What's in a Definition? An Investigation of Semantic Features in Lexical Dictionaries

Luigi Di Caro[a]
Department of Computer Science, University of Turin, Italy

Keywords: Lexical Semantics, Word Meaning, Word Definitions, Backward Dictionaries, Lexical Ambiguity.

Abstract: Encoding and generating word meanings as short definitions for user- and machine-consumption dictionaries is still a usually adopted strategy within interpretable lexical-semantic resources. Definitions have the property of being natural to be created by humans, while several techniques have been proposed for the automatic extraction from large corpora. However, the reversed process of going back to the words (i.e., onomasiological search) is all but simple for both humans and machines. Indeed, definitions show context- and conceptual-based properties which influence their quality. In this contribution, I want to draw the attention to this problem, through a simple content-to-form experimentation with humans in the loop. The results give some first insight on the relevance of the problem from a computational perspective. In addition, I analyzed (both quantitatively and qualitatively) a set of 1,901 word definitions taken from different sources, towards the modeling of features for their generation and automatic extraction.

1 INTRODUCTION

Nowadays, Natural Language Processing is becoming more than one among the research areas within the Artificial Intelligence field. With the advent of large data repositories and new enabling technologies, the fascinating dream of machines interacting through natural language is experiencing renewed attention and horizons.

While research on *Semantics* has been always focusing on the philosophical aspects of meaning and its relationship with language, *Computational Semantics* works towards machine-based encoding of meaning providing human-like capabilities such as similarity and reasoning processes. Being short and approximate, Formal Semantics models inferences at the symbolic level, while Distributional Semantics enables the computation of semantic similarity between symbols (i.e., lexical units) through corpora analysis. In other words, on the one hand, one can define semantics as the way of inferring knowledge given a set of facts and rules expressed through symbolic elements. On the other hand, semantics can be seen as the embodied meaning within such symbols, irrespective of their power to enable logical inferences about some global knowledge. Distantly from this dualism, Lexical Semantics represents a highly studied area aiming at developing semantic resources such as lexical inventories and ontologies to support a huge variety of semantic analysis tasks. However, the encoded meaning in such resources are often still of lexical type, such as definitions, glosses, and examples of use. In this paper, I first study the fragility of such representation of meaning through a simple *content-to-form* experimentation with 20 participants. In particular, I asked some of them to provide individual definitions on few concepts (of different types, as detailed in the paper). Then, the remaining participants had to guess by going back to the described words. This task is often associated with the name of *onomasiologic search*, and it relates with the well-known *tip-of-the-tongue* problem (Brown and McNeill, 1966).

What I found is that definitions resulted to be very fragile encodings of lexical meaning, even with simple concepts, and in a controlled scenario. Then, we looked deeper at the result of the experiment by analyzing each definition in terms of different criteria, trying to make some measurement of their quality, and considering its effectiveness in correctly indicating the unveiled words.

Then, I carried out an experimentation with a dataset of 1,901 word definitions about 300 random concrete concepts, extracted with the help of BabelNet (Navigli and Ponzetto, 2010), belonging to re-

[a] https://orcid.org/0000-0002-7570-637X

Di Caro, L.
What's in a Definition? An Investigation of Semantic Features in Lexical Dictionaries.
DOI: 10.5220/0010116902250232
In *Proceedings of the 16th International Conference on Web Information Systems and Technologies (WEBIST 2020)*, pages 225-232
ISBN: 978-989-758-478-7

sources such as Wordnet (Miller, 1995), Wikipedia[1], Wiktionary[2], OmegaWiki[3], and others. In particular, I quantitatively and qualitatively analyzed the type of information contained in the definitions, highlighting possible features to be used for the creation of better definitions rather than their automatic extraction from large corpora.

2 MOTIVATIONS AND RESEARCH QUESTIONS

Lexical Semantics is about encoding *lexical meaning*. This is crucial in many multilingual NLP tasks and applications. However, interpretable (as opposed to statistical and vector-based) lexical resources mostly rely on the paradox of providing word meanings through, again, the use of words. This forces the Word Sense Disambiguation (WSD) process to work on non-machine-based representations of lexical contexts and definitions. However, some questions arise and have already been approached computationally, e.g., (Muresan and Klavans, 2002; Klavans et al., 2003; Thorat and Choudhari, 2016). Some of them may be the following:

1. How consistent and reliable are the definitions of words?
2. How good are the definitions within lexical resources? Or, more generally, is it possible to measure the quality of a definition?
3. Are there any features that may rule the effectiveness of definitions?
4. Which semantic features and relations are typically covered by textual definitions?

In this paper, I carried out a simple experimentation with humans in the loop to put some light on these aspects.

In my perspective, we can roughly define the general quality of a definition as its ability to let people identify the described concept. If a definition carries to wrong guesses, then we could say it does not make its work properly (and it is difficult to think it could do a better job in automatic machine-based methods).

While there exists significant work on the principles behind definition writing, such as the *genus-differentia* mechanism (e.g., (Strehlow, 1983)), to the best of my knowledge, little effort has been put towards computational-oriented modeling of definitions quality. In this paper, I propose a first model, evaluating it with a restricted set of concepts and through a test with non-expert participants.

[1]https://www.wikipedia.org
[2]https://en.wiktionary.org/
[3]http://www.omegawiki.org/

3 RELATED WORK

Lexical semantics resources usually fall into categories such as computational lexicons, corpus-based models, semantic frames, and common-sense knowledge bases. They all define the meaning of words in terms of (sometimes categorized) textual descriptions. In this section, I briefly overview them under a content-to-form perspective.

3.1 Computational Lexicons

WordNet (Miller, 1995) may be considered as the most referenced and used computational lexicon for English. Counterparts in other languages (Bond and Foster, 2013) are also available. WordNet is produced by humans for humans, as concepts are described through word definitions, and contextualized in terms of paradigmatic relations such as hyperonymy and meronymy. BabelNet (Navigli and Ponzetto, 2010) is the result of a large-scale integration of WordNet with Wikipedia and other sources of semantic information. Most Word Sense Disambiguation (WSD) systems use these resources and their gloss-based model. However, glosses are often short and disclose very few semantic information from which is difficult to go back to the words. In Section 6 I present an experiment that puts some light on this issue.

3.2 Frames

(Fillmore, 1977) proposed semantic frames to encode meanings through slot-filler structures, and FrameNet (Baker et al., 1998) represents the largest frame-based resource available. Slots and fillers represent attributes and values respectively. While such representation could be used to better go back to the described items, actually, frames do not encode the individual meaning of concepts but their *situational* use. An interesting and novel slot-filler approach was presented by (Moerdijk et al., 2008) with the ANW dictionary and the introduction of the concept of *semagram*. A semagram is a conceptual structure that describes a lexical entity on the basis of a wide range of characteristics, defined with a rich slot-filler structure. The semagrams provided in the ANW dictionary are, however, limited in coverage, often expressed with a fragmented set of semantic slots and written in Dutch. In

(Leone et al., 2020), the authors revised the semagram structure to overcome these limitations.

3.3 Corpus-based Distributional Models

Corpus-based semantic models are based on the Distributional Hypothesis (Harris, 1954), i.e., words co-occurring in similar contexts tend to have similar meanings. Latent Semantic Analysis (Dumais, 2004), Latent Dirichlet Allocation (Blei et al., 2003), and, more recently, embeddings of words (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2016) and word senses (Huang et al., 2012; Iacobacci et al., 2015) represent vectorial / geometrical representations of words. However, the relations holding between vector representations are not typed, nor are they organized systematically. While these representations work well for semantic similarity computation, vector dimensions are not interpretable concept descriptions. Conceptual Spaces (Gärdenfors, 2004) provide a geometric approach to meaning where vector dimensions are instead qualitative features (e.g., colors may be represented through hue, saturation, and brightness). However, the encoded knowledge does not define concepts explicitly, and dimensions usually represent perceptual mechanisms only.

3.4 Common-sense Knowledge

Common-sense knowledge resources may be described as a set of shared and general facts or views of a set of concepts. ConceptNet (Speer and Havasi, 2012; Speer et al., 2016) is one of the largest resources of this kind. However, terms in ConceptNet are not disambiguated, which leads to the confusion of lexical-semantic relations involving concepts denoted by ambiguous words. NELL (Carlson et al., 2010) matches entity pairs from seeds to extract relational phrases from a Web corpus, but it is mostly oriented to named entities rather than concept descriptions. Property norms (McRae et al., 2005; Devereux et al., 2014) represent a similar kind of resource, which is more focused on the cognitive and perception-based aspects of word meaning. Norms are based on empirically constructed semantic features via questionnaires asking people to produce features they think as important for some target concept (e.g., a *crocodile* is often associated with the norm *is-dangerous*). The problem with norms is that they do not represent complete descriptions (usually, only immediate and common-sense facts are reported).

4 FEATURES OF DEFINITIONS

As earlier mentioned, the main aim is to understand and evaluate the robustness of definitions as they still lie at the core of most lexical resources, e.g., WordNet. Differently from existing features related to writing quality such as (Witte and Faigley, 1981) and more recently (McNamara et al., 2010), definitions have not been yet analyzed and modeled from a computational perspective and under a backward-dictionary view. However, the actual connection between a word and its meaning is, at least by humans, strictly dependent on the robustness of its definition. Thus, this might also have some impact on computational approaches. In this contribution, I propose a simple model made up of three features:

1. *clarity*;
2. *richness*; and
3. *readability*.

For *clarity*, I mean the (main) feature of a definition of being non-ambiguous with respect to similar concepts (e.g., hypernyms). This is actually what I empirically measured with the experment that will follow. I then define *richness* as the quantity of semantic information[4] contained and *readability* as the lexical and syntactic simplicity and shortness of the definition. While the three features are slightly interconnected, they may have diverging scores within a single definition.

For example, a definition may contain several semantic relations without eliminating ambiguity. For example, the following definition of *terminal* can be erroneously associated with the concept *computer*:

> *In networking, a device consisting of a video adapter, a monitor, and a keyboard. The adapter and monitor and, sometimes, the keyboard are typically combined in a single unit.*

Contrariwise, another (very short) definition of *terminal*, while revealing little semantic content, is instead able to clearly identify the concept:

> *A device communicating over a line.*

In the next sections, I will present the results of two experiments aiming to highlight features and dynamics of definitions from both a human and a computer perspective.

[4]For semantic information I mean the set of semantic relations that can be grasped from the definition, such as paradigmatic ones (e.g., hypernyms, meronyms, etc.), physical (e.g., size, shape, color, etc.), behavioral (e.g., purpose, ways of use, etc.), and others.

5 HUMAN-IN-THE-LOOP EXPERIMENT

In this section, I describe how I conducted the experiment with the 20 participants, and the used criteria for the selection of the concepts. The aim of the experiment was to analyze the descriptions of simple concepts by (even non-expert) people under a computational perspective, and specifically towards an automatic content-to-form approach.

5.1 Participants

Participants were 20-35 years old students with different background (linguists, computer scientists, mathematicians, engineers). They were not aware of the goals of the experiment, while they have been introduced with some knowledge on Computational Linguistics, and specifically, on Lexical Semantics.

5.2 Methodology

The idea of the experiment was to test the capability of word definitions to uniquely identify the underlying concepts. In order to be significantly sure about the independence from single subjective views, I asked 12 out of the 20 participants (*def*-participants, from now on) to create definitions for all the concepts, leaving the remaining 8 (*test*-participants) for the later test phase. This way, by having 12 definitions for each concept, we can be rather confident that the results were not influenced by single and unfortunate definitions (the entire set of definitions has been given to the *test*-participants to make a single choice). I finally asked the *test*-participants to mark the best definition which has been mostly useful for giving the answer. This last step allowed us to correlate some features of the best-selected definitions with the accuracy obtained by the participants during the experiment.

5.3 Concepts Selection

Due to the choice of employing most of the participants in the creation of definitions, I had to limit the number of concepts. I have chosen 8 concepts and identified two criteria for the (hard) task of selecting them: *generality* (as opposed to specificity) and *concreteness* (as opposed to abstractness). Table 1 shows the selected concepts along with their characteristics, while Table 2 shows the obtained definitions for the concept *Screw*.

Table 1: Concepts used for the experiment, along with their characteristics.

	General	**Specific**
Abstract	*politics, justice*	*greed, patience*
Concrete	*food, vehicle*	*screw, radiator*

5.4 Results

In this section I report some insights gained from the analysis of the experiment results. The collected 96 definitions[5] have an average length of 56 characters (16 and 225 for the shortest and longest definitions respectively). No significant difference emerged from the different concept types (abstract / concrete / generic / specific).

In order to capture the effectiveness of the definitions, I carried out different measurement: 1) the percentage of correct guesses given by the *test*-participants on the 12 definitions[6] provided by the *def*-participants, aggregated by concept type; 2) the correlation between definitions features and the obtained accuracy levels.

Wrong guesses were of different types: 1) hypernyms or hyponyms (e.g., *vehicle* → *wheeled vehicle* → *motor vehicle*); 2) sister terms (e.g., *calm* instead of *patience*, and 3) less related concepts (e.g., *stove* instead of *radiator*). The lexical overlap among definitions for the same concepts is very low (less than 20% on average, using stemming and stopwords removal only), as it can be also deduced from the example of Table 2. Although semantic-aware lexical enrichment might increase such value, I left the *test*-participants to directly guess the underlying concepts. Table 3 shows the accuracy values aggregated by concept type. The obtained overall accuracy is 58.75%, meaning that even with more than ten times of lexical context at disposal (12 definitions plus 1 from WordNet), the participants were often not able to make the right guess.

Since the scale of the experiment can only have a limited significance value, this is however indicative of the fragility of definition-based lexical resources. While abstract vs concrete concepts revealed small reciprocal differences in the results, definitions of specific (rather than generic) concepts generally carried to more correct guesses. Intuitively, physical objects could be described in terms of specific words whereas abstractness generally requires steps of generalization involving a larger set of lexical items and syntactic

[5]The dataset will be made available in case of acceptance.

[6]Actually, the definitions given to the test-participants were 13 since I also included the WordNet gloss.

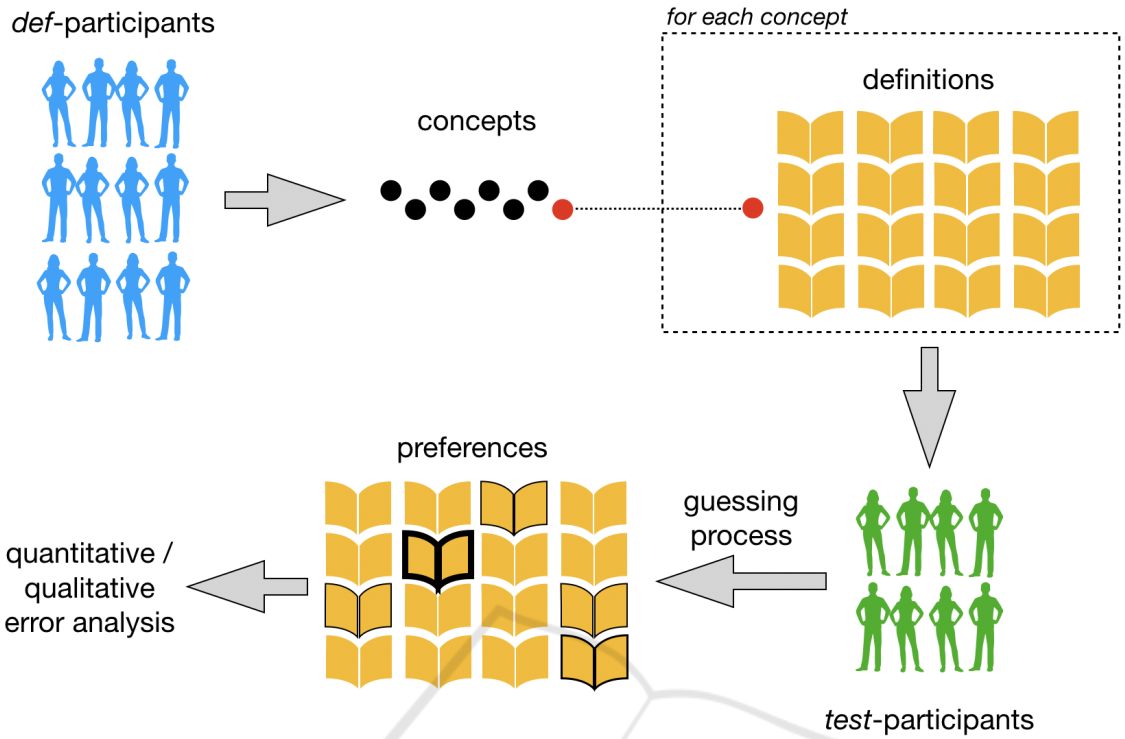

Figure 1: Methodology of the human-in-the-loop experiment.

Table 2: Collected definitions for the concept *Screw*. The most frequent terms (marked in bold) represent taxonomical information (*object, element, item, pin, fastener*), materials (*metal*), and usages as a tool (*used to* *). Other reported semantic information are related to parts (*slotted head*), size (*little*), shape (*helical, spiral*).

Screw (WN 1:06:00)
[WordNet] A fastener with a tapered threaded shank and a slotted head.
Item **used to** connect artificial parts together.
Metal pin with raised helical thread running around it.
Little **metal object** which can be inserted in a support.
Threaded **metal object used to** produce other artifacts.
Metal object used to fix combinable elements.
Object that is **used to** look and join other components.
Object useful to fix other objects on some surfaces, for example a painting on the wall.
Metal object with the shape of a spiral **used to** put things together.
A short, slender, sharp-pointed **metal** pin with a raised helical thread running around it and a slotted head, **used to** join things together by being rotated so that it pierces wood or other material and is held tightly in place.
Structural element needed to fix two parts.
Long and thin pointy item piercing two **objects** to hold them together.

structures.

Finally, by manually looking at the best definitions selected by the *test*-participants, I discovered the following insights. First, they converged, on average, on 4.08 definitions among the 13 at disposal (variance = 1.24). This is indicative of the fact that definitions do embody significant features which are recognized to be effective for clearly identifying the concepts. Second, best definitions contain a variable number of semantic relations (no correlation with *richness*). Third, they are on average 65% longer than non-selected ones (rough estimation of correlation with *readability*).

Table 3: Obtained accuracy values (percentage of correct answers given by the 10 *test*-participants), aggregated by category.

Accuracy	General	Specific	**Total**
Abstract	50%	75%	62.5%
Concrete	45%	65%	55%
Total	47.5%	70%	**58.75%**

6 COMPUTATIONAL EXPERIMENT

In this section, I describe the method and the results of a semantic analysis of a 1,901-sized corpus of word definitions covering 300 concrete (noun) concepts. The aim is to find some insights on the type and statistics related to the semantic information usually contained within definitions.

6.1 Methodology

As already mentioned in the Introduction, I made use of BabelNet (Navigli and Ponzetto, 2010) for the extraction of English definitions associated with an input set of 300 concepts (see the next Section 6.2 for details about the selection process). A total set of 1,901 word definitions have been retrieved, with an average number of definitions per concept of $6,34$ and an average number of tokens per definition of 14,55. Examples of definitions for the concept *salad* are reported below:

> **D.1** (WordNet): *Food mixtures either arranged on a plate or tossed and served with a moist dressing; usually consisting of or including greens.*
> **D.2** (Wikipedia): *A salad is a dish consisting of a mixture of small pieces of food, usually vegetables.*
> **D.3** (WikiData): *Dish of raw vegetables.*
> **D.4** (OmegaWiki): *Any of a broad variety of dishes, consisting of (usually raw) chopped vegetables, most commonly including lettuce.*
> **D.5** (Wiktionary): *A food made primarily of a mixture of raw or cold ingredients, typically vegetables, usually served with a dressing such as vinegar or mayonnaise.*

For each definition, I collected all its nouns and searched for the following semantic information:

- Presence of synonyms (e.g., *plane* $\leftrightarrow$ *airplane*);
- Presence of hypernyms (e.g., *plane* $\leftrightarrow$ *aircraft*);
- Presence of meronyms (e.g., *plane* $\leftrightarrow$ *wing*);
- Presence of purpose-related information (e.g., *plane* $\leftrightarrow$ *transportation*).

For the discovery of the last phenomenon, I made use of simple patterns (e.g., *used for*, *used to*, and *used as*).

6.2 Concepts Selection

As expected, the accuracy values reached in the first experiment were higher for concrete concepts (see Section 5). For this reason, I decided to use concrete concepts only. In particular, I manually created a set of 10 categories[7] covering different conceptual aspects, then picking 30 concepts for each category by making use of the WordNet hierarchy.

6.3 Results

An overview of the results of this experiment is shown in Table 4. As expected, there is a significant use of hypernyms (according to WordNet), which can be easily seen as the *genus* part of the definition. However, only about 30% of the definitions contain direct hypernyms, while most of the times more general hypernyms are used. Despite the selection of concrete objects, only the 10.57% makes use of meronyms. Similarly, almost the 11% contains purpose-related semantic information. Finally, an important issue raised by the experiment regards the usage of synonyms. This phenomenon often indicates the presence of *circularity* in the definitions, which can be problematic (even in not all cases, as demonstrated in (Burgess, 2007)).

7 CONCLUSIONS

In this contribution, I tried to open a discussion over the soundness of lexical word definitions as they still represent the main type of meaning encoding within semantic resources. A simple content-to-form experiment with different concepts is presented, testing their general capability to actually uncover the underlying concepts. Results show that definitions are very fragile means for going back to the concepts. This calls for further research on the quality and the features of definitions, with a need for novel interpretable encoding strategies. In addition, a further quantitative and qualitative analysis on 1,901 word definitions coming

[7]The categories have been taken from (Silberer et al., 2013) and include: animals, musical instruments, tools, artifacts, vehicles, food, clothes, home utensils, appliance, containers.

Table 4: Semantic analysis results of the 1,901 word definitions.

Phenomenon	N. of definitions	Perc. (%)
Presence of synonyms	599 out of 1,901	31.51%
Presence of hypernyms (direct)	583 out of 1,901	30.67%
Presence of hypernyms (2nd level)	996 out of 1,901	52.39%
Presence of hypernyms (3rd level)	1,254 out of 1,901	65,97%
Presence of hypernyms (all)	1,685 out of 1,901	88,63%
Presence of meronyms	201 out of 1,901	10.57%
Presence of purpose-relations	207 out of 1,901	10.89%

from different sources about 300 concrete concepts highlighted possible features for their automatic generation and extraction from large corpora.

REFERENCES

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *Proc. of ACL*, pages 86–90. Association for Computational Linguistics.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Bond, F. and Foster, R. (2013). Linking and extending an open multilingual wordnet. In *ACL (1)*, pages 1352–1362.

Brown, R. and McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of verbal learning and verbal behavior*, 5(4):325–337.

Burgess, J. A. (2007). When is circularity in definitions benign? *The Philosophical Quarterly*, 58(231):214–233.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., H. Jr, E. R., and Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. In *AAAI*, volume 5, page 3.

Devereux, B. J., Tyler, L. K., Geertzen, J., and Randall, B. (2014). The cslb concept property norms. *Behavior research methods*, 46(4):1119–1127.

Dumais, S. T. (2004). Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230.

Fillmore, C. J. (1977). Scenes-and-frames semantics. *Linguistic structures processing*, 59:55–88.

Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought*. MIT press.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proc. of ACL*, pages 873–882.

Iacobacci, I., Pilehvar, M. T., and Navigli, R. (2015). Sensembed: learning sense embeddings for word and relational similarity. In *Proceedings of ACL*, pages 95–105.

Klavans, J. L., Popper, S., and Passonneau, R. (2003). Tackling the internet glossary glut: Automatic extraction and evaluation of genus phrases. In *Proceedings of Semantic Web Workshop, SIGIR*.

Leone, V., Siragusa, G., Di Caro, L., and Navigli, R. (2020). Building semantic grams of human knowledge. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2991–3000.

McNamara, D. S., Crossley, S. A., and McCarthy, P. M. (2010). Linguistic features of writing quality. *Written communication*, 27(1):57–86.

McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behav. r. m.*, 37(4):547–559.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Moerdijk, F., Tiberius, C., and Niestadt, J. (2008). Accessing the anw dictionary. In *Proc. of the workshop on Cognitive Aspects of the Lexicon*, pages 18–24.

Muresan, S. and Klavans, J. (2002). A method for automatically building and evaluating dictionary resources. In *LREC*, volume 2, pages 231–234.

Navigli, R. and Ponzetto, S. P. (2010). Babelnet: Building a very large multilingual semantic network. In *Proc. of ACL*, pages 216–225. Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–43.

Silberer, C., Ferrari, V., and Lapata, M. (2013). Models of semantic representation with visual attributes. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 572–582.

Speer, R., Chin, J., and Havasi, C. (2016). Conceptnet 5.5: An open multilingual graph of general knowledge. *arXiv preprint arXiv:1612.03975*.

Speer, R. and Havasi, C. (2012). Representing general relational knowledge in ConceptNet 5. In *LREC*, pages 3679–3686.

Strehlow, R. A. (1983). Terminology and the well-formed

definition. In *Standardization of Technical Terminology: Principles and Practices*. ASTM International.

Thorat, S. and Choudhari, V. (2016). Implementing a reverse dictionary, based on word definitions, using a node-graph architecture. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2797–2806.

Witte, S. P. and Faigley, L. (1981). Coherence, cohesion, and writing quality. *College composition and communication*, 32(2):189–204.