Sonya S. Nikolova, **Jordan Boyd-Graber**, and Christiane Fellbaum. **Collecting Semantic Similarity Ratings to Connect Concepts in Assistive Communication Tools**. *Modeling, Learning and Processing of Text Technological Data Structures*, 2011, 11 pages.

```
    @inbook{Nikolova:Boyd-Graber:Fellbaum-2011,
Author = {Sonya S. Nikolova and Jordan Boyd-Graber and Christiane Fellbaum},
Title = {Collecting Semantic Similarity Ratings to Connect Concepts in Assistive Communication Tools},
Editor = {Angelika Storrer},
Booktitle = {Modeling, Learning and Processing of Text Technological Data Structures},
Series = {Studies in Computational Intelligence},
Address = {Heidelberg},
Url = {http://umiacs.umd.edu/~jbg//docs/2011_book_chapter_evocation.pdf},
Publisher = {Springer Verlag},
Year = {2011},
}
```

Links:

- Ratings [http://wordnet.cs.princeton.edu/downloads.html]

Contact Jordan Boyd-Graber (jbg@boydgraber.org) for questions about this paper.

Collecting Semantic Similarity Ratings to Connect Concepts in Assistive Communication Tools

Sonya Nikolova, Jordan Boyd-Graber, and Christiane Fellbaum

Department of Computer Science, 35 Olden Street
Princeton University, Princeton NJ 08540, USA
nikolova@princeton.edu
WWW home page: http://wordnet.cs.princeton.edu

Abstract. To compensate for the common inability of people with lexical production impairments to access and express intended concepts, we make use of models of human semantic memory that build on the notion of semantic similarity and relatedness. Such models, constructed on evidence gained from psycholinguistic experiments, form the basis of a large lexical database, WORDNET. We augment WORDNET with many additional links among words and concepts that are semantically related. Making this densely connected semantic network available to people with anomic aphasia through assistive technologies should enable them to navigate among related words and concepts and retrieve the words that they intend to express.

1 Background and Motivation

In this section, we briefly review a debilitating language disorder known as anomic aphasia: the inability to access, retrieve, and produce words. Technology can aid people suffering from the failure to generate the words they wish to express. Our work is motivated by the belief that the effectiveness of such tools can be enhanced by our knowledge of human semantic memory.

1.1 Aphasia

Estimated to affect 1 million people in the United States alone, aphasia is an acquired disorder that impacts an individual's language abilities [1]. It can affect speaking, language comprehension, and writing to varying degrees in any combination in an individual. Rehabilitation can reduce the impairment level, but a significant number of people with aphasia are left with a life-long chronic disability that impacts a wide range of activities and prevents full re-engagement in life. Aphasic individuals often employ different techniques in order to compensate for their inability to communicate; for example, they write notes, gesture, draw, or mimic.

There have been sustained efforts to use technology to help individuals with aphasia communicate. Designing technology that satisfies the needs and expectations of the intended user is a fundamental challenge in the field of human-computer interaction research. This is particularly challenging when designing technology

for people with aphasia due to the variability of impairment. The failure of existing assistive communication tools to address the problems arising from the heterogeneity of the user population has stimulated additional research efforts that show it is essential to seek flexible and customizable solutions [2–4].
Despite efforts to design adaptive assistive tools for elderly and cognitively disabled people, none has proven to be usable by aphasic individuals. Such aids mainly include scheduling and prompting systems that aim to reduce the burden of caregivers [5–8]. On the other hand, most assistive tools for people with aphasia focus on essential therapeutic efforts and the recovery of basic language function. Thus, they do little to leverage the skills of individuals with some residual communicative ability [9]. There have been relatively few systems for non-therapeutic purposes to be used by less severely affected individuals, such as systems that support daily activities like email or social interactions [10, 3].
In addition, the growing ubiquity of personal electronics, whose form factor could address the stigma attached to communicating with the help of a computer, has not benefited individuals with aphasia. In our experience [2], the weakest link is the ability for users to intuitively and quickly select words. Users, particularly those suffering from anomic aphasia, are confused by arbitrary organization of vocabulary terms and terms absent from the vocabulary. The real difficulty is in providing a flexible system in terms of adding new vocabulary items, adapting to users, and minimizing the complexity of navigating the vocabulary. While we are interested in addressing all of these issues, in this work we focus on vocabulary navigation.

1.2 ViVA: Visual Vocabulary for Aphasia

Although initial vocabulary sets can be formed from words frequently needed by the target population, no packaged system has the depth or breadth to meet the needs of every individual. In addition to expressiveness, vocabulary organization and retrieval in existing assistive technology are also problematic. Most of the existing visual vocabularies have a lexical organization scheme based on a simple list of words. The words are organized either in hierarchies which tend to be deep and non-intuitive or in a long list of arbitrary categories. Disorganization and inconsistency result in fruitless scrolling, backtracking, and ultimately frustration. It is important to build an easy to construct and maintain visual vocabulary that rests on a framework of a well-structured computerized vocabulary.
We have developed a multi-modal visual vocabulary that relies on a mixed-initiative design and enables the user to compose sentences and phrases efficiently. The visual vocabulary for aphasia (ViVA) implements a novel approach that organizes the words in the vocabulary in a context-aware network tailored to a user profile that makes finding words faster. ViVA is designed to reorganize and update the vocabulary structure depending on links created between words due to specific user input and system usage.

2 The Design of ViVA

In this section, we aim to describe the design of ViVA to show how semantic similarities can play a role in creating a better communication aid for people with aphasia. Our goal for ViVA is for it to be adaptable, able to be customized by the

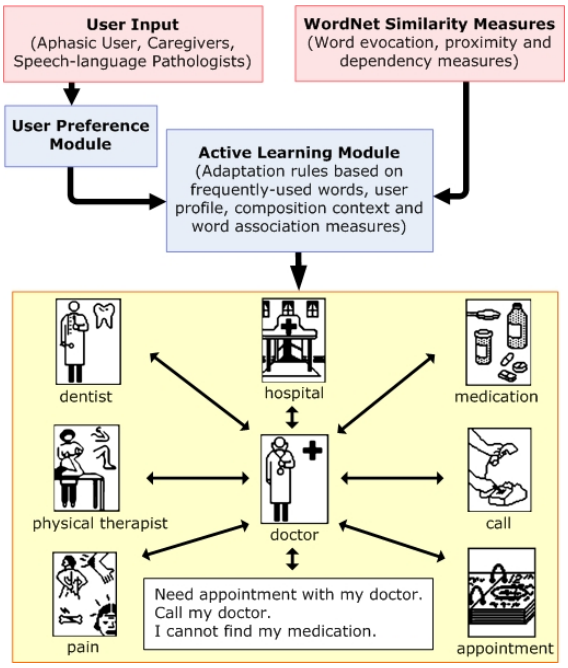

Fig. 1. Schematic of components of a system to assist individuals with aphasia

user, in addition to being adaptive, able to dynamically change to better suit the user's past actions and future needs.

The first component, adaptivity, allows the user to add and remove vocabulary items, group them in personalized categories (for example a "Favorites" folder or ideas related to "Family"), enhance words with images and sounds and associate existing phrases and sentences with a concept. In addition to practical concerns of having sufficient vocabulary terms to express the needed concepts, the ability to adapt a system invests in the user a sense of ownership and empowerment. This attachment to the system, brought about by a sense of accomplishment, is an important aspect of the rehabilitation process [11].

We explain the adaptive component with an example. For example, if the user wishes to compose the phrase "I need an appointment with my doctor." and she searches for `[doctor]` first, the vocabulary network centered on `[doctor]` may look as the one shown in Figure 1. The links between the words may exist because the user has previously composed sentences using `[doctor]` and `[medication]` or using `[doctor]` and `[appointment]`. `[hospital]` and `[doctor]`, for example, may be linked because of a prediction based on known word association measures and usage. In addition, the user may be able to find the phrase "Need appointment with doctor" right away if she had already composed it in the past and had linked it deliberately to `[doctor]`.

This is in contrast to existing systems that have a dichotomy between user-created organization and content and the initial vocabulary. Our goal is to let the user seamlessly add new content and for the organizational structure to change to bet-

ter suit usage needs. However, we still need an initial organization to allow the user to successfully use ViVA from day one. We derive this scaffold from the body of work investigating how the human brain organizes concepts.

2.1 The Organization of Words

To address the fundamental issues which prevent individuals with aphasia from effectively using communication aids, we appeal to the psychological literature on speakers' "mental lexicon," where words are stored and organized in ways that allow efficient access and retrieval. Our goal is to build a system that can help provide the missing semantic connections in the mental lexicon for sufferers of aphasia. Thus, any successful system must provide an ersatz mental lexicon that users can easily and naturally navigate and explore.

The tip-of-the-tongue (TOT) phenomenon is familiar to every speaker: the temporary inability to retrieve from our mental lexicon a specific word needed to express a given concept. This access failure may be due to a variety of factors, including fatigue and interference from a word that is morphologically or phonologically similar to the target word. People with anomic aphasia can be thought of as suffering from a chronic and severe case of TOT, as they have persistent difficulties accessing and retrieving words that express the concepts they wish to communicate.

Experimental evidence — including evidence from TOT states induced in the laboratory — suggests that words are organized in speakers' mental lexicons by various similarity relations, in particular phonological and semantic similarity. For example, subjects in word association experiments overwhelmingly respond with *husband* to the stimulus *wife* [12]. Semantic priming [13], a robust and powerful tool for the experimental investigation of cognitive processes, relies on the semantic relatedness of the prime and an experimental target: responses to the target are faster when it is related to the prime as in the classic case *doctor-nurse*. Spreading network activation models [14] assume that presenting a prime stimulus word activates the corresponding representation in lexical memory and that this activation spreads to other related nodes, thus facilitating the processing of related target words. The semantic network WORDNET [15, 16] is a large-scale lexical database inspired by network theories of semantic memory that accommodate the spreading activation paradigm among related words and concepts.

2.2 WORDNET and Evocation

WORDNET has a rich structure connecting its component synonym sets (synsets) to one another. Noun synsets are interlinked by means of hyponymy, the *super-subordinate* or *is-a relation*, as exemplified by the pair `[poodle]-[dog]`.[1] Meronymy, the *part-whole* or *has-a* relation, links noun synsets like `[tire]` and `[car]` [15]. Verb synsets are connected by a variety of lexical entailment pointers that express manner elaborations `[walk]-[limp]`, temporal relations `[compete]-[win]`, and causation `[show]-[see]` [16]. The links among

[1] Throughout this article we will follow the convention of using a single word enclosed in square brackets to denote a synset. Thus, `[dog]` refers not just to the word dog but to the set – when rendered in its entirety – consisting of {`dog, domestic dog, canis familaris`}.

the synsets structure the noun and verb lexicons into hierarchies, with noun hierarchies being considerably deeper than those for verbs.
We aim to exploit the structure of WORDNET to help "find" intended concepts and words by navigating along the paths connecting WORDNET's synsets. However, WORDNET's internal density is insufficient – there are too few connections among the synsets. Boyd-Graber et al. [17] represents an attempt to create thousands of new links that go beyond the relations specified in WordNet. This measure is called "evocation" as it attempts to measure how much one concept brings to mind another.
In total, evocation [17] aims to add cross-part-of-speech links, connecting nouns to verbs and adjectives. Such syntagmatic relations allow for connections among entities (expressed by nouns) and their attributes (encoded by adjectives); similarly, events (referred to by verbs) can be linked to the entities with which they are characteristically associated. For example, the intuitive connections among such concepts as `[traffic]`, `[congested]`, and `[stop]` should be encoded in WORDNET. This paper [17] also addressed another shortcoming of WORDNET, namely the absence of weights that indicate the semantic distance between the members of related pairs.
These human judgements of evocation were collected via a laborious, expensive method. Undergraduate students were put through a training and vetting process to consistently rate pairs of synsets through a specially designed interface. Because the pairs of synsets were randomly selected, many of the ratings, as expected, were zero. Although we originally hoped that these initial ratings, collected over the course of year and with a significant outlay of time and money, would allow us to automatically label the rest of WORDNET with directed, weighted links, machine learning techniques could not reliably replicate human ratings. In Section 3, we propose a method to collect the same valuable empirical similarity ratings using a far less expensive annotation strategy.

2.3 The Vocabulary

The vocabulary that we used in this application came from two sources: the "core" WORDNET consisting of frequent and salient words selected for the initial collection of evocation data and the visual vocabulary of an assistive device for people with aphasia created by Lingraphicare [18]. We used all synsets from the core 1000 synsets used in our initial evocation study, all verbs in Lingraphicare's vocabulary, and all nouns and adjectives in both Lingraphicare's vocabulary and the core 5000 synsets. Forming the initial vocabulary set from these three sources results in a collection of commonly used words and ones relevant to our target population, people suffering from anomic aphasia.
Lingraphicare's vocabulary pairs a word with a picture and a sound, performing a form of coarse disambiguation. For each concept in Lingraphicare's vocabulary, we selected the corresponding concept from WORDNET to create a single, unified representation of the vocabulary.

3 Collecting Inexpensive Ratings from Untrained Annotators

Many natural language processing tasks such as determining evocation require human annotation that is expensive and time-consuming on large scale. Snow

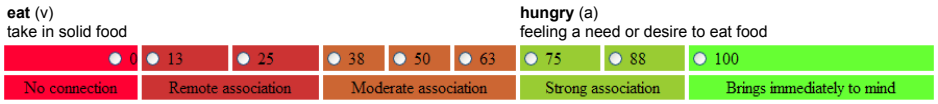

Fig. 2. Example stimulus for collecting evocation ratings. A user rates 50 pairs in a single sitting.

et al. [19] demonstrated the potential of Amazon's Mechanical Turk [20] as a method for collecting a large number of inexpensive annotations quickly from a broad pool of human contributors. Their experiment illustrated that labels acquired through Amazon Mechanical Turk (AMT) from non-expert annotators are in high agreement with gold standard annotations from experts. The positive results of their work motivated us to collect evocation ratings to be used in the visual vocabulary for aphasia through a Mechanical Turk experiment described in this section.

3.1 Method

We used a machine learning algorithm to select the synset pairs to be rated via AMT annotators. We used many of the features found to be predictive of evocation including those based on WordNet connectivity [21], pointwise mutual information based on sentence appearing in the same sentence, and context similarity. We duplicated high evocation pairs (having a median rating of greater than 15) to create a high-recall training set, trained a classifier using AdaBoost [22], and then took the subset of all pairs of synsets in our vocabulary labeled as having a high predicted evocation by our learning algorithm. These pairs were the ones selected to be rated via AMT.

We created 200 tasks consisting of 50 pairs each. The design of the template we posted on AMT was closely modeled after the computer program used by Boyd-Graber et al. [17] to collect ratings from undergraduate annotators. Anchor points on a scale from from 0 to 100 were available to rate evocation (Figure 2). Raters were first presented with the following set of instructions:

1. Rate how much the first word brings to mind the second word using the provided scale.
2. The relationship between the two words is not necessarily symmetrical. For example, "dollar" may evoke "green" more than the reverse.
3. Pay attention to the definition of the words given on the second line; words can have more than one meaning. For example "dog" (the animal) would not bring to mind "bun" (the piece of bread you serve with a hot dog).
4. The letter in parenthesis signifies whether the word is *a*: an adjective, *n*: a noun or *v*: a verb.
5. Don't use information from your personal life. For example, if you had a dog named "bog" you personally would associate "bog" and "dog," but the average person wouldn't.
6. Don't use the spelling of words to make your decisions. For example, even though "bog" and "dog" rhyme, they are not associated.
7. We cannot offer you a big reward for your time, but we greatly appreciate your sincere effort. There are a few pairs with known average ratings embedded in the task. If your ratings for those pairs do not fall in a generously set acceptance bounds, we will have to reject your responses.

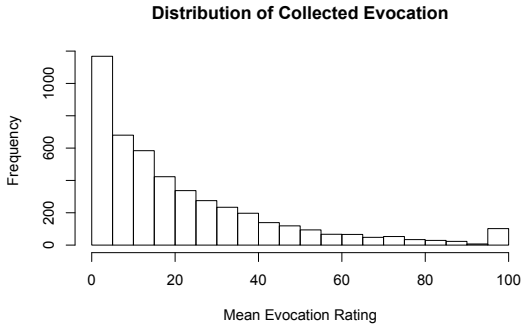

Fig. 3. The distribution of the mean of the evocation pairs collected. The extra bump at 100 is because same synset checks were placed in the tasks to ensure annotator reliability.

Filtering Method	Correlation with Mean	Correlation with Median	Number Ratings
All Checks	0.604	0.563	14850
Most Checks	0.529	0.484	23700
Some Checks	0.355	0.279	24750

Table 1. Correlation of the mean and median against evocation annotations collected by trained undergraduate annotators.

The last instruction was included to forewarn annotators that sloppy contributions such as clicking all zeros will not be rewarded. We embedded five checks, unknown to the annotators, in each task which were later used to determine the validity of the gathered results. Annotators were paid $0.07 to complete a task.

4 Results

We collected 2990 completed tasks in a period of ten days. The average time to complete a task was 4.5 minutes, resulting in an average pay of $0.92 per hour. To ensure the quality of the ratings and a consistency with previous results, we used embedded checks to decide which submitted tasks were valid. The ratings for four of those checks were collected from the dataset provided by Boyd-Graber et al. [17]. The fifth check required annotators to rank a pair consisting of the same synset, for example `[help]` and `[help]`.

We ran three different reliability tests depending on the number of checks we wanted satisfied. If the annotator's rating for the fifth check was 100 and a number of the remaining checks were met within certain acceptance bounds, the annotations were considered valid. The acceptance bounds were defined as follows. As in the task, the scale of 0 to 100 was split into 5 intervals, [0-10), [10-30), [30-70), [70-90),[90-100]. If an annotator's rating fell within the same interval as the corresponding check or in the immediately lower or higher intervals, the rating was considered valid. The first reliability test required **all** checks to be met.

Number of Checks			Trained	Synset 1	Synset 2
All	Most	Some	Undergraduates		
50	10	61	88	trust.v.01	responsible.a.01
39	44	41	44	surgeon.n.01	responsible.a.01
25	18	22	42	deservingness.n.01	exceed.v.02
31	31	28	33	philosopher.n.01	convert.v.03
29	30	30	20	television_receiver.n.01	performance.n.02
46	57	62	19	log.n.01	leaf.n.01
12	12	14	18	subject.n.06	check.v.22
34	33	31	16	diligence.n.02	craft.n.04
25	20	27	16	abundant.a.01	harmony.n.02
21	10	14	1	category.n.02	beginning.n.03
23	19	18	0	eyelid.n.01	wrist.n.01
25	28	26	0	reason.n.02	reference_point.n.01
4	5	9	0	spread.n.05	pill.n.02

Table 2. Examples of mean evocation ratings given three different methods to ensure rater reliability. For comparison, evocation ratings from trained undergraduates are also shown.

For this set, 40.2% of the pairs were rated as having moderate or no association and 3.5% fell in the category immediately brings to mind (see Figure 3).

The second reliability test required **most**, three or more, checks to be met in addition to satisfying the complete-evocation check. The final and most relaxed reliability test required **some**, two or more, checks to be met in addition to the complete-evocation check. Table 1 shows the number of synset pair ratings for each of the reliability levels, and Table 2 has explicit examples of mean evocation ratings for the three levels.

Finally, Table 1 shows mean and median correlation of the three reliability sets against the ratings provided by undergraduate students in [17]. As expected, the results get noisier when less strict checks are applied. The set of synsets where all checks were met results in the highest correlation to the original evocation data. While it is not very high, it is sufficient to show that with good quality control, gathering ratings through AMT was a valid approach. While AMT annotators seem to rate on average evocation lower than the trained annotators, as seen from Figure 4, the inter-annotator agreement in the most reliable set and the original evocation set are comparable.

5 Discussion

While the results may appear less compelling than one might have expected, it is important to bear in mind the difficulty of the task. First, the nature of the task was such that we asked the participants to actively *produce* a rating, rather than to agree or disagree with a pre-set judgment or to select one from a few pre-defined options. Second, the ratings were to be expressed on a scale from 0 to 100, thus allowing for – and in fact, encouraging – very subtle judgments that permitted significant disagreement. Third, while we we controlled for intra-rater reliability, we did not know who our raters were in terms of educational level, literacy, and familiarity with the words and concepts that were presented. Indeed, we had no

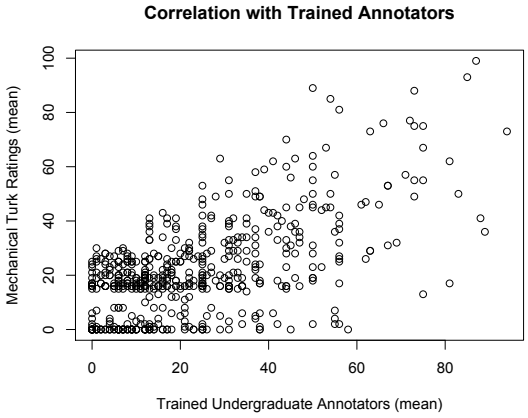

Fig. 4. Ratings from untrained annotators on the web correlated well (0.504) with those collected by trained undergraduate annotators.

way to ascertain that the raters were native or near-native speakers of English. Finally, the raters might have received insufficient training given the cognitive demands of the task.

The results must be compared to those obtained in the carefully controlled study reported by Boyd-Graber et al. [17]. At the outset of that experiment, it was unknown whether any reasonable reliability could be obtained at all, as we were well aware of the difficulty of the task, for which no precedent existed, and we considered the results encouraging. Our raters came from a small, homogeneous pool — Princeton undergraduate students — whose identity we knew and whom we trained carefully and with personal feedback. In light of the different methods of data collection, the results of the current study are comparable. The inherent noise in the evocation data reflects idiosyncrasies in world knowledge; only by accepting this reality and incorporating it into assistive technologies can we hope to build devices that can truly help a heterogenous target population.

Even though our strictest reliability test invalidated half of the collected data, using AMT to gather evocation is still more efficient and economical compared to using trained annotators. This collection of reliable evocation ratings adds on to the scaffolding of our assistive vocabulary by providing meaningful links between words. Such links will compensate for impaired access to the user's "mental lexicon" and assist her in communicating. A network of words whose organization reflects human semantic memory has the potential to help users with anomic aphasia navigate the vocabulary more naturally and thus find what they are trying to express faster.

6 Acknowledgments

This work was partially supported by a 2007 Intelligent Systems for Assisted Cognition Grant by Microsoft Research. We thank Perry Cook for his feedback

and Lingraphicare Inc. for providing us with access to their icon library. The results from our experiments are available for download at:
http://wordnet.cs.princeton.edu/downloads.html.

References

1. The National Aphasia Association: Aphasia: The facts. http://www.aphasia.org.
2. Boyd-Graber, J.L., Nikolova, S.S., Moffatt, K.A., Kin, K.C., Lee, J.Y., Mackey, L.W., Tremaine, M.M., Klawe, M.M.: Participatory design with proxies: Developing a desktop-PDA system to support people with aphasia. In: Proc. CHI 2006, ACM Press (2006) 151–160
3. Moffatt, K., McGrenere, J., Purves, B., Klawe, M.: The participatory design of a sound and image enhanced daily planner for people with aphasia. In: Proc. CHI 2004, ACM Press (2004) 407–414
4. van de Sandt-Koenderman M, M., Wiegers, J., Hardy, P.: A computerised communication aid for people with aphasia. Disability Rehabilitation **27**(9) (2005) 529–533
5. Carmien, S.: MAPS: PDA scaffolding for independence for persons with cognitive impairments. In: Human-computer interaction consortium. (2002)
6. Haigh, K., Kiff, L., Ho, G.: The Independent LifeStyle AssistantTM (I.L.S.A.): Lessons Learned. Assistive Technology (18) (2006) 87–106
7. Levinson, R.: PEAT: The planning and execution assistant and trainer. Journal of Head Trauma Rehabilitation (12(2)) (1997) 769–775
8. Pollack, M.E., Brown, L., Colbry, D., McCarthy, C.E., Orosz, C., Peintner, B., Ramakrishnan, S., Tsamardinos, I.: Autominder: An intelligent cognitive orthotic system for people with memory impairment. Robotics and Autonomous Systems (44) (2003) 273–282
9. Beukelman, D.R., Mirenda, P.: Augmentative and alternative communication: Management of severe communication disorders in children and adults. Brookes Publishing Company (1998)
10. Daeman, E., Dadlani, P., Du, J., Li, Y., Erik-Paker, P., Martens, J., Ruyter, B.D.: Designing a free style, indirect, and interactive storytelling application for people with aphasia. In: INTERACT. (2007) 221–234
11. Allen, M., McGrenere, J., Purves, B.: The design and field evaluation of phototalk: a digital image communication application for people. In: Assets '07: Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility, New York, NY, USA, ACM (2007) 187–194
12. Moss, H., Older, L.: Birkbeck Word Association Norms. Psychology Press (1996)
13. Swinney, D.: Lexical access during sentence comprehension: (Re)consideration of context effects. Journal of Verbal Learning and Verbal Behavior (18) (1979) 645–659
14. Collins, A.M., Loftus, E.F.: A spreading-activation theory of semantic processing. Psychological Review **82**(6) (November 1975) 407–428
15. Miller, G.A.: Nouns in WordNet: A lexical inheritance system. International Journal of Lexicography **3**(4) (1990) 245–264
16. Fellbaum, C.: A semantic network of English verbs. In: WordNet : An Electronic Lexical Database. MIT Press, Cambridge, MA (1998)

17. Boyd-Graber, J., Fellbaum, C., Osherson, D., Schapire, R.: Adding dense, weighted, connections to WordNet. In Sojka, P., Choi, K.S., Fellbaum, C., Vossen, P., eds.: Proc. Global WordNet Conference 2006, Brno, Czech Republic, Global WordNet Association, Masaryk University in Brno (January 2006) 29–35
18. Lingraphicare Inc.: Lingraphica http://www.aphasia.com/.
19. Snow, R., O'Connor, Jurafsky, D., Ng, A.: Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. Proceedings of EMNLP-08 (Jan 2008)
20. Amazon.com, Inc.: Amazon mechanical turk https://www.mturk.com.
21. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings on International Conference on Research in Computational Linguistics, Taiwan (1997)
22. Schapire, R.E.: The boosting approach to machine learning: An overview. In Denison, D.D., Hansen, M.H., Holmes, C., Mallick, B., Yu, B., eds.: Nonlinear Estimation and Classification. Springer (2003)