brighton: One Digital. 94-101.

Longman Language Activator. The World's First Production Dictionary (1999). Summers, D. (ed.). Longman: Harlow.

Manning, Ch. D; Jansz, K.; Indurkhya, N. (2001). 'Kirrkirr: Software for browsing and visual exploration of a structured Warlpiri dictionary'. In *Literary and Linguistic Computing* 16 (2). 135-151.

McRae, K.; Cree, G. S.; Seidenberg, M. S.; McNorgan, C. (2005). 'Semantic Feature Production Norms for a Large Set of Living and Nonliving Things'. In *Behaviour Research Methods, Instruments & Computers* 37 (4). 547-559.

Pantel, P.; Pennacchiotti, M. (2006). 'Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations'. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics* (ACL). Morristown (NJ, USA): Association for Computational Linguistics. 113-120.

Visual Thesaurus [on line] https://www.visualthesaurus.com/ [Access date: 22 Feb. 2010]

WaCky initiative [on line] http://wacky.sslmit.unibo.it [Access date: 22 Feb. 2010]

Wiegand, H. E. (1987). 'Zur handlungstheoretischen Grundlegung der Wörterbuch-benutzungsforschung'. In *Lexicographica* 3. 178-227.

Wiegand, H. E. (1998). *Wörterbuchforschung: Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie*. 1. Teilbd. Berlin - New York: de Gruyter.

Wortschatz-Portal – Universität Leipzig [on line] http://wortschatz.uni-leipzig.de/ [Access date: 22 Feb. 2010]

Wu, L.; Barsalou, L. W. (2009). 'Perceptual Simulation in Conceptual Combination: Evidence From Property Generation'. In *Acta Psychologica* 132. 173-189.