CONCRETEXT @ EVALITA2020: The Concreteness in Context Task

Lorenzo Gregori
University of Florence
lorenzo.gregori@unifi.it

Maria Montefinese
University of Padua
maria.montefinese@unipd.it

Daniele P. Radicioni
University of Turin
daniele.radicioni@unito.it

Andrea Amelio Ravelli
Istituto di Linguistica Computazionale
"Antonio Zampolli" (ILC–CNR) - ItaliaNLP Lab
andreaamelio.ravelli@ilc.cnr.it

Rossella Varvara
University of Florence
rossella.varvara@unifi.it

Abstract

Focus of the CONCRETEXT task is conceptual concreteness: systems were solicited to compute a value expressing to what extent target concepts are concrete (i.e., more or less perceptually salient) within a given context of occurrence. To these ends, we have developed a new dataset which was annotated with concreteness ratings and used as gold standard in the evaluation of systems. Four teams participated in this first edition of the task, with a total of 15 runs submitted.

Interestingly, these works extend information on conceptual concreteness available in existing (non contextual) norms derived from human judgments with new knowledge from recently developed neural architectures, in much the same multidisciplinary spirit whereby the CONCRETEXT task was organized.

1 Introduction

Concept concreteness – that is, how directly a concept is related to sensorial experience (Brysbaert et al., 2014a)– is a fundamental dimension of conceptual semantic representation that has attracted more and more interest and attention in psycholinguistics in the last decade. This dimension is usually assessed by participants ratings on a Likert scale: concrete concepts lie herein on one side of the scale and refer to something that exists in reality and can be experienced immediately through the senses; abstract concepts lie on the opposite side of the scale and are grounded in the internal sensory experience and linguistic information. While concrete concepts have direct sensory referents (Crutch and Warrington, 2005) and greater availability of contextual information (Connell et al., 2018; Kousta et al., 2011; Montefinese et al., 2020), abstract concepts tend to be more emotionally valenced (Kousta et al., 2011) and less imageable (Montefinese et al., 2020; Garbarini et al., 2020).

The CONCRETEXT task challenges participants to build NLP systems to automatically assign a concreteness value to words in context. It is aimed at investigating how the concreteness information affects sense selection: different from past research (Brysbaert et al., 2014b; Montefinese et al., 2014), we are interested in assessing the concreteness of concepts within the context of real sentences rather than in isolation. Additionally, the concreteness score is assumed to be a property of meanings rather than a property of word forms; thus, scoring the concreteness of a concept in context implicitly requires to individuate its underlying sense, by handling lexical phenomena such as polysemy and homonymy.

Ordinary experience suggests that concepts' concrete/abstract status can affect their semantic representation, and lexical access and processing: concrete meanings are acknowledged to be more quickly and easily delivered in human communication than abstract meanings (Bambini et al., 2014). Historically, it has been observed that concrete concepts are responded to more quickly than abstract concepts in lexical decision tasks (Bleasdale, 1987; Kroll and Merves, 1986), although more recent experiments have shown that abstract

concepts might have an advantage when other variables have been accounted for (Kousta et al., 2011). Concrete concepts are also easier to encode and retrieve than abstract concepts (Romani et al., 2008; Miller and Roodenrys, 2009), are easier to make associations with (de Groot, 1989), and are more thoroughly described in definition tasks (Sadoski et al., 1997). Moreover, it takes generally less time to comprehend a concrete sentence than an abstract one (Haberlandt and Graesser, 1985; Schwanenflugel and Shoben, 1983). Thus, it has been proposed that different organizational principles govern semantic representations of concrete and abstract concepts: concrete concepts are predominantly organized by featural similarity measures, and abstract concepts by associative relations, co-occurrence patterns and syntactic information (Vigliocco et al., 2009).

All surveyed features make aspects ingrained in the distinction between concreteness/abstractness a stimulating and challenging field also for computational linguistics. Among the earliest attempts at grasping concreteness, we find works that investigated on concreteness/abstractness information in its interplay with metaphor identification and figurative language more in general (Turney et al., 2011) (and, more recently (Mensa et al., 2018b)). Although concreteness information is acknowledged to be central to, e.g., word-sense induction and compositionality modeling (Hill et al., 2013), the contribution of concreteness/abstractness to semantic representations is not fully grasped and exploited in existing approaches and resources, with the notable exception of works aimed *i)* at learning multimodal embeddings, and how abstract and concrete representations can be acquired by multi-modal models (Hill and Korhonen, 2014); and *ii)* at exploring in how far concreteness information is represented in the distributional patterns in *corpora* (Hill et al., 2013). Moreover, some approaches exist that attempted to create lexical resources by also employing common-sense information (Mensa et al., 2018a; Colla et al., 2018).

Characterizing tokens within sentences with their concreteness requires integrating both word-specific and contextual information. In our view, the CONCRETEXT Task entails dealing with a relaxed form of word sense disambiguation; such aspects were faced by our participants by devising methods relying on both traditional knowledge-based approaches, and more recent language models and sequence-to-sequence models. Finally, like in many real-world cases, the provided trial data is rather scarce, in the order of hundred sentences for the Italian language, and as many for English. This aspect forced our participants to face something similar to a 'cold start' problem. We hope that this edition of the CONCRETEXT task will be the first appointment in a series for those who are interested in the issues posed by the contextual conceptual concreteness to research on natural language semantics.

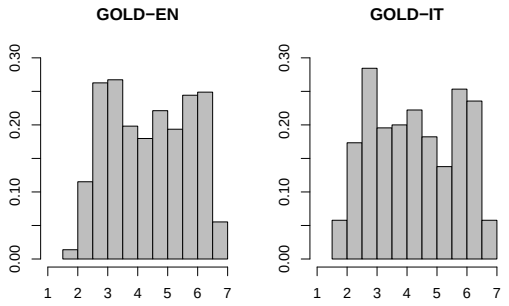
2 Task Definition

The task CONCRETEXT (so dubbed after CONcreteness in conTEXT) focuses on automatic concreteness (and conversely, abstractness) recognition. Given a sentence along with a target word, we asked participants to propose a system able to assess the concreteness of a concept expressed by a given word within a sentence, on a 7-point Likert-like scale where 1 stands for completely abstract (e.g., 'freedom') and 7 for completely concrete (e.g., 'car'). For example, in the sentence "In summer, wheat *fields* are coloured in yellow" the noun *field* refers to an entity that can smell, be touched, and pointed to. In this case, in a scale ranging from 1 to 7 its concreteness may be evaluated as 7, because it refers to an extremely concrete concept. In contrast, the same noun *field* in the sentence "Physics is Alice's research *field*" refers to a scientific subject, i.e., something that cannot be perceived through the five senses, but that can be explained through a linguistic description. In this sentence, the noun *field* may be evaluated 1 because it refers to an extremely abstract concept. Moreover, the task targets can be halfway between completely abstract and completely concrete, as in the case of "Magnetic *field* attracts iron", where the noun *field* refers to something more abstract compared to "wheat *fields*" but more concrete compared to "research *field*". As anticipated, the concreteness score being assigned to the word should be evaluated in context: the word should not be considered in isolation, but as part of a given sentence.

Participants were invited to exploit all possible strategies to solve the task, including (but not limited to) knowledge bases, external training data, word embeddings, etc.

Table 1: Basic statistics on the CONCRETEXT dataset used as gold standard.

	Italian	English
Unique Verb targets	52	44
Unique Noun targets	96	73
Num. Sentences	550	534
Num. Sentences Verb target	189	210
Num. Sentences Noun target	361	324
Avg. sent. length	14.43	14.33
Avg. sent. length (no punct)	13.03	12.87
Avg. full words per sent.	7.14	7.15
Num. Annotators	333	310
Human ratings (HR)	18,726	16,522
Min HR per sentence	30	30



(a) English dataset. (b) Italian dataset.

Figure 1: Distribution of human ratings for the English and Italian datasets.

3 Dataset

The dataset used for this task has been taken from the English-Italian parallel section of The Human Instruction Dataset (Chocron and Pareti, 2018), derived from WikiHow instructions.[1] All such documents had been anonymized beforehand, so that downloaded data present no privacy nor data sensitivity issues.

The dataset is composed of overall $1,096$ sentences, arranged as follows: 562 Italian sentences plus 534 English sentences. Each sentence contains a target term (either verb or noun) with its associated concreteness score (1–7 scale). Such score is derived from the average of at least 30 human judgments from native Italian and English speakers about the concreteness of a target word in a given sentence (see Table 1 for the dataset numbers).

The reliability of the collected data within each language (Italian, English) for the trial and test phases was evaluated separately by applying the split-half correlations corrected with the Spearman-Brown formula after randomly dividing the participants into two subgroups of equal size. All the reliability indexes were calculated on $10,000$ different randomizations of the participants. The mean correlations between the two groups are very high for both the trial and test phases, ranging from a minimum of $r = 0.87$ for English (at the test phase) to a maximum of $r = 0.98$ for Italian (at the trial phase), showing that the resulting ratings are highly reliable and can be used across the entire Italian – and English – speaking populations.

The dataset has been split into trial and test data, with a 20–80 ratio. Trial data has been released with the concreteness scores, while the test data has been provided at the beginning of the evaluation window without any score.[2]

4 Evaluation Measures and Baselines

We chose the Spearman correlation indices as our main evaluation measure; for the sake of completeness, we also report Pearson indices (substantially in accord with the previous metrics). We chose the former measure because the collected ratings are not normally distributed, which makes the Spearman correlation more suited to the data. In fact, by running the Shapiro–Wilk test we obtained a p-value < 0.001. The non normal distribution of data is also confirmed by the plot of the gold standard ratings, as illustrated in Figure 1.

Two baselines have been designed for this task.

Baseline One. The first baseline for the Italian language is derived as follows. The fastText word embeddings have been acquired beforehand by training the model on the Italian dump of the WikiHow instructions. We chose fastText for its support to the handling of OOV terms (Bojanowski et al., 2017), which is a crucial feature in the present setting. The cited norms by Montefinese et al. (2014) (referred to as 'the norms' hereafter) have been used herein. The average score of terms in each input sentence $S = \{t_1, t_2, \ldots t_K\}$ has been

[1] The whole Human Instruction Dataset dataset is freely available on Kaggle, https://www.kaggle.com/paolop/human-instructions-multilingual-wikihow

[2] The dataset employed in the CONCRETEXT task is available at the URL https://lablita.github.io/CONcreTEXT/.

computed by scrolling through the content words of the sentence. Each term t is searched in the norms: if the term is found, the associated concreteness score $c(t)$ is returned; otherwise, if the term is not present in the norms, the ranking of the l ($l = 20,000$) elements most similar to t is generated through fastText. In this case, we scan the whole norms list and employ the concreteness score of the element in the norms closest to those in the fastText ranking. In either case we obtain a score for each and every term in the input sentence, so that the concreteness score of the target token $\hat{t}$ is computed as the averaged score of the terms in the input sentence:

$$c(\hat{t}) = \frac{1}{K} \cdot \sum_{i=1}^{K} c(t_i).$$

The first baseline for the English language is analogous to the Italian one, except for the fact that the English tokens from the norms are accessed in this case. The same strategy governs the handling of the fastText resource, that in this case has been trained on the English dump of the Human Instruction Dataset.

Baseline Two. The second baseline for the Italian language implements a simple lookup function. More specifically, input sentences have been translated into English through the Google Translate ajax API implementation, and then the concreteness scores associated to the terms in the norms by Brysbaert et al. (2014b) are retrieved (in the unlikely case the term is not found, it is dropped, thus not contributing to the final score). The concreteness score of the target term is thus assigned to the average concreteness of terms in the given input sentence. The baseline two for the English language employs the concreteness score —by also employing the norms by Brysbaert et al. (2014b)— associated to all terms in the input sentence, finally assigning to the target token the average concreteness score for the whole sentence.

5 Systems Descriptions

In this Section we briefly describe the systems that participated in the competition. As a first edition, the CONCRETEXT task recorded a good feedback from the community, with 4 teams, overall 7 participants and 15 submitted system runs. In the next Section we report the results obtained by all such systems, while anonymizing a withdrawn participant.

5.1 ANDI

The ANDI team (Rotaru, 2020) proposed a system based on multiple classes of concreteness score predictors. The first class of predictors has been derived from large datasets of behavioral norms, collected for a wide variety of psycholinguistic factors. Beside well known concreteness norms, ANDI takes into account also semantic diversity, age of acquisition, emotional and sensori-motor dimensions, as well as frequency and contextual diversity counts. The vocabulary resulting from the merging of these words collections comprises more than 70K words, and it is the base vocabulary used to extract all the predictors. The second class of predictors has been derived from context-independent distributional models, namely Skip-gram, GloVe, and NumberBatch embeddings, as well as from the concatenation of the three. The third class of predictors has been derived from features obtained through recent transformers models, i.e. context-dependent representations. The models exploited are: BERT, GPT-2, Bart, and ALBERT. The final rating has been computed through a ridge regression over the three classes.

5.2 CAPISCO

The CAPISCO Team (Bondielli et al., 2020) submitted 3 systems for both Italian and English.

NON-CAPISCO. The first system computes a variation of the Baseline Two; that is, the target concreteness is obtained by combining the concreteness value of the target term (taken in isolation), and the average concreteness of the whole sentence. Improvement from baseline comes from considering differently the weight of the concreteness of the target term and of the context.

CAPISCO-CENTROIDS. This system is based on the assumption that close semantic spaces are featured by similar concreteness scores. In this case the authors first build two centroids, one for concrete and one for abstract concepts based on the norms by Brysbaert et al. (2014b) and Della Rosa et al. (2010), by employing fastText pre-trained embeddings. The concreteness score of a term is then computed by averaging the distance of the first 50 lexical substitutes of the target (identified through BERT) from the two polarized centroids. Introducing a list of target substitutes in a given context is thus the gist of this approach.

CAPISCO-TRANSFORMERS. In this variant, the CAPISCO team fine-tuned a pre-trained BERT model on the concreteness rating task, by complementing the CONCRETEXT training data with newly generated training data. The new data generation is twofold: for each original sentence, new sentences are generated by replacing the target term with the first lexical substitutes derived with BERT target masking approach. Then, more sentences are borrowed from Italian and English reference corpora.

5.3 KONKRETIKA

The KONKRETIKA team (Badryzlova, 2020) presented a system that first assigns a concreteness and an abstractness score to the target lemma, and then it adjusts these values based on the surrounding context. In the first step, the system computes semantic similarity between the target vectors and a "seed list" consisting of abstract and concrete words (extracted from the MRC Psycholinguistic Database). In the second step, the values where adjusted to the sentential context considering the mean concreteness index of the entire sentence. The team submitted 4 runs based on a heuristically selected coefficient.

6 Results

Four teams participated in the CONCRETEXT competition: ANDI, CAPISCO, KONKRETIKA, and a withdrawn team. ANDI and CAPISCO developed a system for both languages (English and Italian), while KONKRETIKA participated in the English track only, and the same did the withdrawn participant. Each team was allowed to submit the output of up to 4 system runs; the final ranking has been compiled based on the results of the best run.

In Tables 2 and 3 we present the score of each run for the English and Italian language, respectively. Although, as mentioned, the Spearman indices were adopted as our main evaluation metrics, we also report Pearson correlation indices and Euclidean distance, that may be useful to complete the assessment of the results. The final ranking is provided in Tables 4 and 5.

We can observe a substantial agreement between Spearman and Pearson indices: the averaged delta between such figures amounts to 0.012 and to 0.008 on the English and Italian dataset, respectively. Also the Euclidean distance seems to substantially confirm the results: for the results on English (Table 2) it is minimal for the output of the ANDI system, and it increases while Spearman correlation values decrease. The same trend is also confirmed on Italian results (Table 3).

Table 2: Results for each run on English test set.

System run	Spear	Pears	Eucl.D
ANDI	**0.833**	**0.834**	**15.409**
NON-CAPISCO	0.785	0.787	35.663
KONKRETIKA_3	0.663	0.668	28.613
KONKRETIKA_1	0.651	0.667	29.933
Baseline_2	0.554	0.567	38.451
KONKRETIKA_4	0.542	0.545	29.836
CAPISCO_CENTR	0.542	0.538	48.864
KONKRETIKA_2	0.541	0.545	30.322
CAPISCO_TRANS	0.504	0.501	29.927
Baseline_1	0.382	0.377	31.738
withdrawn_run3	-0.013	0.067	41.109
withdrawn_run1	-0.124	-0.123	44.068
withdrawn_run2	-0.127	-0.129	43.890

Table 3: Results for each run on Italian test set.

System run	Spear	Pears	Eucl.D
ANDI	**0.749**	**0.749**	**19.950**
CAPISCO_TRANS	0.625	0.617	24.367
CAPISCO_CENTR	0.615	0.609	28.608
NON-CAPISCO	0.557	0.557	31.588
Baseline_2	0.534	0.522	40.114
Baseline_1	0.346	0.368	31.046

Tables 6 and 7 report disaggregated Spearman correlations for verbs and nouns. This allows to highlight if and to what extent the participating systems obtained better results on either POS. ANDI obtained the best results on both verbs and nouns in both languages. This system (and NON-CAPISCO as well) obtained analogous results on verbs and nouns. On the whole, the rest of the systems obtained results clearly better on English verbs and slightly better on Italian nouns. In particular, KONKRETIKA (English only) is strongly biased on verbs: its performances on verbs are higher in all 4 runs. CAPISCO systems exhibit the most varied behavior.

7 Discussion

The obtained results confirm transformers as a good device to compute concreteness score for words in context. The virtues of transformers in grasping contextual information are largely

Table 4: Final ranking on English test set.

Team	Spear	Pears	Eucl.D
ANDI	**0.833**	**0.834**	**15.409**
CAPISCO	0.785	0.787	35.663
KONKRETIKA	0.663	0.668	28.613
withdrawn	-0.013	0.067	41.109

Table 5: Final ranking on Italian test set.

Team	Spear	Pears	Eucl.D
ANDI	**0.749**	**0.749**	**19.950**
CAPISCO	0.625	0.617	24.367

Table 6: Spearman rank differences between nouns and verbs on English test set.

	Spear.N	Spear.V	Diff
CAPISCO_TRANS	0.443	0.654	0.211
KONKRETIKA_4	0.502	0.701	0.199
KONKRETIKA_2	0.502	0.683	0.181
CAPISCO_CENTR	0.478	0.659	0.181
KONKRETIKA_3	0.629	0.762	0.133
KONKRETIKA_1	0.611	0.741	0.13
ANDI	0.836	0.857	0.021
NON-CAPISCO	0.779	0.782	0.003

Table 7: Spearman rank differences between nouns and verbs on Italian test set.

	Spear.N	Spear.V	Diff
NON-CAPISCO	0.579	0.507	0.072
CAPISCO_TRANS	0.607	0.667	0.060
CAPISCO_CENTR	0.625	0.591	0.034
ANDI	0.762	0.749	0.013

known, but in the present setting we observe that their output can be further improved by integrating behavioral information (this seems to be one major difference between the systems ANDI and CAPISCO-TRANSFORMERS).

The most important output of this challenge is definitely the great performance of the ANDI system, that proves to be robust and reliable for the considered task: the system obtains the best ranking in both languages, a low deviation from the gold standard and a substantial stability in processing both verbs and nouns. Moreover, the proposed system is ready to be applied in a multi-language environment, given that non-English sentences are automatically translated into English. The ANDI system exploits different kinds of available resources and works with local and contextual information. This shows that deriving the concreteness score of a word in context is a complex task, involving different semantic, cognitive and experiential levels.

The high correlation obtained by the NON-CAPISCO in the English task is somehow surprising, since this system makes use only of the mean concreteness of the sentence (computed from existing norms) as contextual information. This result is thus related to the availability of existing norms, but it shows that there is a link between the concreteness score of a target word in context and the concreteness scores of the words it occurs with. Further analysis are needed, but it suggests that concrete interpretations of a target word are associated with concrete context words. Of course, systems based exclusively on behavioral norms are strongly dependent on the coverage of the considered vocabulary. In fact, the NON-CAPISCO Italian performances (obtained exploiting a $\sim$ 1.2K vocabulary) are lower than all the other systems, while on the English track it ranks second (using a $\sim$ 70K vocabulary).

8 Conclusions

We presented the results of the CONCRETEXT task at EVALITA 2020 (Basile et al., 2020). The task challenges participants to build NLP systems to automatically assign a concreteness score to words in context, evaluating to what extent target concepts are concrete (i.e., more or less perceptually salient) within a given context of occurrence. A novel dataset was developed for this task as a multilingual comparable *corpus* composed of 550 Italian sentences and 534 English sentences, annotated with the concreteness/abstractness rating of target nouns and verbs. Three teams completed their participation to the task, obtaining the following ranking: ANDI (Rotaru, 2020), CAPISCO (Bondielli et al., 2020), and KONKRETIKA (Badryzlova, 2020).

Future work will address the following steps. First of all, we will improve our dataset by including further languages, also from different language families and under-resourced languages. Also the set of considered targets should be expanded, to ensure a broader coverage to the dataset, and more significant results (thanks to the larger experimental base) to its future users as well.

References

Yulia Badryzlova. 2020. KONKRETIKA @ CONCRETEXT: Computing concreteness indexes with sigmoid transformation and adjustment for context. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Valentina Bambini, Donatella Resta, and Mirko Grimaldi. 2014. A dataset of metaphors from the italian literature: Exploring psycholinguistic variables and the role of context. *PloS one*, 9(9):e105634.

Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro. 2020. Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

Fraser A Bleasdale. 1987. Concreteness-dependent associative priming: Separate lexical organization for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(4):582.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information.

Alessandro Bondielli, Gianluca E. Lebani, Lucia C. Passaro, and Alessandro Lenci. 2020. CAPISCO @ CONCRETEXT: (Un)supervised Systems to Contextualize Concreteness with Norming Data. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Marc Brysbaert, Michaël Stevens, Simon De Deyne, Wouter Voorspoels, and Gert Storms. 2014a. Norms of age of acquisition and concreteness for 30,000 dutch words. *Acta psychologica*, 150:80–84.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014b. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

Paula Chocron and Paolo Pareti. 2018. Vocabulary alignment for collaborative agents: a study with real-world multilingual how-to instructions. In *IJCAI*, pages 159–165.

D. Colla, E. Mensa, A. Porporato, and D.P. Radicioni. 2018. Conceptual Abstractness: From Nouns to Verbs. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2253. CEUR.

Louise Connell, Dermot Lynott, and Briony Banks. 2018. Interoception: the forgotten modality in perceptual grounding of abstract and concrete concepts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1752):20170143.

Sebastian J Crutch and Elizabeth K Warrington. 2005. Abstract and concrete concepts have structurally different representational frameworks. *Brain*, 128(3):615–627.

Annette M de Groot. 1989. Representational aspects of word imageability and word frequency as assessed through word association. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(5):824.

Pasquale A Della Rosa, Eleonora Catricalà, Gabriella Vigliocco, and Stefano F Cappa. 2010. Beyond the abstract—concrete dichotomy: Mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 italian words. *Behavior research methods*, 42(4):1042–1048.

Francesca Garbarini, Fabrizio Calzavarini, Matteo Diano, Monica Biggio, Carola Barbero, Daniele P Radicioni, Giuliano Geminiani, Katiuscia Sacco, and Diego Marconi. 2020. Imageability effect on the functional brain activity during a naming to definition task. *Neuropsychologia*, 137:107275.

Karl F Haberlandt and Arthur C Graesser. 1985. Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology: General*, 114(3):357.

Felix Hill and Anna Korhonen. 2014. Learning abstract concept embeddings from multi-modal data: Since you probably can't see what i mean. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 255–265.

Felix Hill, Douwe Kiela, and Anna Korhonen. 2013. Concreteness and corpora: A theoretical and practical study. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 75–83.

Stavroula-Thaleia Kousta, Gabriella Vigliocco, David P Vinson, Mark Andrews, and Elena Del Campo. 2011. The representation of abstract words: why emotion matters. *Journal of Experimental Psychology: General*, 140(1):14.

Judith F Kroll and Jill S Merves. 1986. Lexical access for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12(1):92.

Enrico Mensa, Aureliano Porporato, and Daniele P. Radicioni. 2018a. Annotating concept abstractness by common-sense knowledge. In Chiara Ghidini, Bernardo Magnini, Andrea Passerini, and Paolo

Traverso, editors, *AI*IA 2018 – Advances in Artificial Intelligence*, pages 415–428, Cham. Springer International Publishing.

Enrico Mensa, Aureliano Porporato, and Daniele P. Radicioni. 2018b. Grasping metaphors: Lexical semantics in metaphor analysis. In Aldo Gangemi, Anna Lisa Gentile, Andrea Giovanni Nuzzolese, Sebastian Rudolph, Maria Maleshkova, Heiko Paulheim, Jeff Z Pan, and Mehwish Alam, editors, *The Semantic Web: ESWC 2018 Satellite Events*, pages 192–195, Cham. Springer International Publishing.

Leonie M Miller and Steven Roodenrys. 2009. The interaction of word frequency and concreteness in immediate serial recall. *Memory & Cognition*, 37(6):850–865.

Maria Montefinese, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2014. The adaptation of the affective norms for english words (anew) for italian. *Behavior research methods*, 46(3):887–903.

Maria Montefinese, Ettore Ambrosini, Antonino Visalli, and David Vinson. 2020. Catching the intangible: a role for emotion? *Behavioral and Brain Sciences*, 43.

Cristina Romani, Sheila Mcalpine, and Randi C Martin. 2008. Concreteness effects in different tasks: Implications for models of short-term memory. *Quarterly Journal of Experimental Psychology*, 61(2):292–323.

Armand Rotaru. 2020. ANDI @ CONCRETEXT: Predicting concreteness in context for English and Italian using distributional models and behavioural norms. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Mark Sadoski, William A Kealy, Ernest T Goetz, and Allan Paivio. 1997. Concreteness and imagery effects in the written composition of definitions. *Journal of Educational Psychology*, 89(3):518.

Paula J Schwanenflugel and Edward J Shoben. 1983. Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(1):82.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690.

Gabriella Vigliocco, Lotte Meteyard, Mark Andrews, and Stavroula Kousta. 2009. Toward a theory of semantic representation. *Language and Cognition*, 1(2):219–247.