A Set of Semantic Norms for German and Italian

Gerhard Kremer
CIMeC, University of Trento
Corso Bettini 31
Rovereto, Italy
+39 0464 80 8619
gerhard.kremer@unitn.it

Marco Baroni
CIMeC, University of Trento
Corso Bettini 31
Rovereto, Italy
+39 0464 80 8612
marco.baroni@unitn.it

November 2, 2010

The psychological community frequently investigates semantic norms of properties produced by native speakers after being presented concept words; and these norms are of great value for a wide variety of psychological experiments. This paper presents a new set of norms that includes a collection of properties from a production experiment for the German and the Italian language. Stimuli consisted of 50 concrete objects taken from 10 different concept classes. The data comprises annotations of semantic relation types and several statistical measures, which facilitate the comparison of the two target languages.

Introduction

Semantic features play a central role in studies investigating the mental representation and processing of word meanings, especially in semantic theories about concepts and their categorisation (e. g., Medin & Schaffer, 1978), where semantic features are used as the basis for constructing conceptual representations (see Murdock, 1982).

Typically, researchers aiming to elaborate specific theories in this area empirically collect semantic features through an experimental approach in which participants are presented with a set of concepts and asked to produce features that they think would best describe each of the concepts. The acquired data undergo statistical distribution analyses, and additional measures not based solely on the data collection itself complement the semantic features description. These semantic norms allow researchers to test theories about semantic memory, to construct stimuli for further experiments (while controlling for various variables based on the created measures), and to model human behaviour in computational simulation models.

It is important to understand the capabilities and limits of feature norms. For a fuller discussion see McRae, Cree, Seidenberg, and McNorgan (2005). Feature norms provide valuable information about memory not because there is evidence that semantic knowledge is represented in the brain as a set of verbalisable features, but because semantic representations are used systematically by participants when generating features. Barsalou (2003) assumes that, when generating features, participants simulate a holistic representation of the target category and then interpret this simulation by using featural and relation simulators. Thus, the participant's list of features is a temporary abstraction constructed online, so that the dynamic nature of the feature generation results in substantial variability within and across participants. So, in order to derive a single, averaged representation, responses should be pooled. One limitation of feature norms are that they are linguistically based (participant responses are collected in written or verbal form), and thus some types of information can be transmitted more easily and with more detail than other types of information. For example, that a door is used by people is easier to verbalise than information about where the door handle is attached and how big it is. As a second example, although animals can be recognised by the way they move, the particular movements are hard to verbalise (although for some animals a distinguishing, general movement can be given, e. g., "a frog jumps"). As a consequence, such details are left out by participants and do not appear in the norms. Furthermore, McRae et al. (2005) state that feature norms are biased towards information that distinguishes concepts from each other, either because participants understand this to be the implicit task or because this type of information is actually salient to them. Only few features are listed that are true for a large numbers of concepts. McRae et al. (2005) see this as a strength as general features play only a small role in object identification, language comprehension, and language production.

As more thoroughly reviewed in McRae et al. (2005), research making use of semantic norms include, among many others, Rosch and Mervis (1975) exploring typicality gradients and Ashcraft (1978b) constructing feature verification experiments. Hampton (1979) collected features to test the model of category verification by Smith, Shoben, and Rips (1974) and to predict verification latencies. Wu and Barsalou (2009) used feature norms for the comparison of predictions of a theory involving perceptual symbol systems and one based on amodal semantics. Garrard, Lambon Ralph, Hodges, and Patterson (2001) investigated category-specific semantic deficits, using their norms. Vinson and Vigliocco (2002) used a collection of norms to compare nouns versus verbs in a series of experimental paradigms. Moss, Tyler, and Devlin (2002) used their norms to derive representations for implemented computational models.

Feature norms and derived concept representations have served as the basis for accounts of a number of empirical phenomena, such as semantic similarity priming (e. g., see Cree, McRae, & McNorgan, 1999), feature verification (Ashcraft, 1978a), categorisation (Smith et al., 1974), and conceptual combination (Hampton, 1979). Additionally, they have been used to support modality-specific aspects of representation (Solomon & Barsalou, 2001).

As described above, the research community depends on semantic norms for a multitude of purposes. However, only a few research groups made the norms they collected publicly available (Vinson & Vigliocco, 2008; Garrard et al., 2001; McRae et al., 2005). The data produced by participants is published along with statistical data from analyses regarding psycholinguistic variables including, e. g., familiarity, typicality, production frequency, which are augmented by measures requiring additional sources, such as occurrence frequencies from text corpora and association strength based on these frequencies.

This paper describes a semantic norms collection for 50 concrete concepts from 10 different concept classes. These parallel norms were acquired from native speakers of German and Italian, using a property generation task similar to the one of McRae et al. (2005), and under very similar settings across the two languages. We were moreover careful to follow the transcription and labelling methods of McRae and colleagues very closely, using their norms as our "de facto standard". In this way, the norms are not only highly comparable between German and Italian, but also quite comparable to the McRae English norms. Our data are published and can be accessed online from the Behaviour Research Methods website at `http://www.psychonomic.org/archive`. They are described in section A in the appendix.

The current paper has two purposes. First, we introduce the norms as a resource that, despite its small size, we hope will be useful to the research community. As far as we know, ours are the first publicly available norms for German and Italian (indeed, for any language other than English). As such, the norms, together with the supplementary information we provide, should be useful to researchers working with these languages or interested in cross-linguistic comparisons (also with English).

Second, we present a systematic comparison of our German and Italian data with each other as well as with the related McRae English norms, in order to investigate an important issue that has been somewhat overlooked in the relevant literature, namely to what extent the norms reflect universal (or at least, culturally dependent) properties of concepts that are stable across languages, and to what extent they are instead language-specific. We can only provide a partial answer to this question, given the small number of languages and concepts analysed. As we will see, the data suggest that concept descriptions are in general stable, but language-specific effects are also present.

Method

Participants

Participants were native speakers of the respective target language (German or Italian) attending high school in Bolzano, the capital of South Tyrol, a region in Italy where two groups of native language speakers of Italian and German live together; the two groups are taught the respective other language in intensive foreign language learning courses in schools, where their native language is used in general as teaching language.

We emphasise that inhabitants in this region – at least in the larger urban areas – are generally not bilinguals (which otherwise could be used as an argument to explain emerging similarities in the data results between the two target languages), while they have roughly comparable socio-economic and cultural conditions. Thus, the region is ideal for studying differences due to purely linguistic factors between highly comparable groups.

The current school system promotes contacts within the same language group and discourages contacts with the respective non-native language group, favouring the parallel existence of the two language groups (cf. Forer, Paladino, Vettori, & Abel, 2008). Although there are efforts to socialise these separate groups with each other, appropriate initiatives started only in the last few years. Thus, researchers looking for bilingual speakers must choose participants from smaller cities – and thoroughly verify that they are bilinguals, e. g., by admitting only those whose parents have different mother tongues and who speak both languages at home (see Guagnano, 2010). Several studies make statements about the difference between official bilingualism (a prerequisite for having a public administrative job position, evaluated with a language proficiency test that is passed, on average, by around 50% of the applicants[1]) and the real conditions of the area, namely that ethnolinguistic groups live side by side with only little mutual integration or sociolinguistic contact (see, e. g., Dal Negro, 2005). This view conforms with the opinions of the population itself.[2] Furthermore, the region's statistics institute conducts censuses in which inhabitants are required to declare whether they are German or Italian (or belonging to the small Ladin-speaking minority), acknowledging the rather monolingual reality.[3] A more detailed analysis about the reasons for the lack of a real bilingualism in South Tyrol, viewed from political-institutional, socio-educational, and social relations perspectives, was conducted by Cavagnoli and Nardin (1999).

Each participant in our survey had to fill in a form with information about his/her native language and the native languages of the parents (non-native and mixed background participants were excluded), as well as handedness, gender, and age. The age of the participants was in the range of 15 to 19 years. The average age was 16.7 (standard deviation 0.92) for the German participants and 16.8 (s.d. 0.70) for the Italian participants. Note that similar studies (including McRae's) typically involve older participants, such as university students. In total, 73 German students and 69 Italian students took part in the experiment.

Stimuli

The stimulus set was a collection of 50 concrete concepts from 10 different concept classes (see the table in appendix B). The English concept words were mainly taken from those used by McRae et al. (2005) and Garrard et al. (2001) in their experiments. They were chosen so that their translations into the target languages German and Italian had unambiguous and reasonably monosemic lexical realisations. These target words showed no significant differences in word length for either language. Analysing the corpus frequencies of the target words in German, Italian, and English corpora revealed significantly larger frequencies for words in the "body part" class (across languages) compared to the words in the other classes – It is not surprising that the words *eye*, *head*, and *hand* appear much more often than the other words in the set.

[1]See the brochure at `http://www.provincia.bz.it/astat/de/service/845.asp`
[2]Interviews analysed at `http://asus.sh/oberprantacher.239.0.html`
[3]See `http://www.provinz.bz.it/astat/de/themen/volkszaehlung-sprachgruppen.asp`

Procedure

The experiment was conducted class-wise in schools. Each participant was provided with a set of 25 concepts which were presented on separate sheets of paper. To get an equal number of participants describing each concept, for each participant pair the whole set of 50 concepts was randomised and split into 2 subsets. Thus, each participant saw a random subset of target stimuli in a random order (due to technical problems, the split was not always different across participant pairs). We could not present the whole set of 50 concepts to each participant because of the time limits requested by the schools for the experiment sessions.

Short instructions were provided orally before the experiment and were handed out to each participant in written form. To make the concept description task more natural for the participants and to get mainly those types of descriptions that we aimed at, we suggested that participants imagine a group of alien visitors and assume that each alien visitor knew the meaning of all words of the language except one particular word for a concrete object (the target stimulus) that had to be described.

The participants were instructed to enter one descriptive phrase per line and to try and write at least 4 phrases per target word. The time limit given was one minute per concept, and participants were not allowed to go back to a word they had previously described.

Before the experiment, an example concept (not included in the target set) was presented, and participants were encouraged to describe it and ask clarifications about the task.

Transcription and Labelling

The collected data comprised for each concept, on average, descriptions by 36 German participants (s.d. 1.25) and 34 Italian participants (s.d. 1.73).

The produced descriptions were digitally transcribed and manually checked to make sure that different properties were properly split into separate phrases. Where splitting was necessary, we tried to systematically apply the criterion that, if at least 1 participant produced 2 properties on separate lines, then the properties would always be split in the rest of the data set whenever they appeared in a single line.

Data were then transcribed into English and mapped (by the authors) to a standardised form. These operations were performed by keeping as close as possible to the procedure of McRae et al. (2005) and using their norms as our "annotation guidelines", in order to keep the data comparable between this project's target languages and the McRae's data. Mapping also involved leaving out habitual words (which just express the typicality of the concept description, e. g., "usually", "often", "most", "everybody" – giving typical properties is required implicitly in the task) and merging synonyms.

Translated and mapped phrases were labelled with their respective relation types while following McRae's criteria and using a subset of the semantic relation types described in Wu and Barsalou (2009) – see appendix C. While trying to adapt McRae's annotation style, we encountered dubious cases. For example, in their norms, "carnivore" is classified as a *category*, whereas "eats_meat" is classified as a *behaviour*. To us they seem to convey the same information, which is why we decided to map both to "eats_meat", classified as *behaviour*.

Apart from the semantic relation types described in Wu and Barsalou (2009), the additional semantic relations types we used for annotation comprise *material* (em), *role* (sr), and *episodic property* (iep).[4] Differently from the annotation scheme that McRae et al. (2005) applied, we separated the material something is made of from internal component relations (contrasting, e. g., "made of wood" and "has a leg" and splitting phrases like "has a wooden leg"). The *role* relation was introduced to more appropriately annotate descriptions like "pet" or "one's best friend". Some phrases produced could probably have been annotated best as *systemic property* (esys) in Wu and Barsalou's annotation scheme, but this relation is a quite openly defined relation type, so we decided to use the *episodic property* type (iep) for properties that cannot be directly perceived when encountering a concept (e. g., "is strong", that requires some kind of inference from perceptual data).

During transcription of the produced phrases into English and mapping onto standardised phrases, we observed structural language-dependent differences. For example, in German, expressions denoting a complex meaning (e. g., domesticated animal or pet) are often expressed by noun compositions ("Haustier"), whereas in Italian this would rather be expressed via a noun–adjective combination ("animale domestico"). Since in both languages "animal" was also used separately for other concepts (but not for the same concept), we assumed that such a complex expression was used to convey both parts of the meaning at once, which is why we assigned in this case two relation types: *category* ("an animal") and *role* ("used as pet"). Similarly, "means of transportation" (German: "Transportmittel", Italian: "mezzo di trasporto") was split into the relation types *category* ("vehicle") and *function* ("used for transportation"). In this case, though, the separate German word "Mittel" would not be used separately to adequately describe a vehicle (it has a more abstract meaning), whereas the Italian word "mezzo" can also be used as an ellipsis for expressing the same meaning as in the composed expression above. However, we believe that two meaningful aspects are conveyed here in both language groups, which is supported by the fact that many times German and Italian participants also produced both relation types using separate phrases when they described the same (vehicle) concept. There are also complex expressions that are harder to map to a common phrase, such as "Schwimmhäute" (German) and "piedi palmati" (Italian), both for "webbed feet", where the German expression only refers to the skin (between the fingers) that helps with swimming – some German participants stated explicitly, in addition, that this skin is on the feet. Here, it is hard to come up with a common and accurate mapped phrase. In such (few) cases, we did not attempt to capture the commonalities. Other possible language differences that might have lead to asymmetries in translation and mapping are

[4]We followed the coding scheme of Wu and Barsalou (2009), where the first letter of a type code denotes one of the following 5 general semantic relation types: entity properties (e), taxonomic categories (c), situation properties (s), introspective properties (i), and miscellaneous (m). The remaining letters in a type code denote the specific relation type. See appendix C for the full list of type codes we used in the annotation process.

alternative linguistic constructions to express one meaning, within and across languages (e. g., “quadrupede”, “4-beinig”, and “ha 4 gambe” all refer to the concept of having 4 legs, using a noun, an adjective and a verb phrase, respectively), or semantically similar words used for the same basic meaning (e. g., 4 “paws”/“feet”/“legs”). That is, even though one annotator was solely responsible for the whole German data set, one annotator for the Italian data set, and both tried to come up with a common annotation scheme by using the McRae data set and communicating possible difficult cases, it is likely that there are still inconsistencies in mapping to standardised phrases and mapping of relation types within and across languages.

To test the inter-coder reliability in mapping phrases to relation types, for each target language we asked another native speaker to label 100 randomly sampled standardised phrases and compared the agreement between their labels and the annotated labels in our data set (these secondary annotators were trained using phrases that were not included in the random sample). The agreement between our annotation and that of the secondary annotators was rather high, with kappa values (using Cohen’s kappa) of 0.844 for German and 0.676 for Italian. Cohen’s kappa provides an adjustment of the proportion of agreement for the chance agreement factor, i. e., it is corrected under consideration of the agreement that could already be achieved by chance. A value of 0 means that the obtained agreement is equal to chance agreement, a positive value means that the obtained agreement is higher than chance agreement, with a maximum value of 1 (see Cohen, 1960). Although the lack of consensus on how to interpret kappa values, the two values obtained above are commonly considered as showing a reasonably high agreement (cf. Artstein & Poesio, 2008).

The average number of mapped phrases obtained per participant for a concept is 5.49 (s.d. 1.82) for the German group and 4.96 (s.d. 1.86) for the Italian group. In total, the average number of phrases obtained for a concept is 200.2 (s.d. 25.72) for German and 170.4 (s.d. 25.46) for Italian.

Results and Discussion

Describing the data collected from the experiment, we focus in particular on investigating their cross-language properties, trying to assess to what extent verbally expressed concept descriptions are language-dependent, and to what extent they go beyond language-specific effects. The analysis focuses mostly on our German and Italian data, but we also compare the relation type distribution in our norms to the one attested, for the same concepts, in the English norms provided by McRae et al. (2005).

In total, the collected data amount to 10,010 properties produced by German participants (2,513 distinct properties, if we do not count those repeated across participants) and 8,520 properties produced by Italian participants (1,243 distinct properties). Although slightly more German participants took part in the experiment, it probably does not account for the whole difference in numbers of phrases produced in total and should be subject to future investigations (we have not found an explanation, yet). There were 187 German and 196 Italian concept-property pairs that were produced by at least ten

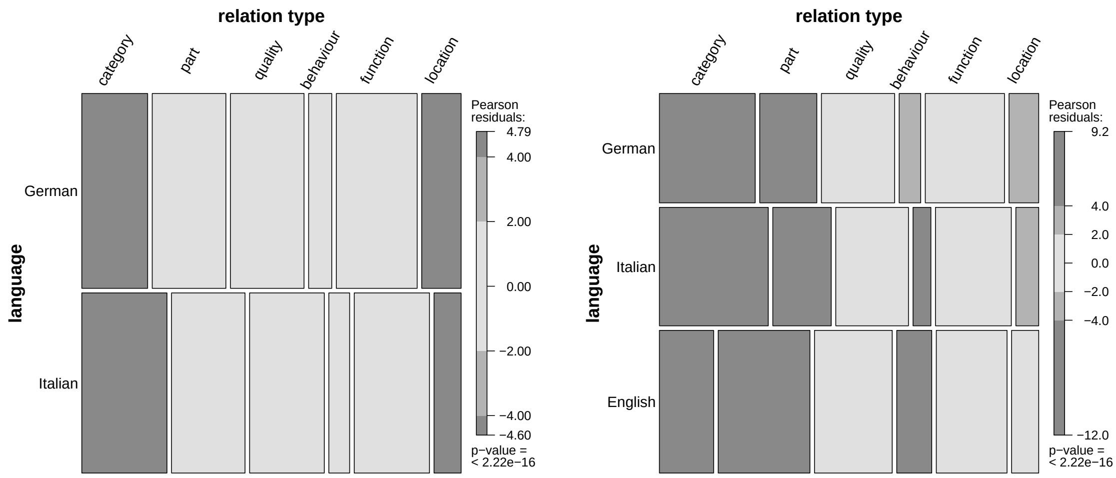

Figure 1: Overall frequency distribution of phrases of one of the 6 relation types that were annotated most frequently for each target language (left). Distributions compared to McRae et al.'s data (English) – including in all languages only phrases produced by at least 5 participants for a concept (right).

participants. Of those, 117 were shared across languages (i. e., 63 % in the German data and 60 % in the Italian data).

The number of properties grouped by the annotated relation types are presented in appendix C. The relation type codes (in the style of Wu and Barsalou) used in the annotation are explained there. The overall frequency distributions of the top 6 relation types are displayed in figure 1. The data subset including only these 6 relation types contains more than 68 % of the whole data set and comprises the relation types *category* (in the Wu/Barsalou coding: ch), *part* (ece), *quality* (ese), *behaviour* (eb), *function* (sf), and *location* (sl). The presented plot is generated via the R statistical computing environment[5], using the vcd package (Meyer, Zeileis, & Hornik, 2006). In this so-called mosaic plot, widths of the rectangles in a row depict the proportions of the total number of phrases produced and mapped to one of the 6 relation types (for the respective language). The height of the set of rectangles in a row represents the proportion of frequency of all relations (of the 6 relation types) produced in a language as compared to the language in the other row. That is, in German, phrases of the relation type *quality* were produced about three times as often as phrases of the relation type *behaviour*, and in total, about the same number of phrases of the top 6 relation types were produced for German and Italian. The grey shades in the mosaic plot code the significance degrees of the differences between the rectangles in a column (comparing the relative frequencies of phrases of a specific type between the two languages) according to a Pearson residual test (see Meyer et al., 2006, for details) – darker rectangles correspond to larger (and more significant)

[5]see `http://www.r-project.org`

deviances from the cross-language distribution.

Both the German and the Italian data had similar distributions, with significant differences only for *category* relations (which were produced less often by German participants than by Italian participants) and *location* relations (which were produced more often by the German participants).

For the difference in *location*, no clear pattern emerges from a qualitative analysis of German and Italian *location* properties. Regarding the difference in *category* relations, we find, interestingly, a small set of more or less abstract hypernyms that are frequently produced by Italians, but never by Germans: "object" (72), "construction" (36), "structure" (16). In these cases, the Italian translations have subtle shades of meaning that make them more likely to be used than their German counterparts. For example, the Italian word "oggetto" (English: "object") is used somewhat more concretely than the extremely abstract German word "Objekt" (or English *object*, for that matter) – in Italian, the word might carry more of an "artifact, man-made item" meaning. At the same time, "oggetto" is less colloquial than German "Sache", and thus more amenable to be entered in a written definition. The "vehicle" (*category*) was more frequent in the Italian than in the German data set. Differences of this sort remind us that property elicitation is first and foremost a verbal task, and as such it is constrained by language-specific usages. It is left to future research to test to what extent linguistic constraints also affect deeper conceptual representations (would Italians be faster than Germans at recognising superordinate properties of concepts when they are expressed non-verbally?).

The mosaic plot on the right in figure 1 shows the distribution of the same relation types for the English data set collected by McRae et al. (2005) in contrast to the data produced by German-speaking and Italian-speaking participants as described in this paper. For uniformity with the available English data, for this plot only relations produced by at least 5 participants for a concept were considered. To achieve the most accurate comparison possible, only concepts which were used both in the English and the German/Italian data sets were considered. For 4 concepts used for German and Italian that did not appear in the English data set, similar concepts were chosen from the English set – *couch*, *blouse*, *gorilla*, and *pyramid* substituted *armchair*, *chemise*, *monkey*, and *tower*, respectively. Furthermore, all concepts from the "body part" class were excluded because this concept class was not represented in the English data set. The most striking aspect of the relation type distribution in the English data set is the low relative number of *category* relations and the high relative number of *part* relations – which distinguishes this set both from the German and the Italian data. These differences might be due at least partially to the following fact. Whereas during the German/Italian data collection participants had a limited time (1 minute per concept, for 25 concepts), the participants in the English norms collection had unlimited time (taking around 40–50 minutes for 20–24 concepts). Having more time to contemplate, participants could come up with more descriptions about a concept's parts (concrete concepts tend to have many parts), whereas in most cases a concept is categorised only into one or two categories independently of time constraints. This time limit difference might also account for the higher total number of produced concept features in the English data set in comparison to the German and Italian sets, as depicted by the height of the rectangles in the plot. Apart from the

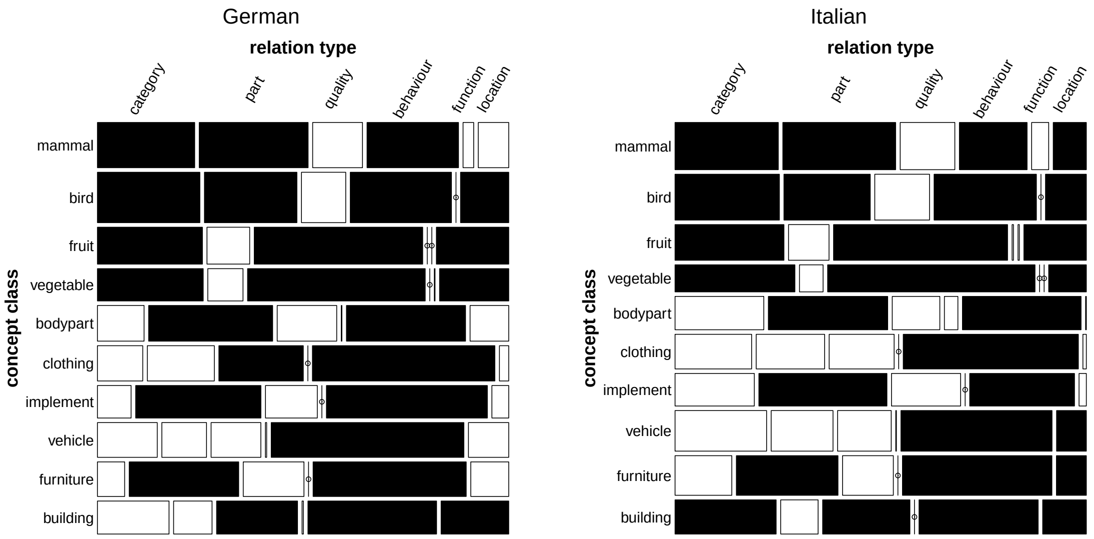

Figure 2: Frequency count deviations from the overall distribution of phrases of the 6 relation types considered for the German (left) and the Italian data (right). Black/white cells indicate over-/underrepresented counts; zero frequencies are indicated by circles.

differences in category and part relations, the relative distributions are roughly rather similar between the three languages.

We additionally investigated the differences between German, Italian and English when considering only the number of *distinct* features produced (participants of the different language groups might produce similar numbers of features for each relation type, but the variety of features used might differ across languages). The relative numbers of distinct features were not differing significantly for any of the 6 relation types analysed across languages. Counting the number of distinct concept–feature pairs, the only significant differences were for the relation type *category*, overrepresented in Italian and underrepresented in English. These additional analyses further stress the commonalities in concept descriptions across languages.

Next, relation type distributions for each of the concept classes are shown in separate mosaic plots for German and Italian (see figure 2). Here, a binary colour coding indicates overrepresented (black) and underrepresented (white) counts for a relation type within a particular concept class, compared to the overall distribution as seen in the left plot of figure 1. A relation type for a specific concept class is overrepresented/underrepresented if the sign of the Pearson residual is positive/negative, i. e., if the relative frequency of relations of that relation type and in that concept class is higher/lower than the relative frequency of phrases of that relation type across all concept classes. Comparing the two languages, we can observe that the proportions are roughly similar, i. e., the relation type of the widest and of the narrowest rectangles match across languages. Furthermore, some concept classes have similar distributions withing a language, most evidently “fruit”

Table 1: Tests of statistical significance of difference between German and Italian global concept measures

Measure	Test	p-value
No.Features	Wilcoxon	0.58
No.DistinguishingFeatures	Wilcoxon	0.66
Percent.DistinguishingFeatures	Student's T	0.84
MeanDistinctivenessAcrossFeatures	Student's T	0.78
MeanCueValidity	Student's T	0.19
No.IntercorrelatedFeatures	Wilcoxon	0.43
Percent.IntercorrelatedFeatures	Student's T	0.46
IntercorrelationalDensity	Student's T	0.16
Prod.Frequency	Wilcoxon	0.52
ConceptSimilarities	Wilcoxon	0.85

and "vegetables" in the German data, which makes sense given that they both can be subsumed under the broad class of "eatable plants"; other classes have markedly different distributions, e. g., compare "fruit" and "implements", where for "implements" a lot of relations of types *part* and *function* were produced in contrast to the "fruit" class, which in turn is characterised by a larger number of *category* and *quality* relations than in the "implement" class. Further research on this data set investigating the cognitive salience of semantic relations is presented in Kremer, Abel, and Baroni (2008).

In addition to the analyses based on relation types, we compared the German and Italian data with respect to various measures that are used in the literature to capture global properties of concept norm productions (and that we include in the data we make available), to see whether they were significantly different across the languages. Measures that were tested comprise, for each concept, the number of different features produced, the numbers of different distinguishing features produced, the percentage of distinguishing features compared to the number of all different features produced, the average distinctiveness across a concept's features, the average cue validity across a concept's features, the number of intercorrelated features, the percentage of intercorrelated features compared to all pairs of features, and the production frequency of a feature for a concept. Furthermore, concept similarities within concept classes were compared, using the cosine similarity values to rank the pairs of concepts within each concept class and compare those ranks between German and Italian. Please refer to appendix A for a more detailed description of these measures.

Two different tests were applied: In case of integer scores, numbers or ranks in the input vectors, the paired Wilcoxon test was used, whereas in case of continuous measures, the paired student's t-test is more appropriate. As can be seen in table 1, the effect of language is far from statistically significant for all the considered measures, i. e., based on these data, there is no evidence for linguistic effects on concept description production. This suggests that the concept description task is mostly tapping into language-independent representations of concepts in semantic memory.

In summary, although there are language-dependent differences in expressing concept descriptions, the analysis conducted on the basis of relation types reveals overall similar type distribution patterns across the languages German, Italian, and also English. Furthermore, the analyses of various global measures of concept description production across German and Italian showed no significant differences between the languages.

Compared to data sets in similar studies, the norms presented here are based on a small set of concepts, which limits the number of experiments they could possibly be used for to a subset of those for which larger norms can be useful. Restricted by our costs for this work this is the maximum of data we could gather. Still, 10 concepts per concept class should be sufficient for many experiments, and considering broad concept classes (e. g., combining "mammals", "birds", "fruit", and "vegetables" into the macro-class "natural" and the remaining classes into the macro-class "artifact"), larger classes can be obtained. Furthermore, we propose here a general annotation scheme and format that should facilitate expanding the norms in future studies.

Conclusion

A data set of highly comparable parallel semantic norms for 50 concrete objects is provided for German and Italian. These are, to the best of our knowledge, the first publicly available semantic norms for these languages, and facilitate an accurate comparison of aspects of concept representations (as mediated by concept description production) in cross-lingual studies. Basic analyses comparing these two languages (and a less detailed comparison of these languages to similar data for English) indicate no remarkable differences across languages, although the distribution of property types used to describe concepts is at least in part affected by language.

Among other purposes, the norms can serve in further studies about semantic memory and concept representation, in particular with German and Italian subjects (separately or together), and possibly involving also English speakers, when our data are complemented with norms from other studies.

Acknowledgements

We thank the German and Italian schools in Bolzano that took part in the collection process, and Andrea Abel, who collaborated to experimental design and data collection. Financial support has been provided by the European Academy of Bolzano (EURAC), Provincia Autonoma di Trento, and Fondazione Cassa di Risparmio di Trento e Rovereto.

References

Artstein, R., & Poesio, M. (2008, December). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596. doi:10.1162/coli.07-034-R2.

Ashcraft, M. H. (1978a). Property dominance and typicality effects in property statement verification. *Journal of Verbal Learning and Verbal Behavior*, *17*, 155–164. doi:10.1016/S0022-5371(78)90119-6.

Ashcraft, M. H. (1978b). Property norms for typical and atypical items from 17 categories: A description and discussion. *Memory & Cognition*, *6*(3), 227–232. Available from http://www.psychonomic.org/backissues/17120//mc/vol6-3/mc-6-227.pdf

Barsalou, L. W. (2003). Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society of London: Series B*, *358*(1435), 1177–1187. doi:10.1098/rstb.2003.1319.

Cavagnoli, S., & Nardin, F. (1999, January). Second language acquisition in South Tyrol: Difficulties, motivations, expectations. *Multilingua – Journal of Cross-Cultural and Interlanguage Communication*, *18*(1), 17–46. doi:10.1515/mult.1999.18.1.17.

Cohen, J. (1960, April). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. doi:10.1177/001316446002000104.

Cree, G. S., McRae, K., & McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science: A Multidisciplinary Journal*, *23*(3), 371–414. doi:10.1016/S0364-0213(99)00005-1.

Dal Negro, S. (2005). Minority languages between nationalism and new localism: The case of Italy. *International Journal of the Sociology of Language*, *2005*(174), 113–124. doi:10.1515/ijsl.2005.2005.174.113.

Forer, D., Paladino, M. P., Vettori, C., & Abel, A. (2008, July). Il bilinguismo in Alto Adige: Percezioni, osservazioni e opinioni su una questione quanto mai aperta. *Il Cristallo – Rassegna di Varia Umanità, Anno L*(1), 49–62. Available from http://www.altoadigecultura.org/rivista.html

Garrard, P., Lambon Ralph, M. A., Hodges, J. R., & Patterson, K. (2001). Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*, *18*(2), 125–174. doi:10.1080/02643290125857.

Guagnano, D. (2010). *Bilingualism and cognitive development: A study on the acquisition of number skills.* Unpublished doctoral dissertation, Centro Interdipartimentale Mente/Cervello (CIMeC), Università degli Studi di Trento. Available from http://eprints-phd.biblio.unitn.it/203

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *18*(4), 441–461. doi:10.1016/S0022-5371(79)90246-9.

Kremer, G., Abel, A., & Baroni, M. (2008, August). Cognitively salient relations for multilingual lexicography. In *Proceedings of the workshop on cognitive aspects of the lexicon (*COGALEX *2008)* (pp. 94–101). Manchester, United Kingdom: Coling 2008 Organizing Committee. Available from http://www.aclweb.org/anthology/W08-1913

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559. Available from http://brm.psychonomic-journals.org/content/37/4/547

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning.

Psychological Review, *85*(3), 207–238. doi:10.1037/0033-295X.85.3.207.
Meyer, D., Zeileis, A., & Hornik, K. (2006, October 19). The strucplot framework: Visualizing multi-way contingency tables with vcd. *Journal of Statistical Software*, *17*(3), 1–48. Available from http://www.jstatsoft.org/v17/i03
Moss, H. E., Tyler, L. K., & Devlin, J. T. (2002). The emergence of category-specific deficits in a distributed semantic system. In E. M. E. Forde & G. W. Humphreys (Eds.), *Category specificity in brain and mind* (pp. 115–147). East Sussex, UK: Psychology Press.
Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*(6), 609–626. doi:10.1037/0033-295X.89.6.609.
Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605. doi:10.1016/0010-0285(75)90024-9.
Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, *81*(3), 214–241. doi:10.1037/h0036351.
Solomon, K. O., & Barsalou, L. W. (2001). Representing properties locally. *Cognitive Psychology*, *43*(2), 129–169. doi:10.1006/cogp.2001.0754.
Vinson, D. P., & Vigliocco, G. (2002). A semantic analysis of grammatical class impairments: Semantic representations of object nouns, action nouns and action verbs. *Journal of Neurolinguistics*, *15*, 317–351. doi:10.1016/S0911-6044(01)00037-9.
Vinson, D. P., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, *40*(1), 183–190. doi:10.3758/BRM.40.1.183.
Wu, L., & Barsalou, L. W. (2009). Perceptual simulation in conceptual combination: Evidence from property generation. *Acta Psychologica*, *132*(2), 173–189. doi:10.1016/j.actpsy.2009.02.002.

A. Archived Materials

The data files described below are provided in simple text format for German and Italian separately, indicated by the file name ending in `_de.txt` and `_it.txt`. The variables in each file are arranged in columns separated by a tabulator space; the names of the variables are defined in the first line of each file. The archived materials comprise the annotated experiment data, separate measures for concepts and features, measures for each concept's features, and concept pair similarities. All data files can be downloaded from `http://www.psychonomic.org/archive`. Furthermore, the archive contains the instructions sheet that was handed out to participants, translated into English (`instructions.pdf`).

A.1. Production Data

The file contains concept stimuli, phrasal responses, and related data as described below. File names:
`production-data_de.txt`, `production-data_it.txt`

Table 2:

Variable Name	Description
Concept(EN)	English translation of the concept name which was presented in the target language in the experiment.
ConceptClass	One of the 10 concept classes that were used in the experiment and which the concept belongs to.
Concept(DE)/Concept(IT)	Concept name in the target language as it was presented to the participants.
SubjectCode	Unique subject number code for each participant (`s1`–`s73` for German participants, `s74`–`s142` for Italian participants).
No.Concept	Number (1–25) defining the order in which concepts were presented to a participant.
No.Phrase	Number defining the order in which the participant produced a descriptive phrase for a concept.
Feature	Phrase in standardised form and translated into English, referring to a feature of a concept (containing underscores instead of spaces).
RelationType	Type of the semantic relation between a concept and the produced feature, as a code of the slightly modified taxonomy of Wu and Barsalou (2009); see the table in section C for the description of the full list of types used here.
Phrase	Original phrase as it was produced by the participant and with minor modifications (reducing the phrase to the pure meaningful content by omitting quantifiers, concept names/pronouns, and correcting spelling errors). Spaces in the phrase are substituted by underscores.

A.2. Concept Measures

Several measures for the set of concepts used in the experiment are provided. Again, we took care to generate measures analogously to the methods used by McRae et al. (2005) for data comparability.
File names:
`concept-measures_de.txt`, `concept-measures_it.txt`

Table 3:

Variable Name	Description
Concept(EN)	English translation of the concept name which was presented in the target language in the experiment.
ConceptClass	One of the 10 concept classes that were used in the experiment and which the concept belongs to.
Concept(DE)/Concept(IT)	Concept name in the target language as it was presented to the participants.
ConceptLemma	Lemma word form of the concept name as used in corpus queries.
FreqWaCKy	Occurrence frequency of the concept name in the WaCKy Web corpus.
logFreqWaCKy	Natural logarithm value of the concept name's occurrence frequency.
No.Letters	Number of letters contained in the concept name.
No.Syllables	Number of syllables contained in the concept name.
No.Features	Number of features that were produced by at least 5 participants for a concept.
No.DistinguishingFeatures	Number of distinguishing features named for a concept – features that were produced only for one or two concepts in the set (from those that were produced by at least 5 participants for a concept).
Percent.DistinguishingFeatures	Percentage of distinguishing features of a concept – number of distinguishing features of a concept divided by the number of features produced by at least 5 participants for that concept.

... continued on next page.

Variable Name	Description
MeanDistinctivenessAcrossFeatures	Average distinctiveness across a concept's features. A feature's distinctiveness is defined as the inverse of the number of concepts for which a feature was produced (calculated across all concepts, rather than only across concepts within a category).
MeanCueValidity	Average cue validity across a concept's features. Cue validity is the conditional probability of a concept given a feature. It was calculated as the production frequency of a feature for a particular concept divided by the production frequency of that feature for all concepts. As above, only features were considered which were produced by at least 5 participants for a concept.
No.IntercorrelatedFeatures	Number of feature pairs of a concept for which features are intercorrelated, considering only those features produced by at least 5 participants for a concept and of these features only those appearing with at least 3 concepts. Correlation computation was based on a vector for each feature comprising the production frequencies of the respective feature for each of the concepts. Intercorrelation of feature (vectors) was calculated using the Pearson product moment. Feature pairs were counted as significantly correlated if the features shared at least 6.5 % of their variance.
Percent.IntercorrelatedFeatures	Number of intercorrelated features divided by all possible pairs of features for a concept.
IntercorrelationalDensity	Sum of the percentage of shared variance for each concept's intercorrelated feature pairs.

A.3. Feature Measures

The measures in this file include only those features for each concept that were produced by at least 5 participants for that concept.
File names:
`feature-measures_de.txt`, `feature-measures_it.txt`

Table 4:

Variable Name	Description
Feature	Phrase in standardised form and translated into English, referring to a feature of a concept (containing underscores instead of spaces).
RelationType	Type of the semantic relation between the produced feature and a concept, as a code of the slightly modified taxonomy of Wu and Barsalou (2009); see the table in section C for the description of the full list of types used here.
No.Concepts	Number of concepts for which the feature was produced.
(Non)Distinguishing	Label that divides the set into distinguishing features (`D`) and non-distinguishing features (`ND`), as described above.
Distinctiveness	Distinctiveness value of the feature, calculated as explained above.

A.4. Concepts-Features

This file includes most of the variables and measures from above for each concept's feature:

Table 5:

Concept(EN)	No.DistinguishingFeatures
ConceptClass	Percent.DistinguishingFeatures
Concept(DE)/Concept(IT)	MeanDistinctivenessAcrossFeatures
Feature	MeanCueValidity
RelationType	No.IntercorrelatedFeatures
FreqWaCKy	Percent.IntercorrelatedFeatures
logFreqWaCKy	IntercorrelationalDensity
No.Letters	(Non)Distinguishing
No.Syllables	Distinctiveness
No.Features	

The additional measures in this file not mentioned earlier are described below. The lines in the file are sorted by English concept name, then by feature rank.
File names:
`concepts-features_de.txt`, `concepts-features_it.txt`

Table 6:

Variable Name	Description
Prod.Frequency	Number of Participants who produced the feature for the respective concept.
Rank	Rank of the feature within the set of features of a concept according to the production frequency (features with the same production frequency were assigned the same rank number within the set of features for a concept).

A.5. Concept Similarities

Similarities between concepts were computed by calculating the cosine distances between each pair of vectors (for two concepts) on the basis of feature production frequencies for these concepts. Values range from 0 (vectors are orthogonal, no similarity) to 1 (identical concepts). The file contains pairs of English concept names and the cosine similarity value, separated by a tabulator space. For the ease in looking up a similarity value for a specific concept pair <`concept1`, `concept2`>, the file contains additionally the line with the concept pair in the opposite order <`concept2`, `concept1`> and the similarity value.
File names:
`concept-cosines_de.txt`, `concept-cosines_it.txt`

B. Concept Stimuli

The set of stimuli used in the experiments (50 concepts from 10 concept classes):

Table 7:

concept class	concepts
bird	goose, owl, seagull, sparrow, woodpecker
bodypart	eye, finger, hand, head, leg
building	bridge, church, garage, skyscraper, tower
clothing	chemise, jacket, shoes, socks, sweater
fruit	apple, cherry, orange, pear, pineapple
furniture	armchair, bed, chair, closet, table
implement	broom, comb, paintbrush, sword, tongs
mammal	bear, dog, horse, monkey, rabbit
vegetable	corn, onion, peas, potato, spinach
vehicle	airplane, bus, ship, train, truck

C. Semantic Relation Types

This is the set of semantic relation types used in the annotation process, the total number of phrases of the respective relation type which were produced in each language (German: DE, Italian: IT), and the percentage based on all phrases produced in the respective language. The first letter of the type code denotes the general semantic relation type, which divides the relation types into one of the 5 groups: entity properties (e), taxonomic categories (c), situational properties (s), introspective properties (i), and miscellaneous (m):

Table 8:

Code	Definition	Example	Lang	No.	%
sf	function	sweater – is worn	DE	1492	14.91
			IT	1284	15.07
ch	superordinate ("higher")	bus – a vehicle	DE	1215	12.14
			IT	1453	17.05
ese	surface property (external)	bear – is large	DE	1358	13.57
			IT	1274	14.95
ece	component (external)	broom – has a brush	DE	1360	13.59
			IT	1247	14.64
sl	location	seagull – lives by the ocean	DE	727	7.26
			IT	462	5.42
eb	behaviour	horse – jumps	DE	427	4.27
			IT	355	4.17
sa	action	spinach – is edible	DE	362	3.62
			IT	331	3.88
se	(associated) entity	chair – used at the table	DE	380	3.80
			IT	280	3.29
em	material made of	socks – made of wool	DE	321	3.21
			IT	272	3.19
eci	component (internal)	cherry – has a pit	DE	307	3.07
			IT	257	3.02
sp	participant	skyscraper – used by humans	DE	308	3.08
			IT	166	1.95
iep	episodic property	hand – is flexible	DE	276	2.76
			IT	161	1.89
eq	quantity of entity	leg – humans have two	DE	196	1.96
			IT	122	1.43
esi	surface property (internal)	pineapple – is yellow inside	DE	185	1.85
			IT	132	1.55
ie	evaluation	bed – comfortable	DE	162	1.62

. . . continued on next page.

Code	Definition	Example	Lang	No.	%
			IT	129	1.51
io	(cognitive) operation	sword – like a long knife	DE	195	1.95
			IT	64	0.75
ic	contingency	airplane – requires pilot	DE	133	1.33
			IT	101	1.19
eae	(associated) abstract entity	rabbit – Easter	DE	125	1.25
			IT	86	1.01
ew	(larger) whole	garage – part of a house	DE	79	0.79
			IT	92	1.08
st	time	owl – found at night	DE	87	0.87
			IT	62	0.73
cl	subordinate ("lower")	finger – thumb	DE	89	0.89
			IT	24	0.28
sr	role	dog – is domestic	DE	61	0.61
			IT	48	0.56
sor	origin	potato – is from America	DE	42	0.42
			IT	31	0.36
cc	coordinate	monkey – relative of humans	DE	18	0.18
			IT	49	0.58
mm	meta-comment	shoes – I own some	DE	62	0.62
			IT	0	0.00
in	negation	eye – without we are blind	DE	21	0.21
			IT	19	0.22
cs	synonym	ship – boat	DE	0	0.00
			IT	14	0.16
ir	representational state	bus – is popular	DE	13	0.13
			IT	0	0.00
ia	affect/emotion	bear – is frightening	DE	6	0.06
			IT	0	0.00
ssw	state of the world	train – is late	DE	0	0.00
			IT	5	0.06
iq	quantity of introspection	bear – has only one young	DE	1	0.01
			IT	0	0.00
sq	quantity of a situation	apple – there are many here	DE	1	0.01
			IT	0	0.00
ss	spatial relation	airplane – flies upwards	DE	1	0.01
			IT	0	0.00