Doctoral Dissertation

A Crow's Beak is not Yellow – Investigations on Cognitively Salient Concept Properties

Gerhard Kremer

26th November 2010

Centro Interdipartimentale Mente/Cervello (CIMeC)
Università degli Studi di Trento

Acknowledgements

First and foremost, I thank Marco Baroni for doing a great job as tutor and research colleague, and for being a pleasant fellow, too. I am also grateful to Andrea Abel from EURAC for acting as my co-advisor, for contributing on issues in lexicography, and for establishing connections to other researchers. More people directly or indirectly supported my PhD studies, and to a larger or smaller extent. They either (make your own choice) took care of the appropriate education for a kick-off start into PhD time, made good examples I like following, engaged with me in technical discussions, gave away a bunch of practical hints, tuned in to a comfortable working atmosphere, were essential during creative breaks, consulted in intercultural peculiarities, generated the feeling of feeling at home, added a little craziness every now and then, or simply kept me up and running:

Hermann Achmüller, Steffen Ade, Stefanie Anstein, Eduard Barbu, Aoife Cahill, Alfonso Caramazza, Federica Cavicchio, Alessandro Chinello, Alejandro Devalle, Stefanie Dipper, Grzegorz Dogil, Stefan Evert, Sara Fabbri, Mauro Felice, Arne Fitschen, Michele Furlan, Ulrich Heid, Amaç Herdağdelen, Robert Hlawatsch, Hans Kamp, Lisandro Kaunitz, Manuel Kirschner, Michael Klein, Mamma and Pappa Kremer, Gianluca Lebani, Anke Lüdeling, Verena Lyding, Brad Mahon, Mara Mazzurega, Andrea Mognon, Brian Murphy, Eduardo Navarrete, Yavor Nenov, Tim Nonner, Nathaniel Obadia, Hugo Gonçalo Oliveira, every participant in my experiments, Francesca Perini, Stefanos Petrakis, Emmanuele Pianta, Franziska Plathner, Massimo Poesio, Magdalena Putz, Giorgio Revolti, Uwe Reyle, Anne Schobert, Antje Schweitzer, Kati Schweitzer, Amy Spencer, Egon Stemle, Luigi Tamè, Genny Tartarotti, Sabine Schulte im Walde, Michael Walsh, Wolfgang Wokurek, Roberto Zamparelli, and Heike Zinsmeister.

These persons can only form a subset of the actual party concerned; that includes only those who were "cognitively most salient" to me at the moment of writing this. No excuse. The others not appearing here most certainly will (or should) know, anyways, that they made my day and were equally important to me and my PhD years.

Contents

List of Tables

List of Figures

1. Introduction

Subject-generated concept descriptions in terms of properties of different kinds (category: *rabbits* are *mammals*, parts: they have *long ears*, behaviour: they *jump*, ...) are widely used in cognitive science as proxies to feature-based representations of concepts in the mind (Garrard et al., 2001; McRae et al., 2005; Vinson and Vigliocco, 2008). These *feature norms* (as collections of subject-elicited properties are called in the relevant literature) are used in simulations of cognitive tasks and experimental design. Moreover, vector spaces that have subject-generated properties as dimensions have been shown to be a good complement or alternative to traditional semantic models based on corpus collocates (Andrews et al., 2009; Baroni et al., 2010).

Since the concept–property pairs in feature norms resemble the tuples that semantic relation extraction algorithms extract from corpora (Hearst, 1992; Pantel and Pennacchiotti, 2006), recent research has attempted to extract feature-norm-like concept descriptions from corpora (Almuhareb, 2006; Baroni et al., 2010; Shaoul and Westbury, 2008). From a practical point of view, the success of this enterprise would mean being able to produce much larger norms without the need to resort to expensive and time-consuming elicitation experiments, leading to wider cognitive simulations and possibly better vector space models of semantics. Lexical resources incorporating semantic relations between lexical entries (e. g., WordNet, see Fellbaum, 1998) would profit likewise from such automatic extraction methods that would facilitate extending the lexical resource with relation instances that are prominent to speakers from a cognitive perspective. From a theoretical point of view, a corpus-based system that produces human-like concept descriptions might provide cues of how humans themselves come up with such descriptions.

The general goals of this dissertation are (i) to report empirical investigations of the cognitive salience of semantic relation types and (ii) to present a case study about extracting cognitively salient concept properties from text corpora, namely composite expressions for constitutive parts of concepts (e. g., *crow*: has a *black beak*).

The next section in this introductory chapter gives an overview on dictionaries and lexical databases that incorporate semantic relations. None of these resources used an empirical approach as the basis for collecting the included relations. The subsequent section discusses a few efforts to build semantic feature norms and their usefulness for the psycholinguistic community. Semantic norms have so far been collected for the English language, mainly. After that section, previous methods for extracting semantically related words for given concepts from text corpora are presented. Furthermore, the section motivates the focus of this research on extracting (modifiers of) composite *part* relations for given concepts. Finally, the introduction chapter concludes with the dissertation outline.

1.1. Dictionaries and Lexical Databases

This section gives an overview on dictionaries and other lexical resources relevant to this research. Particularly, it focuses on the selection and the types of semantically related words that are included along with the lexical entries.

In most paper-based general and learners' dictionaries only some information about synonyms and sometimes antonyms is presented. Newer dictionaries, such as the "Longman Language Activator" (see Summers, 1999), are providing lists of related words. While these will be useful to language learners, information about the *kind* of semantic relation is usually missing.

Semantic relations are often available in electronic resources, most famously in WordNet (see Fellbaum, 1998), an electronic lexical database, where synonymous words are combined into semantically related synsets which are linked to each other. However, WordNet comprises only taxonomy-related semantic relations. While WordNet's target language is English, similar projects emerged for other languages, such as GermaNet[1] (for German) or MultiWordNet[2] (aiming to build parallel word nets for a set of languages). Related lexical resource projects for learners including semantic relations have been developed (e. g., KirrKirr[3] or ALEXIA[4]) — but their entries were either linked to each other manually, or the method for collecting semantically related words has not been made transparent. In general, the salience of the relations incorporated is not verified experimentally. Furthermore, these resources tend to include few relation types (such as hypernymy, meronymy, or antonymy), and the same set of relations is used for all words with the same part-of-speech. The results of this dissertation, as well as work by Vinson and Vigliocco (2008), indicate that different concept classes should, instead, be characterised by different relation types (e. g., *function* is very salient for tools, but not at all for animals).

The ELDIT[5] dictionary (Electronic Learners' Dictionary German–Italian; for details see Abel et al., 2003) provides the possibility to explore the semantic neighbourhood of a word meaning by browsing a set of closely related words, such as hyponyms, co-hyponyms, or (quasi-)synonyms. The relations that define these so-called "word fields" in ELDIT have been chosen on didactic and theoretical lexico-semantic grounds (in particular, structural semantics and word field theory — see, e. g., Geckeler, 2002; Hoberg, 1970) rather than being based on experimental data determining which relations are more salient for native speakers. "Word field" input in ELDIT has been manually carried out by lexicographers who used data sources such as online lexical resources[6], WordNets, or synonym dictionaries[7], resulting in a rather small set of entries (currently,

[1] see URL `http://www.sfs.uni-tuebingen.de/GermaNet`
[2] see URL `http://multiwordnet.fbk.eu`
[3] see URL `http://nlp.stanford.edu/kirrkirr`
[4] see Chanier and Selva (1998)
[5] The present work is a cooperation with the ELDIT project that has been developed at the European Academy of Bozen/Bolzano (EURAC). The dictionary is accessible at URL `http://www.eurac.edu/eldit`.
[6] e. g., URL `http://wortschatz.uni-leipzig.de`
[7] e. g., Müller (1997), or Stopelli (1999)

a few hundred). One future application of this dissertation's research might be the enrichment of electronic learners' dictionaries, such as the ELDIT dictionary, on the basis of cognitively motivated decisions.

ConceptNet[8] has a slightly different approach for harvesting facts for its database (that, in turn, is used for a commonsense reasoning engine) consisting of pairs of related words and the type of relation between these. Entities in the database (embedded in natural language templates, e. g., "*conceptA* is for *conceptB*") are presented to contributing users on the web who may use the given templates to create more facts for the database. Additionally, they may rate facts entered by other users as true or false. Frequencies of the facts entered and ratings create a ranking of facts representing the facts' truth reliability. Currently, the classification of word pairs uses a closed set of 23 types of relations, such as "IsA", "PartOf", and "AtLocation". That is, these semantic relation types restrict what participants may produce. Moreover, participants may take their time in consciously thinking about ever more facts for a concept — thus, the collection method does not substitute an experiment with a controlled environment and produced relations are not necessarily cognitively salient.

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

1.2. Semantic Feature Norms

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 and Schaffer, 1978), where semantic features are used as the basis for constructing conceptual representations (see Murdock, 1982).

Typically, researchers who aim 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 et al. (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. In search of an explanation for the participants' systematic use of features, e. g., Barsalou (2003) assumes that, when generating features, participants simulate a holistic representation of the target

[8] see URL `http://csc.media.mit.edu/conceptnet`

category and then interpret this simulation by using featural and relation simulators. According to this view, 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 is 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 et al. (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 et al. (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 et al. (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 et al., 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 and 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 (Garrard et al., 2001; McRae et al., 2005; Vinson and Vigliocco, 2008). Garrard et al. (2001) instructed subjects to complete phrases ("*concept* is/has/can. . . "), thus restricting the set of producible feature types. McRae et al. (2005) instructed their subjects to list concept properties without such restrictions, but providing them with some examples. The produced features were then normalised and

classified into categories such as *part* and *function* by the experimenters. The published norms include, among other kinds of information, the production frequency of each feature listed for a concept by the participants. Vinson and Vigliocco (2008) gave similar instructions, but explicitly asked subjects not to freely associate, aiming to exclusively collect concept descriptions that actually concern properties of the concept (e. g., *apple*: "on trees", instead of "seven dwarfs"). Typically, the data produced by participants is published along with statistical data from analyses regarding psycholinguistic variables, such as 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.

Norms have been collected mainly for the English language. One question at this point is whether feature-based concept representations are language-dependent or, instead, generalisable across languages. De Deyne and Storms (2008) conducted a feature elicitation experiment for the Dutch language and translated part of the resulting feature norms into English. Nevertheless, comparison to the English norms would be inaccurate as experiment designs are different (De Deyne and Storms (2008) asked their participants for the first three associations for each concept) and participant groups grew up in different environments (and thus made different experiences that might influence which concepts they know and which details these comprise). This dissertation describes the acquisition of parallel norms for German and Italian from participants living in Bolzano in South Tyrol (a region in Italy where Germans and Italians coexist without being in general bilinguals). Experiment design and transcription of the data follow McRae et al. (2005).

1.3. Relation Extraction

In addition to an empirical study on cognitively salient concept properties, a second investigation in this dissertation explores methods to automatically harvest such properties from text corpora. This section overviews research works in semantic relation extraction that address those kinds of tasks.

Many approaches focus on the acquisition of semantically similar nouns. In one of the first approaches, Hindle (1990) used the annotated structure of a parsed text to analyse predicate–argument structures. To find similar nouns, he relied on the distributional hypothesis (cf. Harris, 1985) and applied distributional similarity metrics.

One of the early approaches to acquire word pairs with a particular semantic relation is described in Hearst (1992). She used lexico-syntactic patterns ("*noun*, such as *noun*, ...") to extract noun pairs for the semantic relations hyponymy and hypernymy from a POS("part-of-speech")-tagged text corpus. In a similar approach, Almuhareb and Poesio (2004) used pure word-based patterns (e. g., "the *feature* of the *concept* [is|was]"), thus circumventing the need of a POS-tagged corpus. Following the approach of Hearst (1992), Pantel and Pennacchiotti (2006) used seed instances across parts of speech with a known semantic relation to acquire generic lexico-syntactic patterns. After applying a reliability measure, additional instances were extracted from the POS-

tagged corpus with a bootstrapping procedure.

Almuhareb (2006) was the first to attempt to reproduce subject-generated features with text mining techniques. He computed precision and recall measures of various pattern-based feature extraction methods using Vinson and Vigliocco's norms for 35 concepts as a gold standard. The best precision was around 16 % at about 11 % recall; maximum recall was around 60 % with less than 2 % precision, confirming how difficult the task is. Importantly for our purposes, Almuhareb (and, more recently, Devereux et al., 2010) removed the modifier from composite features before running the experiments ("one wheel" converted to "wheel"), thus eschewing the main characteristic of subject-generated concept descriptions that we tackle below. Shaoul and Westbury (2008), Baroni et al. (2010), and Baroni and Lenci (2010) used corpus-based semantic space models to predict the top 10 features of 44 concepts from the McRae norms. The best model (Baroni et al.'s Strudel) guesses on average 24 % of the human-produced features, again confirming the difficulty of the task. And, again, the test set was pre-processed to remove modifiers of composite features, thus sidestepping the problem we will focus on. It is worth remarking that, by removing modifiers, previous authors are making the task easier in terms of feature extraction procedure (because the algorithms only need to look for single words), but they also create artificial "salient" features that, once the modifier has been stripped of, are not that salient anymore (what distinguishes a monocycle from a tricycle is that the former has one wheel and the latter three, not simply having wheels). It is conceivable that a method to assign sensible modifiers to features might actually improve the overall quality of feature extraction algorithms.

The corpus-based models proposed for this task up to this point overlook the fact that participants in experiments very often produce *composite* properties: Participants state that rabbits have *long* ears, not just ears; cars have *four* wheels; a calf is a *baby* cow, etc. Composite properties are not multi-word expressions in the usual sense. There is nothing special or idiomatic about *long ears*. It is just that we find it to be a remarkable fact about rabbits, worth stating in their description that their ears are long.

In the feature norms described in section 2.1, *part* relations were frequently encountered (1,667 of the 10,010 phrases produced in total were *parts*), and these were often composite expressions (625, i. e., more than one third of the *part* relations). From that set of composite *part* relations, 404 were composed of an adjective and a noun (the target of chapter 3 on suitable extraction methods for such composite expressions). Looking at the distinct *parts* that were elicited, 92 were always produced with a modifier, 280 only without modifier, and 122 both with and without modifier. That is, for about 43 % of the *parts* at least some speakers used a composite expression of adjective and noun. Note that while our focus is on feature norms, a similar point about the importance of composite properties could be made for other knowledge repositories of significance to computational linguistics, such as WordNet (Fellbaum, 1998) and ConceptNet (Liu and Singh, 2004), approximately 68,000 (36 %) of the entries and 1,300 (32 %) of the part entries being composites, respectively.

1.4. Dissertation Outline

This work has two main parts: The first part concentrates on acquiring experiment data regarding salient semantic relations that serve in the second part for tuning and evaluating the computational extraction of such data from corpora.

In more detail, chapter 2 empirically investigates the cognitive salience of semantic relations for a set of (concrete) basic-level concepts (cf. Murphy, 2002). A first experiment collected descriptions for a set concepts that are presented to participants. The aim was to find out which concept properties are prominent to native speakers, and to use these results later as a basis for the extraction of cognitively salient relations. Such semantic relations can be useful in psycholinguistic research using feature-based concept representations, and also for extending lexical resources. To systematically extract properties for concepts, the collected production data is categorised into types of semantic relations and classes of concepts. A complementing perception experiment tests if relation types that were prominently produced are also perceived to be salient by speakers. To our knowledge, no such approach that investigates empirical evidence for cognitively salient relations with the purpose to automatically extract appropriate relations for extending a lexical resource has been reported, yet. Furthermore, we collected our feature norms in parallel for two target languages (German and Italian) to discover possible generalisations across languages (in contrast to previous studies that mostly concentrated solely on the English language).

In chapter 3, the focus are composite expressions of *part* relations of concepts, as they are commonly encountered in participant-produced concept descriptions, and because they are commonly composite. The automatic extraction of this example for a semantic relation type is simplified by assuming that salient *part* nouns for given concepts have already been identified in a preceding step (using an already existing algorithm — e. g., see Girju et al., 2006). The purpose of the methods described is to select appropriate adjectival modifiers for the *part* nouns according to rankings based on co-occurrence frequencies in text corpora. A set of the five best modifiers is output per concept–part pair. The method performances are evaluated first on the collected German and Italian *production* norms. Alternatively, to evaluate the extracted pairs based on which of these are *perceived* by speakers as being reasonable, plausibility ratings of the list of modifiers serve as the gold standard for a second performance evaluation. Furthermore, the best method is evaluated on a set of concept–part pairs that were not seen during the tuning of the selection algorithm. A separate section in that chapter describes the futile attempts to improve method performances.

The concluding chapter summarises the work presented.

2. Empirical Investigations

To explore which properties are salient to native speakers for a given concept, an empirical study was conducted. This chapter describes two experiments investigating the cognitive salience of relations. One experiment yields empirical results based on properties participants produced prominently for a given concept, and the second experiment analyses which properties they perceived as being salient of a concept. As the goal is to use these findings for an automatic extraction approach (described in the next chapter), systematic characteristics are sought for concept *classes* and relation *types*. Furthermore, the production data collected in the first experiment is used for tuning the extraction algorithm and as the evaluation basis for its performance analysis.

2.1. Production of Concept Properties

The aim of the first behavioural experiment described below is to investigate which properties most participants use when describing a selected set of concrete concepts. Investigations include analyses of differences across and within concept classes, and a comparison between the two target languages German and Italian (in turn compared to the data from a separate study on the English language).

2.1.1. Experiment Design

Similar to previous approaches in other studies (Garrard et al., 2001; McRae et al., 2005; Vinson and Vigliocco, 2008), a property elicitation experiment was conducted, which is described in this section. For a given concept in the stimuli set, participants described its properties. The collected data set was then annotated with relation types holding between concepts and the produced properties.

Stimuli

The stimulus set was a collection of 50 concrete concepts from 10 different concept classes (displayed in table 2.1). 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 differences in corpus frequencies of the target words in German, Italian, and English

Table 2.1.: The set of stimuli used in the production experiment — 50 concepts from 10 concept classes

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	aeroplane, bus, ship, train, truck

corpora revealed significantly larger frequencies for words in the *bodypart* 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.

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.

To emphasise this fact, 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 et al., 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 (e. g., 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 their mother tongue is German or Italian (or if they belong 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 the experiment survey presented here 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 from the analysis), 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 the study by McRae et al. (2005), typically involve older participants, such as university students. In total, 73 German students and 69 Italian students took part in the experiment.

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 two 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). The time limits requested by the schools for the experiment sessions restricted the number of concepts to be presented to each participant, which is why no participant was given the whole set of 50 concepts.

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 this study aimed at, participants were asked to 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 four phrases per target word. The task time was set to 1 minute per concept, and participants were not allowed to go back to a word they had previously described.

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

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.

2.1.2. Transcription and Labelling

The collected data comprised for each concept, on average, descriptions by 36 German participants (standard deviation 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, the transcribers tried to systematically apply the criterion that, if at least one participant produced two 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 manually mapped 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 the study's "annotation guidelines", in order to keep the data comparable between this project's target languages and 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.

Relation Type Mapping

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 A. While trying to adapt the annotation style of McRae et al. (2005), dubious cases were encountered. For example, in their norms, "carnivore" is classified as a *category*, whereas "eats_meat" is classified as a *behaviour*. As these seem to convey the same information, both were mapped to "eats_meat", classified as *behaviour*.

Apart from the semantic relation types described in Wu and Barsalou (2009), the additional semantic relation types in the annotation scheme of the present study comprise *material* (em), *role* (sr), and *episodic property* (iep).[4] Differently from the annotation scheme that McRae et al. (2005) applied, the *material* something is made of was separated 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*

[4] Following the coding scheme of Wu and Barsalou (2009), the first letter of a type code denotes one of the following five 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 A for the full list of type codes we used in the annotation process.

property (esys) in Wu and Barsalou's annotation scheme, but this relation is a quite openly defined relation type, so the present study provided the *episodic property* type (iep) for properties that can not be directly perceived when encountering a concept (e. g., "is strong" requires some kind of inference from perceptual data).

Language-Dependent Differences

During transcription of the produced phrases into English and mapping onto standardised phrases, structural language-dependent differences were observed. For example, in German, expressions denoting a complex meaning (e. g., domesticated animal or pet) are often expressed by noun compositions ("Haus|tier"), 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), the assumption was that such a complex expression produced was to convey both parts of the meaning at once, which is why in this case two relation types were assigned: *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, two meaningful aspects are assumed to be 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, no attempt was made to capture the commonalities.

Other possible language differences that might have lead to asymmetries in translation and mapping are 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").

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.

Inter-Coder Agreement

To test the inter-coder reliability in mapping phrases to relation types, for each target language another native speaker labelled 100 randomly sampled standardised phrases. We compared the agreement between their labels and the annotated labels in the original data set (these secondary annotators were trained using phrases that were not included in the random sample). The agreement between the original annotation and that of the secondary annotators was rather high, with kappa values (using Cohen's κ) of 0.844 for German and 0.676 for Italian. Cohen's κ 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). Despite 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 and 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.

2.1.3. Analysis

When describing the data collected from the experiment, the focus here is 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 the collected German and Italian data, but it also compares the relation type distribution in these 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 not counting 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, which should be subject to future investigations (an explanation has not been found, yet). There were 187 German and 196 Italian concept–property pairs that were produced by at least ten participants. Of those, 117 were shared across languages (i. e., 63 % in the German data and 60 % in the Italian data).

Distribution of Relation Types

The number of properties grouped by the annotated relation types are presented in appendix A. The relation type codes (in the style of Wu and Barsalou) used in the

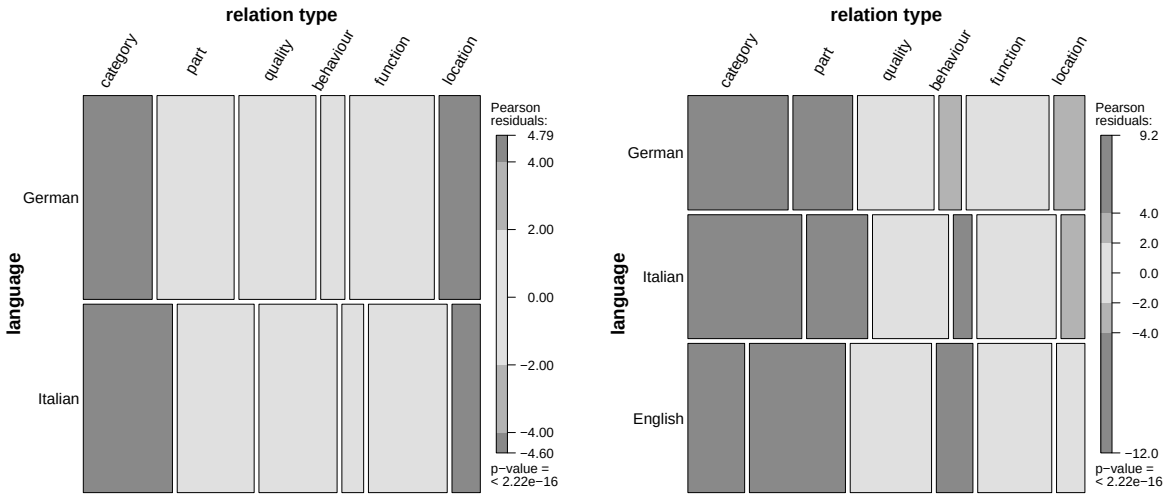

Figure 2.1.: Overall frequency distribution of phrases of one of the six relation types that were annotated most frequently for each target language (left); distributions compared to McRae et al.'s data (English) — including in all languages only phrases produced by at least five participants for a concept (right)

annotation are explained there. Figure 2.1 displays the overall frequency distributions of the top six relation types. The data subset including only these six 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 (see Meyer et al., 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 six 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 six 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 six 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 (for details see Meyer et al., 2006) — darker rectangles correspond to larger (and more significant) deviations from the cross-language distribution.

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

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

more often by German participants than by Italian 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, interestingly, a small set of more or less abstract hypernyms 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" (relation type *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 2.1 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 the present work. For uniformity with the available English data, for this plot only relations produced by at least five 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 taken into account. For four 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 *bodypart* 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 differences in *category* and *part* relations, the relative distributions are roughly rather similar between

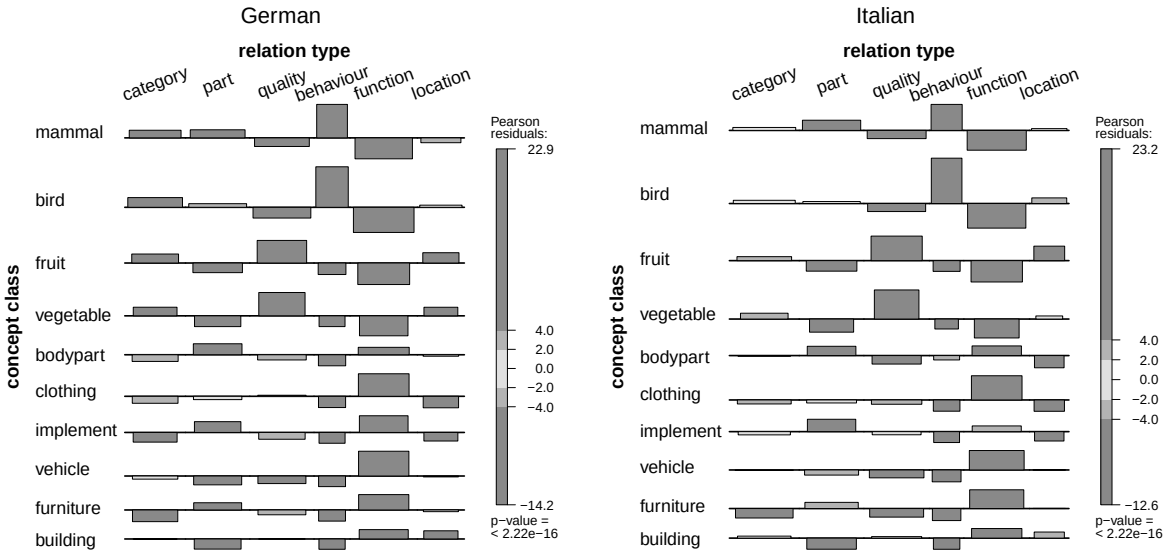

Figure 2.2.: Frequency count deviations from the overall distribution of phrases of the six relation types considered for the German (left) and the Italian data (right). Rectangles above/below the horizontal lines indicate over-/underrepresented counts; magnitudes of significance are coded by different shades of grey.

the three languages.

Additionally, the differences between German, Italian and English were investigated 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 six 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.

Relation Type Distributions per Concept Class

Next, relation type distributions for each of the concept classes are shown in separate association plots for German and Italian (see figure 2.2). Here, the position of the rectangles relative to the horizontal lines indicates overrepresented (above the line) and underrepresented (below the line) counts for a relation type within a particular concept class, compared to the overall distribution as seen in the left plot of figure 2.1. A relation type for a specific concept class is over-/underrepresented 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. The width of a rectangle is a measure for the value expected from the overall distribution;

the height of a rectangle is a measure for the degree of the deviation from the expected value. Similar to the mosaic plots described above, the magnitude of the statistical significance is coded by shades of grey: the more significant a deviation is, the darker the shade of the rectangle.

Comparing the two languages, we observe that the deviations are roughly similar, i. e., the positions of the rectangles relative to the baseline are the same for most cells across languages. Furthermore, some concept classes have similar deviations within a language, most evidently *fruit* 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 deviations, e. g., compare *fruit* and *implements*, where for *implements* a lot more relations than expected (from the overall distribution) of types *part* and *function* were produced in contrast to the *fruit* class, which in turn is characterised by larger positive deviations of *category* and *quality* relations than in the *implement* class.

The type patterns associated with specific concept classes are not particularly surprising, and they have been already observed in the literature (Vinson et al., 2003; Baroni and Lenci, 2008). In particular, living things (animals and plants) are characterised by paucity of functional features, that instead characterise all man-made concepts. Within the living things, animals are characterised by typical *behaviours* (they bark, fly, etc.) and, to a lesser extent, *parts* (they have legs, wings, etc.), whereas plants are characterised by a wealth of *qualities* (they are sweet, yellow, etc.) Differences are less pronounced within man-made objects, but we can observe *parts* as typical of tool and furniture descriptions. *Behaviour*, not surprisingly, pertains to vehicles only. Finally, *location* is a more typical definitional characteristic of buildings (for clothing, nothing stands out, if not, perhaps, the pronounced *lack* of association with typical *locations*). *Bodyparts*, interestingly, have a type profile that is very similar to the one of implements — manipulable objects are, after all, extensions of our bodies.

Hierarchical Clustering of Concepts

The distributional analysis presented above confirmed the main hypotheses — that particular relation types are salient properties of concepts that differ from a concept class to the other, but are robust across languages. However, skewing effects associated to specific concepts were not taken into account so far (e. g., it could be that, say, the property profile observed for *bodyparts* in figure 2.2 is really a deceiving average of completely opposite patterns associated to, say, heads and hands). Moreover, this analysis already assumed a division into classes — but the type patterns, e. g., of mammals and birds are very similar, suggesting that a higher-level *animal* class would be more appropriate when structuring concepts in terms of type profiles. Both issues are tackled in an unsupervised (hierarchical) clustering analysis of the 50 target concepts based on their property types. If the postulated classes are not internally coherent, they will not form coherent clusters. If some classes should be merged, they will cluster together.

Concepts were represented as 6-dimensional vectors, with each dimension corres-

ponding to one of the six common relation types discussed above, and the value on a dimension given by the number of times that concept triggered a response of the relevant type. Using the functions implemented in R, Euclidean distances between concepts based on the described vectors were calculated. These were the input for the hierarchical cluster analysis using the complete linkage method. Complete linkage means that the distance between two clusters is defined by the longest distance between any two members of the clusters.

Looking at the clustering results, both in German (see figure 2.3 on the following page) and in Italian (see figure 2.4 on page 21), the best solution is a three-way partition of the concept set into animals (*mammals* and *birds*), plants (*vegetables* and *fruit*), and objects (*clothing*, *implements*, *vehicles*, *furniture*, and *buildings*) plus *bodyparts* (that, as observed above, have a distribution of types very similar to the one of tools).

The type profiles associated with animals, plants and objects plus *bodyparts* have enough internal coherence that they robustly identify these macro-classes in both languages. Interestingly, a three-way distinction of this sort — excluding *bodyparts* — is seen as fundamental on the basis of neuro-cognitive data by Caramazza and Shelton (1998). On the other hand, more granular distinctions could not be made based on the few (six) and very general types used.

Finally, the peculiar object-like behaviour of *bodyparts* stresses that concept classification is not a trivial task, once trying to go beyond the most obvious categories typically studied by cognitive scientists — animals, plants, implements.

2.1.4. Summary

This section described a multi-lingual concept description experiment. Participants produced different semantic relation type patterns across concept classes. Moreover, these patterns were robust across the two languages studied in parallel. Similarities in overall distribution of relation types were found also for the English data set from a previous study by McRae et al. (2005). A closer look at the data suggested that linguistic constraints might affect verbalisations of conceptual representations (and thus, which properties are produced), but in general, language-independent aspects were found. In summary, the result of this study is promising and could be used in the procedure for automatically harvesting semantically related words for a given entry in a lexical resource: Knowing the corresponding (broad) concept class, those semantic relations types should be focused on for extraction that proved to be salient for that class an those that were in general produced frequently.

However, so far only concrete concepts were considered. To be able to cover more concept classes, the stimuli set in a future experiment will have to be expanded to include, e. g., abstract concepts — although the hope is to mine some abstract concept classes on the basis of the properties of the present concept set (colours, for example, could be characterised by the concrete objects of which they are typical).

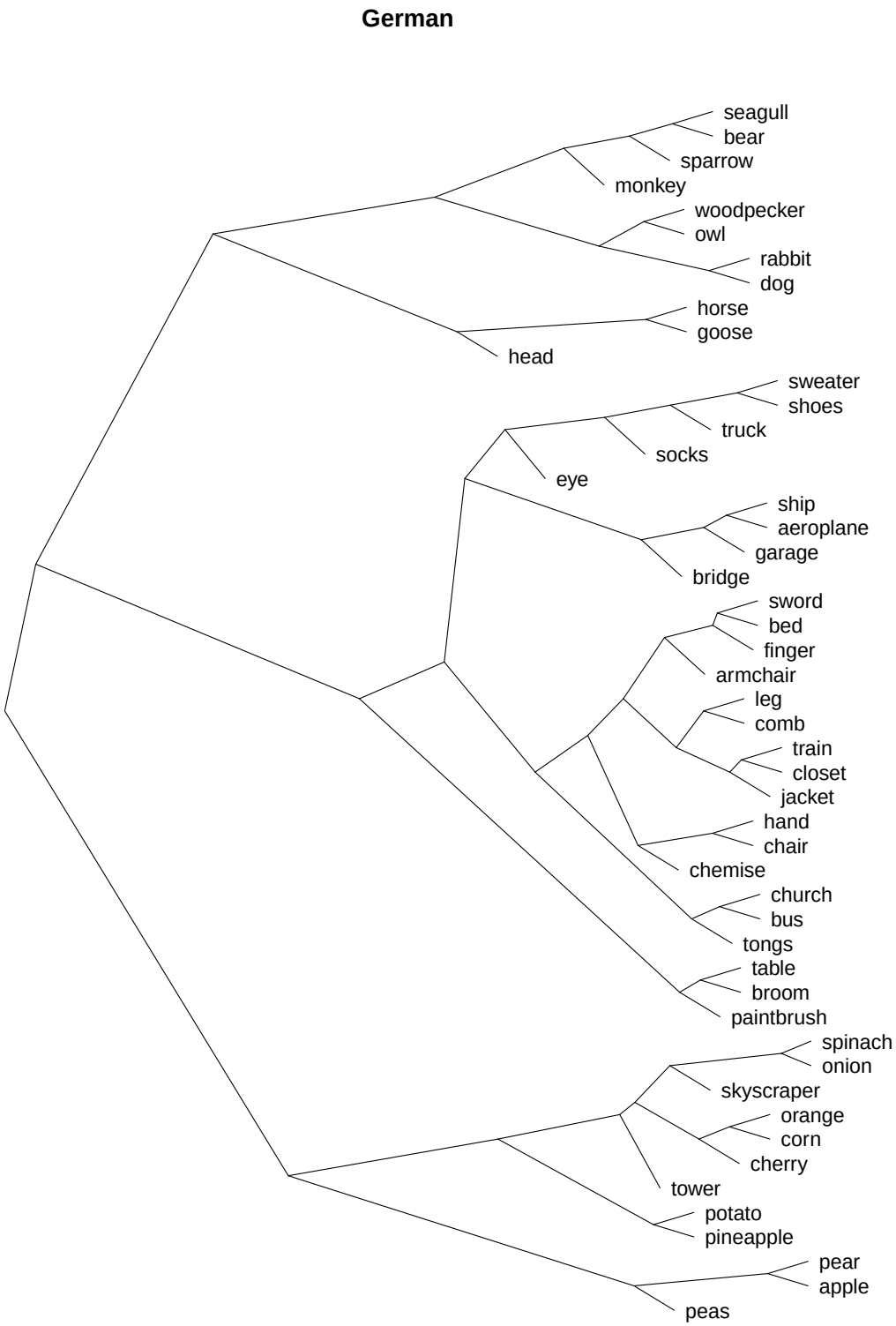

Figure 2.3.: Hierarchical clustering dendrogram for German concepts clustered by numbers of the top six relation types produced

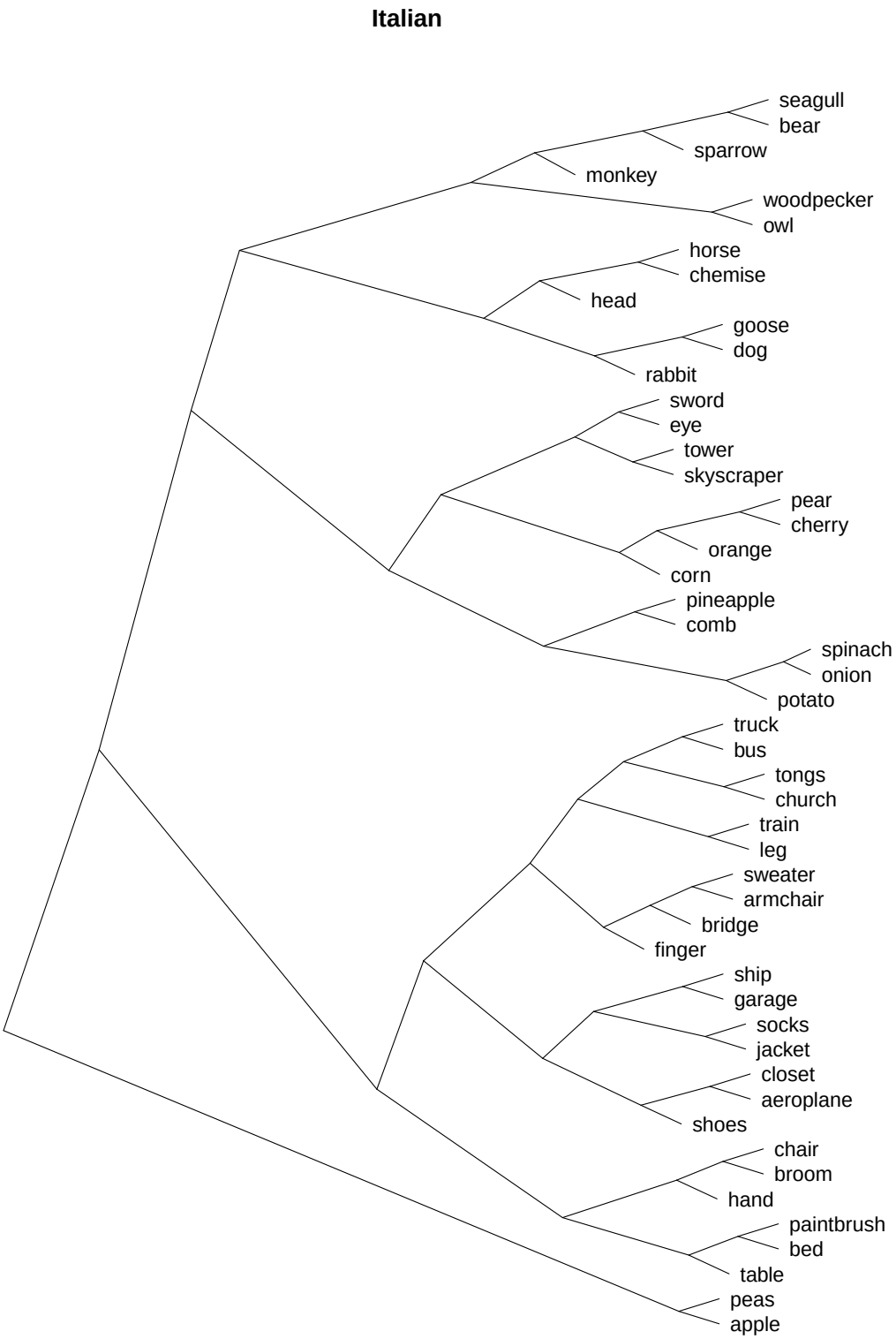

Figure 2.4.: Hierarchical clustering dendrogram for Italian concepts clustered by numbers of the top six relation types produced

2.2. Perception of Concept–Property Pairs

The behavioural experiment described in this section addresses the question if the differences in cognitive salience that were found in the production experiment can be confirmed when testing the participants' perception performances. If so, the evidence for concept-specific sets of salient relation types would be stronger and invite further investigation in that direction. If, on the other hand, the results from the previous experiment are contradicted by the findings in this follow-up experiment, they should be reinterpreted and considered with limitations.

Response times and error responses from the perception experiment are analysed statistically using a mixed effects model and visually by comparing differences in error rates.

2.2.1. Experiment Design

This section describes the feature verification experiment paradigm used (see, e. g., Cree et al., 2006). For the analysis, reaction times and response errors were recorded. In line with the preceding production experiment, the target languages were German and Italian.

Stimuli

The stimuli set contained word pairs consisting of a concept word and a word semantically related. For each of the 10 concept classes from the production experiment (*mammal*, *bird*, *fruit*, *vegetable*, *bodypart*, *clothing*, *implement*, *vehicle*, *furniture*, *building*), 5 concept words were in the set. As in the production experiment, concepts were taken mainly from the sets used in the experiments of McRae et al. (2005) and Garrard et al. (2001). Each concept word was paired with six semantically related words that covered all six relation types used in the analysis of the production experiment (*category*, *part*, *quality*, *behaviour*, *function*, *location*). This set of stimuli word pairs was extended with a control set of the same size (300 pairs) including the same concepts paired with words of the six mentioned relation types, but for which the related words were invalid for the corresponding concepts (e. g., "songbird" for the concept "goose").

Regarding the selection of appropriate stimuli word pairs for the experiment, two conditions had to be met. First, the semantic relation type for the semantically related word of a concept should be reasonably easy to infer, as there was no explicit type indication provided for participants during the experiment. Second, participants should be prevented from (possibly unconsciously) developing their own strategies of differentiating valid from invalid word pairs other than consciously thinking about the relatedness (besides, this makes the cognitive task more demanding, and thus it spreads the ranges of response times, which facilitates a more significant statistical result). To hinder participants from responding solely on the basis of the existence of a strong association between the valid pairs' items as opposed to a weak association between items of invalid pairs, both types of pairs had to have a high association

strength. Although all word pairs were selected carefully by hand to account for both issues, not in every case ideal pairs could be found.

As cognitive association strength of words was shown to correlate with their co-occurrence frequency in corpora (cf. Spence and Owens, 1990), co-occurring words from the German WaCky corpus[6] served as the basis for stimuli pair collection from a ranked list: concept words and all those words co-occurring in a 5-word window around the concept words (within sentences) were counted. Both single-word frequencies and co-occurrence frequencies were determined. Instead of ranking the list according to frequency, association measures were calculated[7] from the frequencies, as these also take into account the frequencies of the single words of a pair co-occurring in other pairs, which is expected to result in a more accurate ranking (although perhaps not appearing very frequent in that composition, a pair of words intuitively has a stronger association if each word does not co-occur with many other words). For each concept, this list was ordered according to the association measure values of the log-likelihood statistic (for details on the calculation, please see equation 3.1 on page 41). From the ordered list, appropriate valid and invalid word pairs that ranked highest were chosen for the stimuli set. For the Italian stimuli set, a native Italian speaker translated the German word pair stimuli. The full set of stimuli (translated into English) is presented in appendix B.

Participants

Participants were students at high schools in Bolzano (South Tyrol) where either German or Italian is used as the teaching language. Please see section 2.1 for the discussion about the monolingual status of these participants. None of the participants in this experiment had taken part in the production experiment. The age of the participants across language groups was in the range between 14 and 18. In total, 70 Italian mother tongue speakers and 72 German mother tongue speakers participated in the experiment.

Procedure

Participants were instructed to decide if a word that was presented on a laptop monitor — and followed the presentation of a concept word — could be used in a description of that concept. On the laptop's touchpad, the left button should be pressed with the left index finger for "no" ("It can not be used in a description of the concept") and the right button should be pressed with the right index finger for "yes" ("It can be used in a description").

Before the real experiment, all participants did a short experiment run with 22 example trials to get used to the task. Subsequently, some of the given examples were discussed to make sure the task was understood as intended. In the real experiment, the first five trials were considered as starter trials. All words used in the short run or

[6] see the WaCky project at URL `http://wacky.sslmit.unibo.it` for details

[7] using the UCS framework; see URL `http://www.collocations.de/software.html`

as starters did not appear in the stimuli set and were not taken into account for the analysis. Furthermore, participants filled out an anonymous questionnaire, specifying their mother tongue, the mother tongue of each of their parents, other languages spoken, age, sex, and handedness.

Two very similar laptop models were used for presenting the stimuli and to simultaneously run two experiments at a time. The software presenting the stimuli and gathering the data was Vision Egg (see Straw, 2008), a module for the python programming language[8]. Words appeared in white letters on a black screen, with a font size of 70 pixels. In each trial, stimuli were presented in the following order: First, a red fixation spot was shown in the middle of the screen (for 30 frames, equalling 500 ms for the given monitor refresh rate of 60 Hz). Subsequently, the spot was removed and the concept word was presented (for 500 ms) at the horizontal centre and vertically shifted 105 pixels from the centre of the screen to the top. It was followed by a blank screen (500 ms). After that, the semantically related/non-related word (i. e., in a valid/invalid relation condition) was shown vertically centred and shifted 105 pixels from the centre of the screen to the bottom. From the beginning of the presentation of this word, the participant could press one of the two touchpad buttons, which instantly lead to the next screen: A white fixation spot in the middle of the screen (shown for random durations between 10 and 70 frames, in steps of 10 frames, equalling a duration between 166,$\bar{6}$ ms and 1166,$\bar{6}$ ms). If the participant did not press a button within 2,000 ms (the maximum response time), the presentation moved to the next screen. Every participant saw 300 randomised trials (150 valid and 150 invalid concept–relation pairs), for which an experiment run took between 15 and 18 minutes.

2.2.2. Data Storing

Before the analysis, data were excluded from participants which were not clearly mother tongue speakers by checking in the questionnaire if the target language was indicated as the mother tongue of both parents. In addition, all cases were removed where no response had been given during a trial.

The data gathered in the experiment were saved in a text file and comprised for each trial a line with: the participant code including an incremental count number (`subjNNN`), the code for the laptop model (`D620`/`D630`), the trial number (`trialNNN`), the concept word presented, the (valid/invalid) relation word presented, the validity of the presented word pair (`valid`/`false`, and `starter` for the example trials), the mouse button(s) pressed (`MB_LEFT`, `MB_RIGHT`, `MB_MULTIPLE`, `MB_NONE`), the reaction time of the response measured from the onset of the presentation of the relation word (in seconds), the concept class, the relation type, and the English translation of the concept word. Furthermore, configuration data of the experiment and frame refresh statistics were saved during each experiment run to check for possible differences in laptop performances.

[8] see URL `http://python.org`

2.2.3. Analysis

A statistical analysis using mixed effects modelling as described in Baayen et al. (2008) was conducted, using the R statistical computing environment. Moreover, response errors were analysed.

Mixed Effects Model Analyses

Models that incorporate both fixed and random effects (so-called mixed effects models, see Baayen et al., 2008) help to discover dependencies of an experimental measure from controlled variables, even if this measure might have been influenced by other, random effects that can not easily be controlled for (e. g., individual participant performances).

For the following mixed effects model analyses, all correct responses from German and Italian participants were taken into account. However, all trials concerning the relation types *function* and *behaviour* were excluded: There are no appropriate functions for animals and no appropriate behaviours of a typical tool. This is why many word stimuli for these relation types were dubious and thus might have influenced analysis and complicate interpretation. Furthermore, extreme outliers were excluded by leaving out those trials where the reaction time is below the 0.5-percentile or above the 99.5-percentile.

Reaction Time Analysis

Incremental inclusion of factors and their ANOVA comparisons lead to a model including as relevant factors *concept class*, *relation type*, and their interaction (as fixed effects), and *subject*, *relation word length* and (the English translation of the) *concept word* (as random effects). These factors were used to model the reaction times of the responses given. The aim was to see if reaction time depended on the factors *concept class* and *relation type* that were controlled for. Individual performances of subjects, the length of the relation words to recognise, and particular concepts are all variables that can not be easily controlled for and make the experiment not exactly reproducible (with other subjects and possibly other words for the given concept classes and relation types) — the subsets of subjects, relation words and concept words in this experiment are all from much larger populations. As these could have concealed the effects of the independent variables in the analysis, they were considered by adding them to the model as random effects. The analysis result of the model fit is shown in table C.1 on page 67.

The most important information there are the t-values: As many hundreds of observations (21,082, to be exact) were given as input for the model, an absolute t-value above 2 can be assumed to have a p-value below the 5 % significance level (cf. Baayen et al., 2008) — which represents the probability that the null hypothesis is true ("the two tested distributions are from the same population"). Additionally, the sign of the value gives the direction of the difference of reaction time means. The reference level for the model's estimates were the concept class *building* and the relation type *quality*. These were chosen on the basis of separate analyses of concept classes

from relation types as the approximate mean of positive and negative distributional differences.

Compared to the reference level, 10 out of 27 interactions between concept class and relation types were significantly different — both for shorter and longer reaction times. However, comparing these results with the results from the production experiment, no consistent picture emerged: For example, given that produced descriptions of the relation type *category* for animals (comprising *mammals* and *birds*) were over-represented, one would have expected a shorter reaction time in the current experiment, but this was not the case — responses within the concept class *mammals* seemed to have a longer reaction time, and the t-value for the concept class *birds* was not significant. On the other side, for *parts*, for both concept classes *mammals* and *birds*, reaction times were significantly shorter, just as could have been expected from the over-representation of relations that were produced for these concept classes and relation type.

However, there is neither a single concept class nor a single relation type for which no interaction is significant. That is, even though there is no general pattern arising for which an intuitive interpretation could be given, there were distinct reaction time distributions depending on differing combinations of concept classes and relation types.

Adding *language* as a fixed effects factor (in a separate analysis) changed the values only slightly, and the *language* factor had a non-significant t-value (0.59). That is, on the basis of the present data set, possible differences across languages could not be detected statistically. The same as above is true for the analysis with the additional fixed effects factor *laptop model* and the factor *valid/invalid relation*: no big differences were found when comparing the models' t-values, and the additional factors showed no significant t-values.

Separate analyses for valid relations and invalid relations resulted in differing sets of nine significant interactions, but no clear pattern emerged whatsoever (cf. the analysis results in table C.2 on page 68 and table C.3 on page 69).

Visual Reaction Time Analysis

We turn to the visual analysis of the reaction time distributions, considering the original concept classes separately again. Figure 2.5 on the facing page shows the box-and-whisker plot of the reaction times for correct responses grouped by concept classes and the four analysed relation types. They represent the full data set including both languages, as language had no significant effect on response times in the preceding analyses.

The boxplot indicates the tendency of reaction time medians to be different across combinations of concept classes and relation types. Some can be explained by (and thus confirm) the production experiment results. For example, it seems intuitive to assume slower reaction times for *quality* relations of *birds* compared to *part* relations of *birds*, as *quality* was found to be under-represented and *part* slightly over-represented in the production data. Nevertheless, other reaction time medians are not differing in the directions that would have been expected, so there is no consistent picture to be drawn on the basis of this analysis. Furthermore, all reaction time distributions are

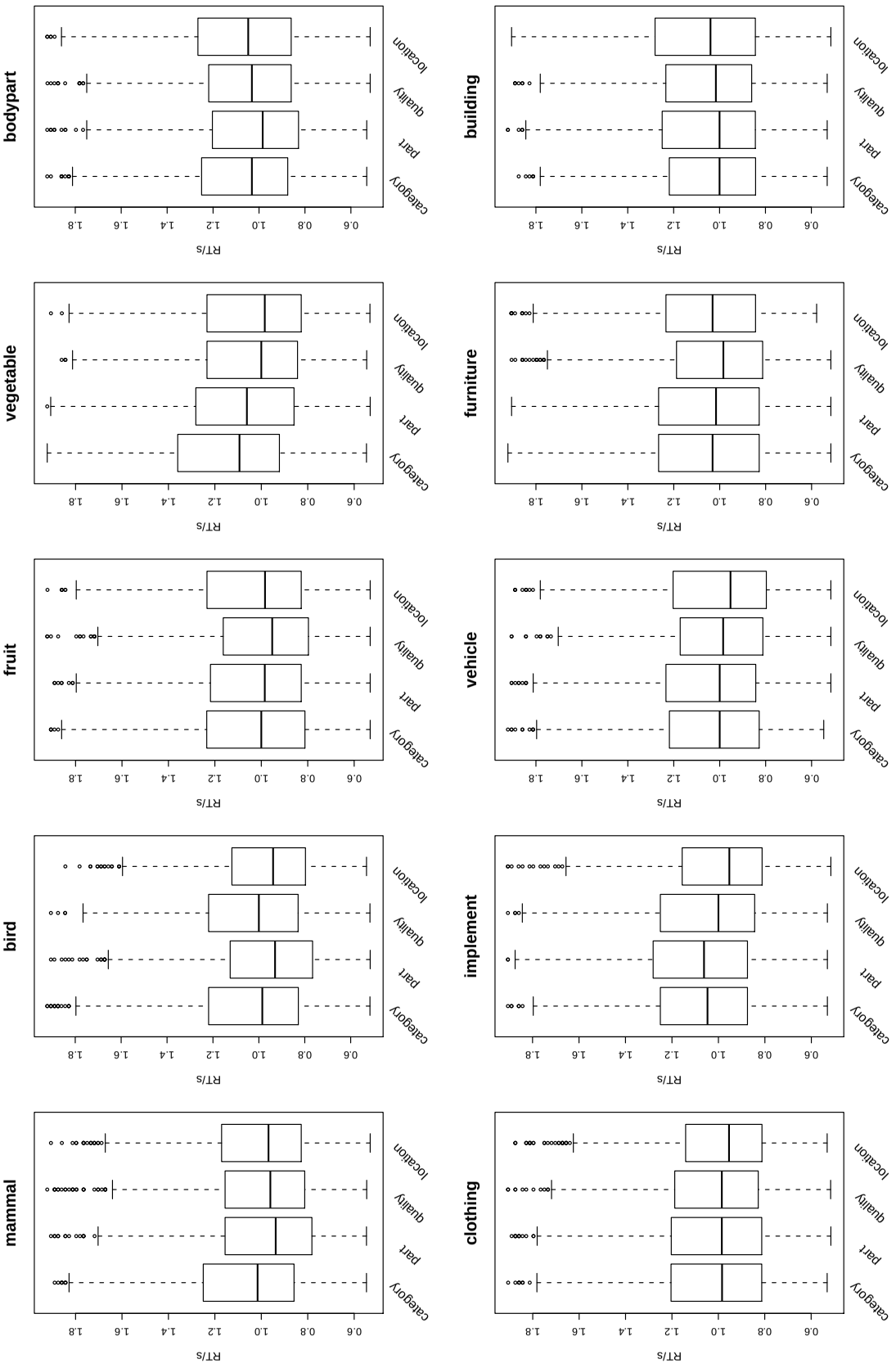

Figure 2.5.: Boxplots for reaction times grouped by concept classes and relation types

almost equally spread — the spreads are wide compared to the little differences in their medians, indicating that the reaction times are not clearly different from each other. For that reason, these differences should be rather looked at as tendencies.

Reaction Time Analysis for Super Classes

Another mixed effect model analysis was conducted, this time replacing the factor *concept class* with *super class* (i. e., broader concept classes, namely, animals, plants, and man-made objects including body parts). Still, significant differences in reaction times showed up for some interactions, but no consistency was observable by comparing these with the results of the production experiment. Details are shown in table C.4 on page 70.

Error Analysis

Next, we conducted a logistic regression analysis where the binary dependent variable was whether the response was correct or not (cf. Baayen, 2008). The same factors as above were used, but leaving out the random effects factor *relation word length* (as including it leads to convergence failure during model estimation). Again, running the analysis with the additional fixed effects factor *language* showed no statistically significant differences between the two language groups. Table C.5 on page 71 shows z-scores and p-values. The z-scores indicate if the probability increases (with a positive sign) or decreases (with a negative sign). The statistical significance is again determined from the corresponding p-values. About half of the interactions (14 out of 27) are significant at the 5 %-level, i. e., half of the combinations of concept classes and relation types were influencing when participants responded correctly.

Introducing the additional fixed effects factor *valid/invalid relation* into this model resulted in a highly significant p-value ($p < 2 \cdot 10^{-16}$). Because of that result, separate models for valid relations and invalid relations were fitted and analysed. Comparing these two models, more interactions with significant p-values and at a lower significance level are observed for the model with only valid relations in the data set (see table C.6 on page 72 and table C.7 on page 73). This conforms to what can be expected for invalid relations given that relation types were not explicitly indicated in this experiment design — there should be less influence of specific concept classes or relation types and less false responses, which leads to a smaller number of less significant p-values. Looking at the data, the number of incorrect responses was about three times higher for valid relations of concepts (8,729, i. e., 32 % of the responses) than for invalid relations (2,961, i. e., 11 % of the responses). Besides this observation, no other patterns are prominent.

Visual Error Analysis

The analysis of the number of incorrect responses given is assessed in the association plot in figure 2.6 on the next page. The data used here comprise all wrong responses to word pairs with a valid relation. They are grouped by concept class and relation

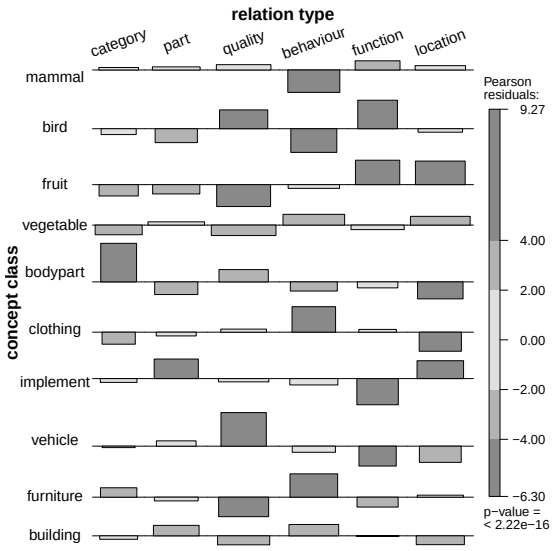

Figure 2.6.: Association plot for error numbers of responses to valid relations

type. The darkness of rectangles represents the significance degree of deviations for the overall distributions of errors. It is based on the p-value of a Pearson residual test. Analogously to the description in section 2.1.3, rectangles above the baseline indicate that participants made more verification errors than expected (marking valid relations as not useful for a description about the concept), and rectangles below the baseline indicate less verification errors than expected.

Looking at the cells in the plot shows, e. g., that for the relation type *quality* and the concept class *fruit* significantly less errors were made than expected, whereas the cell for *quality* and *bird* indicates a significantly high number of errors. Both examples are conform with what was found in these cases in the production experiment: For *fruit*, significantly more properties of the relation type *quality* were produced (which may lead to the assumption that this should trigger less errors when verifying such a relation). The contrary is observed for *bird* and *quality*.

Significant values can be observed for a number of combinations spread over all concept classes and relation types — some make sense when comparing them with the results of the production experiment, whereas others do not seem to be explicable on that basis.

2.2.4. Summary

A perception experiment (using a feature verification method) was carried out and analysed to compare results to the previously conducted production experiment (using a property generation approach). The underlying research question was if differences in participants' performances could be discovered depending on concept class and relation type of the respective word pair stimuli.

Mixed effects models were analysed using reaction time or response correctness as the

dependent variable. Both showed significant differences depending on which concept class the presented concept word belonged to, and which relation type the presented relation word had. This result was restricted to a subset of all possible combinations of concept classes and relation types, with no specific pattern that might be interpreted more generally. Furthermore, some single results of each analysis could be interpreted on the basis of the production experiment, others could not. Analyses with concept super classes, a visual check of the reaction time distributions, and the analysis of the response errors all drew the same picture. In summary, although not as consistent and generalisable as in the production experiment, significant differences in the mental processing of relation types for specific concept classes were found. Between languages, no significant differences were detected. These results do not confirm the results of the production in detail, but they also do not contradict them.

A few modifications in the experiment design might support clearer analysis results. To better compare the results with those of the production experiment, a second experiment could be conducted with only word pairs that appeared in that experiment. The range of reaction times measured was reaching up to the maximum response time allowed. This could mean that there was not sufficient time for the participants to process the word stimuli deeply enough for the intended task and that they had to develop another strategy in some cases.

It might as well be that there are more differences in reaction time distributions which were not detectable as some effects of concept classes and relation types are less prominent. For that case, the collection of a bigger amount of data should give more insight.

The stimuli collection was made on the basis of web corpus data, but appropriate word pairs were hard to find. Collecting them from other sources or from (other) participants is an alternative. Moreover, the Italian stimuli were translations — they were not collected in the corpus using the same method — because the aim was to provide the same concepts in both languages. It seems also reasonable to collect the words separately for each language while trying to adapt to similar familiarity measure values.

A second issue about the stimuli word pairs concerns the relation type. In the experiment presented, participants were not given the type of relation. Some word pairs could be interpreted with a relation type different from the intended one, but which is more prominent to the participant (in particular, invalid word pairs, such as “hand – instrument”, where participants might think that the relation type is *function* instead of the intended *category* relation). To exclude this alternative interpretation, one solution is to indicate the relation type in the experiment — in the present experiment this had not been done in favour of imposing a bigger mental processing load, which was hoped to result in greater differences in reaction times.

The results of the present experiment show that concept classes and the relation types of word pairs have an influence on the mental processing of their semantic relation. Future experiments will show if these tendencies can be confirmed in more detail and if salient relation types can be defined for groups of concept classes as in the production experiment.

3. Cognitively Salient Composite Part Relations

In the area of semantic relation extraction, much research has been done already. However, composite expressions of semantic relations have not been in the focus of these works, so far — in particular, when the targets are cognitively salient semantic relations. To simplify the task for the first attempt in developing an extraction method, this section focuses on *part* relations, in particular those composed of an adjective and a noun. Such concept properties are commonly produced by participants when describing concepts during feature elicitation experiments like the one presented in the previous chapter. In the feature norms collected for the German language (10,010 descriptive phrases in total), of the 1,667 *parts* produced, more than one third (625) were composite *parts*, and 404 were composed of an adjective and a noun. This high proportion motivates our work and is not surprising, given that, for describing a specific concept, one will tend to come up with whatever makes this concept special and distinguishes it from other concepts — which (considering *parts*) sometimes is the *part* itself (elephant: trunk) and sometimes something special about the shape, colour, size, or other attributes of the *part* (elephant: big ears).

The concept–part pairs in the described feature norms (see section 2.1) served on the one hand as input to our algorithm — on the other hand, its output (the set of selected modifiers from the corpus) could be evaluated against those modifiers that were produced by the participants. Furthermore, the bilingual nature of the norms allowed us to tune our algorithm on one language (German) and evaluate its performance on the other (Italian), to assess its cross-lingual generalisation capability.

Assuming that for a given concept its cognitively salient (constitutive) *parts* have already been identified (e. g., applying the whole–part extraction method described by Girju et al., 2006), this section presents the approaches explored for ranking and extracting modifiers of composite *part* relations. The goal is to collect a small, reasonable set of modifiers for each concept–part pair, from which subsequently a human selects the best candidates for the respective purpose. The performance of three different extraction methods are evaluated, adopting the production norms for German and for Italian as the gold standard. Acceptance rate data from a follow-up judgement experiment and a new gold standard set for previously unseen concept–part pairs complete the evaluation set. Eventually, section 3.3 describes a series of failed attempts to improve the performance of the selection algorithm (and thus may be skipped as they are not the core part of this chapter).

The data set for tuning the modifier extraction algorithm and for subsequent evaluation comprises all the concept–modifier–part triples (e. g., *onion*: *brown peel*) produced

by at least one participant, taken from the German and the Italian norms. The German (Italian) speakers described 41 (30) different concepts by using at least one out of 80 (45) different *parts* in combination with one out of 62 (50) different modifiers, totalling to 229 (127) differently combined triples.

3.1. Modifier Selection Based on Corpus Co-Occurrences

Based on the idea that the co-occurrence of words in a text corpus reflects to some extent how strong these words are associated in speakers' minds (cf. Spence and Owens, 1990), the extraction approach described below works on co-occurrence frequencies in the lemmatised and POS-tagged German WaCky[1] web corpus of about 1.2 billion tokens.

Using co-occurrence statistics for words in certain contexts to hypothesise a meaningful connection between the words has a very long tradition in computational linguistics (Church and Hanks, 1990). In this respect, the approach proposed below is not different from common methods to extract and rank collocations, multi-word expressions or semantically related terms (Evert, 2008). From a technical point of view, the innovative aspect is that we do not just look for co-occurrences between two items, but for co-occurrences in the context of a third element, i. e., modifier–part pairs that are related when predicated of a certain concept. The method applied to the extraction of modifier–part pairs when they co-occur with the target concept in a large window is similar to the idea of looking for partially untethered contextual patterns proposed by Garera and Yarowsky (2009), that extract name–pattern–property tuples where the pattern and the property must be adjacent, but the target name is only required to occur in the same sentence.

3.1.1. Modifier–Part Frequencies

Using the CQP[2] tool, corpus frequencies were collected for all co-occurrences of adjectives with those *part* nouns that were produced in the experiment described in section 2.1. A possible gap of up to three tokens between the pair of adjective and noun allowed to extract also adjectives that are not directly adjacent to the nouns in the corpus (but in a sequence of adjectives, for example). For each *part* noun, the five most frequent adjective modifiers from the ranked modifier–part list were selected under the assumption that the preferred usage of these modifiers with the specific *part* indicates the most common attributes which that *part* typically has.

[1] see the WaCky project at URL `http://wacky.sslmit.unibo.it`

[2] Corpus Query Processor (part of the IMS Open Corpus Workbench; see URL `http://cwb.sourceforge.net`)

3.1.2. Modifier–Part Frequencies in Concept Context

The previous method does not necessarily yield generally atypical modifiers that are however typical of a *part* when it is attributed to a specific concept. For example, birds' beaks are typically brown, orange or yellow, but aiming to extract modifiers for a crow's beak, *black* would be one of the desired modifiers — which does not appear at a high frequency rank as a generic beak modifier. The method described so far did not take the concept into account when generating the modifier–part pairs, i. e., for all concepts with a specific *part* the same set of modifiers would be extracted.

To address this issue, a second frequency rank list was prepared in the same manner — with the only difference that the *part* noun had to appear within the context of the concept noun. That way, also modifiers for specific concepts' *parts* that deviate from the most typical *part* modifiers appear at a high rank. However, these data are sparser, which is the reason for using a wide context of 40 sentences (20 sentences before and after the *part*) within which the concept had to occur (i. e., a paragraph-like context size in which the topic, presumably, comprises the concept). Further on, ranked lists of modifier–part pairs that do not take the target concept into account are referred to as contextless lists, and lists within the span of a context as in-context lists.

Due to the already mentioned data sparseness problem, not all modifiers used for a *part* noun in the production norms could be extracted with the latter method, as some of the obvious modifiers for specific parts are just not written about. For these, there is a higher chance that they appear, if at all, in the contextless rank list. For example, *thin bristles* does not appear in the context of *broom*. In the in-context list, 33 % of the 229 triples extracted from the German norms were not found (in the contextless list, only 9 % of the triples are missing after the lacking modifier–part pairs were matched to the appropriate concepts). Additionally, particular concepts, *parts*, or concept–part pairs (within the 40 sentence span) might be missing from the corpus, as well. From the German norms collection, all concepts appeared in the corpus, but one *part*[3], and six concept–part pairs[4] were missing (rare or colloquial nouns). In the evaluation to follow, all the modifiers pertaining to these missing data from the corpus will be counted as positives not found by the algorithm.

The example excerpt in table 3.1 shows modifiers that the current algorithm selected for *bear* and *fur*, using the two frequency rank lists described above. Although in this example many modifiers (thick, dense, soft) are found in both lists, two arguably reasonable modifiers are just in the contextless set (black, long), and one only in the in-context set (white). A disadvantage of selecting modifiers from the in-context rank list is that many modifiers have the same low frequency, but they should nevertheless have differing ranks. In such cases, they were assigned ranks according to alphabetic order of modifiers.

[3] "Löffelohr", a noun–noun compound for *rabbit's ear*

[4] namely, "Fruchtkörper" for *pineapple*, "Pratze" for *bear*, "Reißer" for *broom*, "Plastikteilchen" for *comb*, and "Stingel" for *cherry* and *corn* (*part* nouns for *fruiting body*, *paw*, *bristles*, *teeth*, and *stem*)

	contextless		in concept context	
rank	frequency	modifier	frequency	modifier
1	507	thick	16	thick
2	209	dense	14	white
3	204	soft	11	small
4	185	black	11	soft
5	175	long	9	dense

Table 3.1.: Top five modifiers from frequency rank lists for part *fur* and concept *bear*

3.1.3. Productwise Combination of Frequencies

As an approach to improve performance, the raw frequencies were combined productwise into a new list (for those modifier–part pairs missing in the in-context list, the frequency of the pair in the contextless list was taken alone, instead of multiplying it by zero; i. e., the in-context term was $\max(\text{freq}, 1)$). This achieves a sort of "intersective" effect, where modifiers that are both commonly attributed to the *part* and predicated of it in the context of the target concept are boosted up in the list, according to the intuition that a good modifier should be both plausible for the *part* in general, and typical for the concept at hand.

3.2. Performance Evaluations for Differing Gold Standards

This section reports the evaluation results for the set of modifiers that were yielded for given concept–part pairs by the selection algorithm from rank lists created via three different methods. The evaluation is based on different gold standards adopted: the data produced in the German and Italian norms, concept–modifier–part–triples that were rated most plausible in a judgement experiment, and translated data from English norms that were not in the set when tuning the algorithm.

3.2.1. Production Norms

The feature norms data represented the gold standard for the evaluation of all sets of modifiers chosen by each of the described methods for the given concept–part pairs. Note that, even if a modifier–part pair was produced only once in the feature production norms (e. g., *aeroplane*: "has round windows"), the corresponding concept–modifier–part triple was included in the gold standard — which contains 41 different concepts, 80 different *parts*, and 62 different modifiers, totalling to 229 concept–modifier–part triples. As in the German corpus there are 154,935 adjective–part-noun pairs, the baseline (random guessing) for finding these 229 pairs is approaching 0 (similarly for Italian and the judgement data set).

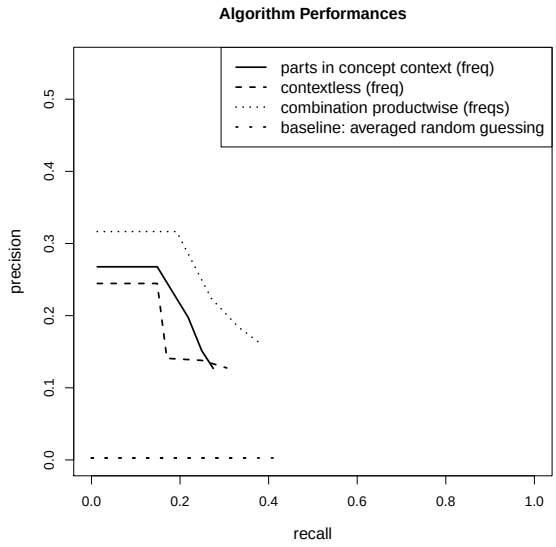

Figure 3.1.: Evaluation on German norms

Figure 3.1 displays the performance of the methods on German in the form of a recall–precision graph. For each rank (1–5), overall recall and interpolated precision values are given for all modifier–part pairs up to this rank. As expected, extracting modifiers of *parts* within a concept context (the in-context list) achieves low recall. In contrast, modifiers that were extracted by querying the corpus for *parts* without considering the concept context have a higher recall. But this method has a lower precision in general. The performance for the method combining frequencies productwise is substantially better. Not only the precision is much higher at all recall levels, but also the maximum recall value is higher than those of the contextless lists, i. e., it was worth combining the complementing information in the two lists. Note that all methods perform distinctively well above the baseline.

Qualitatively analysing the data collected with the described methods did not give definite clues about why some performed not as good as expected. As a comprehensible example, the modifier *short* for *legs* is at rank 5 in the contextless list, but because of the frequent co-occurrence with *monkey* it rises to rank 2 in the productwise combination of these lists. An understandably bad performing example is the modifier *yellow* for the *eyes* of an *owl*: Although it appears in the in-context list at rank 2, it is a quite infrequent modifier for *eyes* in general (i. e., low in the contextless list), and thus it is not contained in the top five modifiers in the productwise combined rank list. For all methods, collected modifiers include undesired words for attributes not describing the *part*, but other, rather situational aspects, e. g., *own*, *left*, *new*, *protecting*, and *famous*. Furthermore, some modifiers in the list are reasonable for the respective concept–part pair, but they are counted as false because they did not occur in the production experiment (that represented the evaluation basis), e. g., for the *blade* of a *sword*, not only *large* is acceptable, but also *long* and *wide*, essentially making the

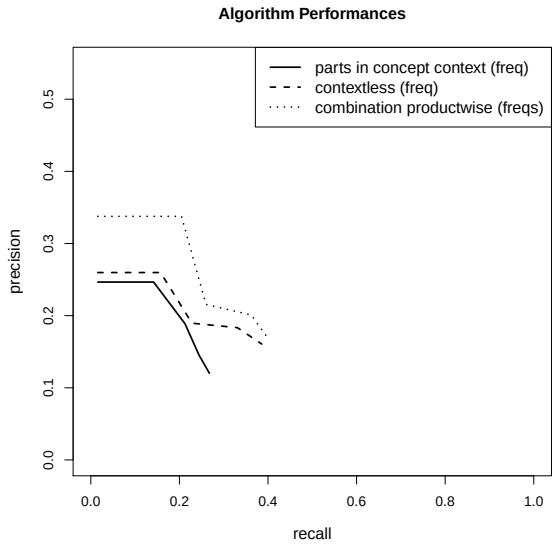

Figure 3.2.: Evaluation on Italian norms

same assertion about the size of the *blade*. This issue is addressed further below by creating a new evaluation standard based on plausibility judgements.

To evaluate the cross-lingual performance of the extraction approach, the Italian norms were explored similarly to the German norms for composite *parts*. The gold standard here comprised 127 triples (from combinations of 30 different concepts, 45 *parts*, and 50 different modifiers). The same methods described above were used to extract modifiers from the Italian WaCky web corpus (more than 1.5 billion tokens), with one difference regarding the query for adjectives near nouns: As in the Italian language adjectives in a noun phrase can be used both before and after the noun (with differences in their meaning), and given that most of them were produced after the noun, the query included all adjectives occurring up to two words from the left of the noun and up to four words to the right.

Figure 3.2 shows the performance curves of the methods for the Italian data. Like in German, the in-context method yields a low recall, in contrast to the method not considering the presence of concepts in context. Again, productwise-combination of frequencies outperforms both of the other methods.

Summarising, the comparison of various corpus-based ranking methods to the feature production norms, both in German and Italian, suggests that composite *parts* produced by participants are best mined in corpora by making use of both general information about typical modifiers of the *parts* (the contextless rank) and more specific information about modifiers that co-occur with the *part* near the target concept. Moreover, it is advantageous to combine the two information sources productwise, which suggests an intersecting effect (the most likely modifiers are both well-attested out of context and seen near the target concept).

By looking at the overall performance, the results seem somewhat underwhelming,

with precision around 20 % at around 30 % recall for the best models in both languages. A natural question at this point is whether the modifiers ranked at the top by the best methods and treated as false positives because they are not in the norms are nevertheless sensible modifiers for the *parts*, or whether they are truly noise. In order to explore this issue, as well as looking at how the methods generalise to concepts that were not in the original norms, a plausibility judgement experiment was set up.

3.2.2. Plausibility Judgements

Focusing from now on the productwise-combination method, the purpose of this judgement experiment was to see which concept–modifier–part triples the majority of participants would rate as acceptable. It allows us to investigate two topics:

- the comparison of what people produce and what they perceive as being a prominent modifier for a concept–part pair (the selection algorithm might actually provide good candidates which were just not produced, as mentioned above), and
- the performance of the best algorithm (productwise-combination) on new concepts that were not in the data set the algorithm was tuned on (by chance this set could be special from other concept–part pairs).

The sets to test were created by first applying the best performing method (productwise-combination) to the concept–part pairs from the German feature norms and to the new pairs translated from the English norms of McRae et al. (2005) for which composite *parts* expressions were produced. From the resulting rank lists, for each concept–part pair the five highest ranked modifiers were selected for the judgement experiment. The test set created from the German norms contained 692 triples, comprising 41 concepts and 71 *parts*; the set with the new pairs (i. e., unseen by the algorithm during tuning) summed up to 318 triples, comprising 45 new concepts and 20 *parts*.

From the unified set of pairs a set of triples was chosen randomly for each of the 46 participants (recruited by e-mail among acquaintances of the author). The triples were presented to participants embedded into a natural-sounding sentence of the form "The [part] of a [concept] is [modifier]". Each participant rated 333 sentences that were presented on separate lines of a text file (this set of sentences presented comprised additional triples which were intended for other purposes — for the current evaluation, a subset of 110 of these from each participant was used, on the average). Participants were instructed to read the sentences as general statements about a concept's *part* and mark them by typing a letter ("w" for wonderful and "d" for dubious — to facilitate one-handed typing and easy memorisation) at the beginning of the line, if they thought it plausible/unlikely that someone used the sentence to explain an aspect of the relevant *part*. In total, 5,525 judgements were collected; each sentence in the set was judged on the average by eight persons.

The performance evaluation was based on the acceptance rate of the participants: Modifiers accepted by a majority of at least 75 % of the raters were considered plausible. Figure 3.3 on the following page shows the recall–precision graph for the productwise-

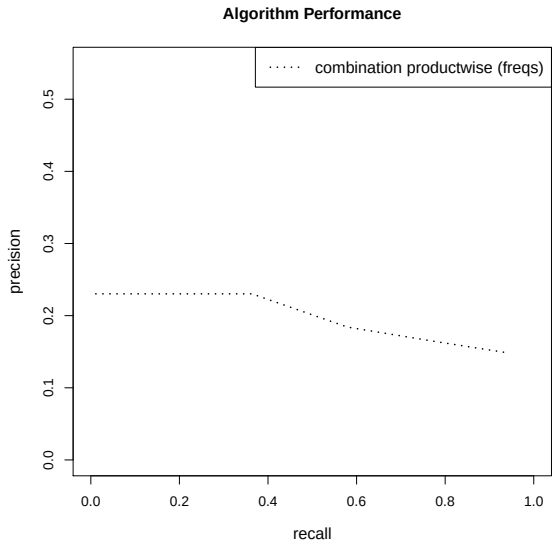

Figure 3.3.: Evaluation on judgements (German)

combination method tested on the concept–part pairs from the German norms. From the 692 triples judged, around 13 % were accepted by the majority of speakers. The precision rate is comparable with the one resulting from the evaluation on the basis of the modifiers produced by participants (highest recall is 1, of course, because all modifiers to be judged were exclusively from the data set selected by the productwise-combination method).

Turning to the qualitative comparison of production and perception, there was a relatively small overlap of triples (46) contrasting with modifiers only produced but not accepted (53), and modifiers accepted but not produced (42). Intuitively, one could have expected that what was produced will be also accepted by the majority of people. Possibly, some participants in the judgement experiment found a few of the triples produced questionable (*goose*: *long beak*) — such triples were in the gold standard because, deliberately, composite *parts* were not excluded even if produced by only one speaker — whereas participants producing *parts* for given concepts probably just did not think of specific *parts* or modifiers (e. g., *aeroplane*: *small windows* and *bear*: *dense fur*). The important fact regarding this difference is, however, that the algorithm presented captures both kinds of modifiers.

3.2.3. Rated Modifiers of New Concept–Part Pairs

As mentioned in the previous section, the stimuli set of the judgement experiment additionally included concept–part pairs that were not taken from the production experiment. That way, the performance could be evaluated on data that was not used during the tuning of the selection algorithm, verifying if the algorithm was overtrained and if the former concept–part pair set of the gold standard was possibly a special

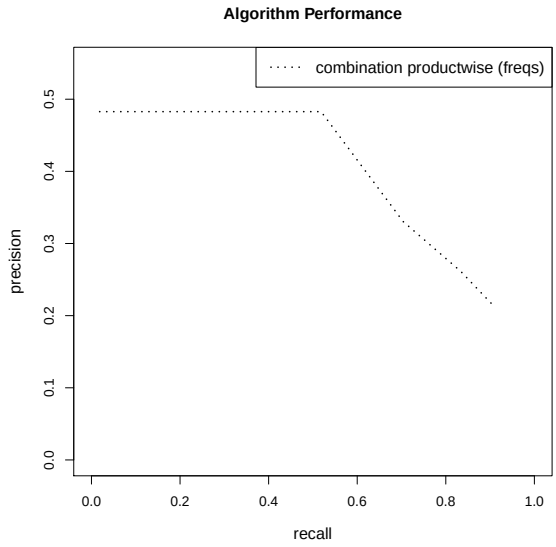

Figure 3.4.: Evaluation of new concepts (German)

group that might have influenced the results. Again, modifiers of the concept–part pairs that were selected from the production experiment data of McRae et al. (2005) and translated into German were only incorporated into the new gold standard, if they were judged as plausible for the corresponding pairs by at least 75 % of the participants.

Figure 3.4 shows the performance for these concept–part pairs that were not in the German norms. Like in the other performance evaluations discussed, the precision value at the highest recall is roughly around 20 % (although slightly higher than for the other evaluations), and precision has notably higher values at lower recall values. To see if there was a significant difference in performance on the old and new concept–part pairs, the distributions of their acceptance rates for the productwise-combination method were compared. A t-test on the acceptance rates resulted in a non-significant p-value of 0.275. That indicates that the presented algorithm generalises well to items that were not in the initial set we originally focused on.

3.3. Further Attempts in Improving Performance

The following collection of attempts aimed to improve the performance of the modifier selection algorithm. Despite the justified ideas of why each of them might be of advantage for the algorithm, none of these was more successful (or at least not essentially better — see table 3.2 on the next page — and much more costly to prepare) than the best method so far, described in section 3.1.

Table 3.2.: Evaluations based on the German production norms for alternatives to rankings according to co-occurrence frequencies (first row). The values presented pertain to the selection of the top five candidates from the rank lists.

Ranking Method	Recall (%)	Precision (%)
Productwise combination of frequencies	43.23	14.45
Productwise combination of frequency logarithms	41.05	13.72
Summed log-rescaled frequencies	35.37	11.83
Productwise combination of log-likelihood values	34.50	11.53
Cosine-distance similarity (default parameters)	1.75	0.58
cosine re-ranking: compare to 300 modifiers	0.44	0.15
cosine re-ranking: compare to 30 modifiers	31.88	10.66
cosine re-ranking: re-rank 500 modifiers	1.75	0.58
cosine re-ranking: re-rank 50 modifiers	4.37	1.46
cosine re-ranking: maximum cosine value	43.23	14.45
cosine re-ranking: average cosine value	38.43	12.85
cosine re-ranking: sum of cosine values	42.36	14.16
cosine re-ranking: compare to log-likelihood list	17.90	5.99
cosine re-ranking: compare to in-context list	28.38	9.49
cosine re-ranking: singular value decomposition matrix	3.06	1.02
cosine re-ranking: multiply by average entropy value	43.67	14.60

3.3.1. Re-Ranking Based on Frequency Transformations

One series of attempts (instead of ordering the modifier rank list according to co-occurrence frequencies) tried several values based on the transformations of the frequencies. This should lead to a re-ranking of modifiers, with more appropriate candidates at the top five positions. Table 3.2 shows the performance values for each method when selecting these candidates. The gold standard for the evaluation comprised the set of produced concept–modifier–part triples from the German production experiment.

Logarithmic Values of Frequencies

The rank lists contained only few modifiers with very high co-occurrence frequencies and a huge set of modifiers with very low frequencies (both in the in-context and in the contextless list). To account for these unbalanced differences of frequencies between high and low ranks, the logarithms of the frequencies of the in-context lists and the contextless lists were taken before combining them productwise. This should prevent modifiers at the highest ranks with a high frequency value in the contextless list to appear at the first rank in the combined list, even though they never appeared in the in-context list (as described, such modifiers were assigned a frequency of 1 in the in-context list).

Table 3.3.: Contingency table scheme defining variable names for observed co-occurrence frequencies of a given modifier–part pair: (*modifier*, *part*)

No. of pairs that contain...	*modifier*	other modifiers	Row totals
part	O_{11}	O_{12}	O_{r1}
other parts	O_{21}	O_{22}	O_{r2}
Column totals	O_{c1}	O_{c2}	N

Summed Log-Rescaled Frequencies

As a variant for combining information from the in-context and the contextless list, the scaled frequencies for the concept–modifier–part triples appearing in both lists were added. Scaling was done because the frequencies in the contextless list are in general much higher than in the in-context list. Furthermore, to account for the fact that at high ranks the difference in frequency between subsequent ranks is much higher than at lower ranks, scaling was done by taking the logarithms of the frequencies: For each concept–modifier–part triple, its frequency logarithm value was divided by the logarithm value of the maximum corpus frequency of all *parts* in the corpus (in the contextless list) or of all concept–part pairs co-occurring within 40 sentences (in the case of the in-context list).

Log-Likelihood Values of Frequencies

This alternative method calculated[5] the log-likelihood association value for each modifier–part pair in the contextless list and ranked the modifiers according to these values. Given the observed co-occurrence frequencies O_{ij} (as defined in table 3.3) and the total frequency of pairs N, for each modifier–part pair the log-likelihood association measure was calculated using the formula

$$\text{log-likelihood} = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}} \tag{3.1}$$

with expected frequencies

$$E_{ij} = \frac{O_i O_j}{N} \tag{3.2}$$

and marginal frequencies

$$O_{ri} = O_{i1} + O_{i2} \tag{3.3}$$

$$O_{cj} = O_{1j} + O_{2j} \tag{3.4}$$

[5] using the UCS toolkit, described at URL `http://www.collocations.de/software.html#UCS`

Log-likelihood weighting should account for typical modifiers which have a low co-occurrence frequency (in the list of pairs) but do generally not occur often in the corpus, and with not many other *parts* — their log-likelihood value will be higher, and so will be their rank (e. g., *two-sided blade* in contrast to *long blade*). Again, the in-context (frequency) values were multiplied with the contextless (log-likelihood) values.

Cosine Distance Similarities

Another attempt to further improve performance is based on the idea that *parts* are described by some specific types of attributes. For example, a *leaf* would be characterised by its shape or consistency (e. g., *long*, *stiff*), whereas for *beak* rather colour should be considered (e. g., *yellow, orange, red*). If an algorithm was able to cluster modifiers for their attribute type and find out which attribute types are in particular important for a specific *part*, those could get a preference in the rank list and be moved towards the top. This way, those modifiers which were low in the frequency-based list get a chance to move to the top. For example, the modifier *black* for *beak*, being perhaps still below rank 5 in the productwise-combined list, should be moved towards the top of the list after realising that *beaks* are often co-occurring with colour modifiers. At the same time, modifiers specifying other attributes types that co-occurred less frequently (e. g., *open*, *hungry*, or *full*) should sink to lower ranks.

To approach this in a simple way, a re-ranking method was used which is supposed to cluster and choose the right cluster of modifiers implicitly: The modifiers in the (productwise-) combined list were tested for their similarity by looking if they co-occur with the same relative frequency with the same set of nouns. In case of high similarity (in this respect) of a modifier to a single other modifier, or if the modifier was similar to a lot of modifiers, it should be re-ranked to a higher position. In more detail, a vector was created for each modifier, denoting its co-occurrence frequencies with each noun in the corpus within a window of four tokens (on the left side of the noun). Random indexing helped to reduce the vector dimensionality from 27,345 to 3,000 elements (cf. Sahlgren, 2005). These vectors served for calculating the cosine distance similarities between modifiers.

Cosine distance similarity is defined by the cosine of the angle between the two vectors $\vec{a}$ and $\vec{b}$ that are pointing to A and B (points in a multi-dimensional space created from frequency information for a pair of modifiers m_A and m_B), which can be calculated from the dot product of these vectors, divided by the Euclidean norms (the vector lengths) of $\vec{a}$ and $\vec{b}$:

$$\text{cosine-distance}(m_A, m_B) = \cos(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|} \tag{3.5}$$

For each of the top 200 modifiers in the combined frequency rank list (covering 84 % of the triples from the German norms), the cosine distance was calculated to each of the top 100 modifiers in the contextless rank list. A constant of 1 was added to each of the computed cosines, thus obtaining a quantity between 1 and 2. The

original combined frequency value was multiplied by this quantity (thus leaving it unchanged when the original cosine was 0, increasing it otherwise). From the re-ranked list resulting from this operation, the algorithm selected, again, the top five modifiers of each concept–part pair. For example, suppose that *black* is among the modifiers of a *crow*'s *beak* in the combined list. The cosine distance similarity of *black* with the top 100 modifiers of *beak* (in any context) is computed, and, for each of these cosines, the original combined value of *black* is multiplied by cosine+1. Since the colour is a common attribute of beaks, the presence of modifiers like *yellow* and *brown*, high on the contextless *beak* list, helps re-ranking *black* high in the *crow*-specific *beak* list. This method was hoped to help out concept-specific *values* (e. g., *black* for *crow*) of *attributes* that are in general typical of a part (*colour* for *beak*).

Parameter Variation for Cosine Re-Ranking

Several parameters for the cosine re-ranking step were modified to investigate the impact on the method performances. Comparisons might be not have been made to a big enough set of modifiers (or they were possibly made to a too large one). Thus, instead of calculating similarities of the modifiers in the rank list to 100 modifiers in the contextless rank list, comparison with a higher number (300) and a lower number (30) was tried. Similarly, the number of modifiers to be re-ranked was varied. The default (re-ranking 200 modifiers) was substituted by the numbers 500 and 50 (while keeping the default value of 100 similarity comparisons for each of these modifiers). Re-ranking more modifiers opens the possibility of lower ranked modifiers to be re-ranked (hopefully, high enough), and a smaller number might help by excluding more inappropriate modifiers.

Next, instead of multiplying the rank list values for each comparison by the modified cosine value (augmented by 1), three alternative methods were tried: First, using only the highest modified cosine to multiply the rank value with, second, using the average of the modified cosines from the comparisons, and third, using the sum of modified cosines. As above, only positive cosines were considered.

In the original cosine-based re-ranking method, the modifiers to be re-ranked were compared to the top 100 modifiers from the contextless list. As this list might not include the most appropriate modifiers for the concept–part pairs, alternative modifier lists served as the basis for the similarity comparisons. One was the contextless list, but ordered for log-likelihood values of modifier–part pairs, which was expected to include the most typical modifiers for (attributes of) *part* nouns. The other alternative list used for comparisons was the in-context list, assumed to include more modifiers that are typical of concept–part pairs.

The matrix used to calculate the cosine distance similarities in the default re-ranking method had been created by applying the random indexing to the full noun–modifier-frequency matrix. Here, the alternative was to perform a singular value decomposition. In contrast to random indexing aiming to just recode the data to reduce matrix dimensions, a singular value decomposition selects those data subsets which account for the main information contained — and does not include the information contained

in the remaining data when building the new matrix. The modifier vectors in the new matrix were reduced to 300 elements.

The qualitative analysis of the cosine re-ranking method showed that it re-ranked mainly those modifiers to high positions that can in general be used with many nouns, e. g., *simple*, *whole*, *own*, and *new*. This is, the cosine re-ranking method worked as expected by pulling up those modifiers which have much in common with many other modifiers (regarding which nouns they co-occur with). Next, the goal was to exclude those modifiers (from ranking them up) that are not highly similar with most of the other modifiers. To achieve this, the value calculated from cosine distance similarities was additionally multiplied by the entropy value for the respective modifier in the re-rank list. Entropy is a measure of uncertainty, or disorder, introduced by Shannon (1948). A low value indicates low disorder (0 being the lowest possible value), whereas higher values indicate higher disorder. Those modifiers that have similarly high cosine distances to most of the comparison modifiers have a low disorder and correspondingly a low entropy value (near 0). Thus, they will be pushed down in the rank list when multiplying the old rank value with the entropy value.

The (average) entropy H for a modifier X to be re-ranked was calculated from the percentage p of each cosine distance value in relation to the sum of all cosine distance values for the modifier according to formula (3.6); note that "log" is the natural logarithm to the base e.

$$H(X) = -\sum_{i=1}^{n} p(x_i) \ \log p(x_i) \tag{3.6}$$

Applying this formula to modifiers with a wide variety in the range of the cosine distances to other modifiers will result in higher values for the average entropy (caused by the high logarithmic values of small cosine value percentages) than for modifiers whose cosine distances do not vary that much as they are equally similar to many other modifiers. The rank values of the former will thus be augmented when being multiplied with the entropy value, and for the latter, values that will be lower or even below zero will sink lower in the re-ranked list.

3.3.2. Larger Gold Standard Set

After various attempts to improve performance, the productwise-combination method still performed best. This section re-evaluates that method, targeting at a more accurate estimation of its performance by extending the gold standard set. The initial gold standard set taken from the production experiment data for evaluating the modifier selection algorithm was arguably small, possibly confounding the performance estimation of the selection algorithm. To remedy this and see if that causes a difference in the evaluation results (and if, in what direction?), a new, larger set of concept–modifier–part triples was gathered to be used for another evaluation.

For this purpose, three different, openly available databases were exploited for concrete concepts and constitutive concrete *parts* with an adjective specifying an aspect

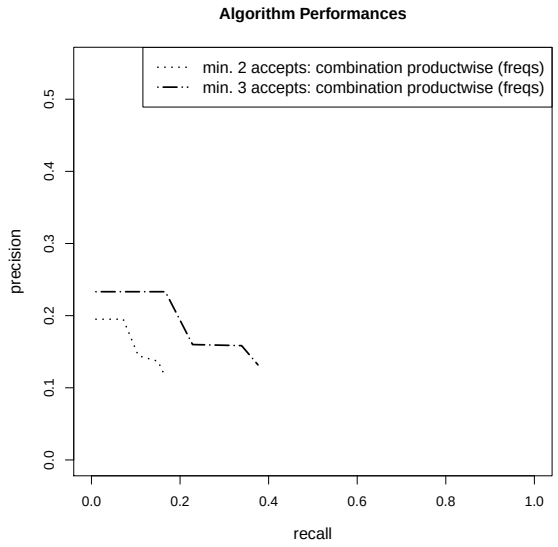

Figure 3.5.: Evaluation on a new, larger data set (translated from English to German)

of the *part* in more detail. Manually scanning the WordNet database for composite expressions for constitutive *parts* of concepts yielded only a small set of concept–modifier–part triples (nine triples of different concepts, modifiers, and *parts*). Looking for appropriate triples in ConceptNet yielded 147 instances, combined from 97 concepts, 74 *parts*, and 68 modifiers. There were 126 different modifier–part pairs in this set. The third source of new triples for the evaluation standard was the Leuven database (described in De Deyne and Storms, 2008), which is a collection of norms data for the Dutch language. Besides the Dutch norms, the authors made also publicly available the English translations of the results of a typicality ratings experiment based on these norms. In the ratings experiment, four participants had to state for each of the features in a set (from the Dutch norms collection) if it was typical (or not) for each of the given concepts. For the new evaluation standard, those triples of modifier, *part*, and concept were collected for which at least two participants rated the feature to be typical for the concept. All the collected triples were subsequently translated from English into their German equivalents. The resulting data extracted from the Leuven database consisted of 955 triples, combined from 193 concepts, 38 *parts*, and 24 modifiers (65 different modifier–part pairs were in this set). When restricting the data from the Leuven database by the condition that at least three participants had to accept the concept–property pairs, the data amounts to 371 triples, combined from 149 concepts, 37 *parts*, and 23 modifiers (61 different modifier–part pairs were in this set).

Figure 3.5 shows the performance curves for both gold standards (including both triples accepted by at least two and by at least three participants). The performance is better for the evaluation based on the gold standard where at least three participants had to accept the composite *part* for a given concept in the data set from the Leuven database. In that case, the performance is similar to the performance based on

production norms — but this Leuven data subset is not much bigger than the set from the production norms. Thus, the question of what difference it would make using a much larger gold standard set is still open. Comparing the two data sets collected from the Leuven database, one observes that they mainly differ only in the number of concepts (i. e., the additional concepts were combined with the same composite *part* relations), which explains why performance is much worse for the larger data set. Unfortunately, no other data sets to extend the evaluation gold standard have been found so far.

3.3.3. Concept-Specific Web Corpus

As an alternative to the (supposedly) wide coverage of text domains of the German WaCky corpus, the following study examined the acquisition of a concept-specific corpus. The idea was to collect text content only from web pages containing a given concept word, and thus, creating a corpus where information about the concept are less sparsely distributed. Additionally, a smaller but concept-specific corpus might be sufficient to yield similar results as a large, unrestricted web corpus.

The corpus was prepared in several steps using the scripts from the BootCaT toolkit[6] (see Baroni and Bernardini, 2004). First, a script collected URL addresses from Yahoo via their Search API. This script was set to return for each search term 100 addresses of web pages written in German. For each request to the Search API, the script received as input (the search term to look for in web pages) one of the words in the concept set that was used in the production experiment. Each concept word was used in its singular and in its (manually created) plural form (in separate requests). From the complete set of returned URLs, duplicates were removed, and a second script downloaded the web pages (only if in html format) and heuristically extracted the content-rich page parts (the raw text). The next script removed duplicate documents from the collection — some documents might just be a copy from a different web page. The whole collection was then converted to the common ISO-8859-1 text encoding standard. After that, the tagger script tokenised the text stream and assigned lemmata and part-of-speech labels. As a last pre-processing step for building the actual corpus from this document collection, very long sentences (more than 150 tokens) were excluded. The final script created the CQP-readable corpus, which comprised around 4.5 million tokens. Compared to this, the German WaCky corpus had around 350 times its size (around 1.6 billion tokens).

As expected, despite the smaller size, concept words appeared relatively often in the concept-specific corpus. On the average, concept words occurred with a relative frequency of 128 parts-per-million (ppm) and a standard deviation of 106 ppm. In contrast, the same concept words appeared in the German WaCky corpus, on the average, with a relative frequency of 2 ppm (s.d. 3 ppm).

But regarding the relative corpus frequencies of the *part* words, the two corpora do not differ that much (concept-specific corpus: average 27 ppm, s.d. 60 ppm; German

[6] retrieved from URL `http://bootcat.sslmit.unibo.it`

WaCky corpus: average 18 ppm, s.d. 37 ppm). The relative frequency data speak even more in favour of the WaCky corpus when looking at the occurrences of modifier–part pairs (queries were composed of the same elements, allowing the modifier to appear within four tokens before the *part* noun). Many pairs (61 %) were not even found once in the concept-specific corpus, whereas in the WaCky corpus only 8 % of the pairs were missing. As these first analyses indicated that much data from the evaluation set were missing in the concept-specific corpus, whereas in the WaCky corpus already more of these data were found, the modifier selection algorithm was not run on the concept-specific corpus data.

In summary, the attempt of building a small concept-corpus in a simple way for the purpose of extracting composite *part* relations for these concepts did not have an advantage over the large WaCky web corpus (although that was not built with the aim to contain information about specific concepts). One improvement might be to aim for a corpus with more tokens than presented here, as the WaCky corpus was still 350 times the size of the concept-specific corpus.

A further (but less promising) attempt to improve corpus coverage of the evaluation set data could be to collect web pages containing word forms of the *full* set of inflected concept words — the corpus described above was created by only searching for pages containing (nominative) singular and plural concept word forms. Additionally, one would intuitively try and also collect web pages containing both concepts and their *part* nouns on the same page. This would make the corpus preparation more complex (considering all combinations of concept and *part* word forms), although it would be feasible. But foremost, this approach would only be reasonable for this very task of finding salient modifiers for given concept–part pairs. Regarding the future goal of identifying and extracting first the *part* relations for a given concept, and finding the best modifier candidates for these, the described approach would lead to a circular problem: One would need these *part* nouns first to build an appropriate corpus for the task. How to build a concept-specific corpus (that is rich in information about the concepts) without requiring beforehand to have a set of words semantically related to the concepts is a possible topic for a separate study.

3.3.4. Ranking Based on Web Search Page Hits

Possibly, the web corpus that was used in the evaluation described in section 3.2.1 included an unbalanced set of selected web pages that do not represent accurately which composite *part* relations are used prominently by native speakers. This follow-up study is a nearly identical repetition of the previous analysis of the German data, but, in contrast to counting occurrences in a web corpus, a larger amount of web texts was searched via a web search service application. The goal was to see if this approach lead to higher precision and recall values (promoting the use of a bigger corpus) or if results were similar or worse (in which case the future focus should be on a more elaborated extraction and ranking approach while relying on the currently used corpus).

Among the currently most well-known web search service APIs (application program-

ming interfaces), namely, Yahoo BOSS[7], Microsoft Bing[8], and Google AJAX[9], Yahoo's API provided the most convenient way to set up, and according to the documentation it served best for the purpose of this study given the query syntax possibilities.

Query Adaptations to Web Search

Analogously to the previous investigation on the web corpus, the procedure in this web-search-based study was to look up co-occurrence frequencies of modifier–part pairs within and out-of concept context — in a manner compatible to the web corpus approach. To achieve this, minor adaptations to the querying procedure were necessary — web search services operate on (indexed) raw text data, and as such the query syntax does not facilitate the usage of part-of-speech tags, lemmata, and sentence boundaries (for the wide concept context) that the work on the web corpus had benefited from. Instead, word forms have to be given, and these can be intermixed with wildcards for words and be restricted to appear within a phrase (otherwise they may appear anywhere in the same website text). Considering these differences in the query language, the following paragraphs describe the procedure used for the web-search-based ranking of modifier candidates.

As a web search service does not support looking for words with a specific part-of-speech tag and returning a list of corresponding words (which had been possible in the work on the WaCky web corpus), the idea was to search specifically for the good modifier candidates defined in the evaluation standard and the set of modifiers that had been extracted from the WaCky web corpus (see section 3.1) but were not in the evaluation standard. That way, the selection algorithm could choose the best modifiers from a large set of candidates. Furthermore, it facilitates to accurately compare the performance of this web search approach to that of the (WaCky) web corpus approach as both had to rank the very same set of modifiers. However, there might be modifiers in the world wide web that were not in the WaCky web corpus (and as such these were not queried from the world wide web according to the procedure just described), and thus were not in the list made available to the selection algorithm. In case these missing modifiers (that are neither in the WaCky corpus nor in the evaluation set) have high frequencies in the web and would be included in the ranked modifier lists, they would rank at high positions and thus lead to a worse performance measure (although it is improbable that high-frequency modifiers in the web do not occur in the WaCky web corpus at all). In the opposite case, assuming those missing modifiers (that were not in the query set but are present in the web) have low occurrence frequencies and would be included in the ranked modifier lists, they would rank at low positions (i. e., at least not within the top five ranks) and thus would not change the performance results. In conclusion, possibly missing modifiers from the rank lists that can be found in the web would not improve the algorithm performance, anyways; it could be just worse than evaluated here.

[7] see URL `http://developer.yahoo.com/search/boss`
[8] see URL `http://msdn.microsoft.com/en-us/library/dd900818.aspx`
[9] see URL `http://code.google.com/apis/ajaxsearch`

The set of triples to be used in the queries originated from three rank lists generated from the German WaCky corpus — in-context, contextless, and combined in-context and contextless. For each concept–part pair therein, the top 100 modifiers were collected. The resulting set of unified, unique triples included approximately 89 % of the target triples (produced in the experiment and used as evaluation standard) that were also found in the web corpus.

In a web search service, lemmata can not be used in the queries to find all corresponding word forms. Thus, a second adaptation was to create a set of queries for each concept–modifier–part triple that included all combinations of inflectional word forms of the word triples. For each of the 50 concepts in the set, the lemma form as it was used in the queries on the web corpus was manually augmented with the unique set of suffixes for the German cases (nominative, genitive, dative, accusative) in their plural and singular forms (considering alternative word stems) to constitute the corresponding inflected word forms. For example, the German word *Apfel* ("apple") can appear in text in one of the following word forms: *Apfel*, *Apfel-s*, *Äpfel*, *Äpfel-n*. Analogously, for each *part* word of the modifier–part pairs from the production experiment, a set of modifier–part suffix pairs was manually collected that should yield all possible inflectional combinations of modifiers with the respective part word.

This resulted in the 16 different suffix pair sets shown in table 3.4 on the following page. Again, in search queries only the subset of unique word form pairs created from such a suffix pair set was used. So, for example, for the *part* word *Arm* with the modifier *lang*, the following set was used to search the web for all corresponding word forms (given that for *Arm* the suffix pair set number 7 in table 3.4 had been identified to map to its appropriate set of inflected word forms): *lang-e Arm*, *lang-en Arm-s*, *lang-en Arm*, *lang-em Arm*, *lang-er Arm*, *lang-en Arm-e*, *lang-en Arm-en*, *lang-e Arm-e*.

Note that some modifier word stems are not the same as their lemmata (e. g., *sauer* has the word stem *saur-*), and that some modifiers that were aimed to be extracted in their lemma forms from the WaCky corpus were in fact inflected word forms. Both cases were not corrected as they were estimated to occur rarely (after a visual check of a data subset) and the manual correction work would have been disproportionally high. This inaccuracy would actually lead to a better performance in case that there were more of these wrong word forms in the data than expected. That is because queries including wrong word forms produce low or zero occurrence numbers leading to a low ranking and thus would not be selected by the algorithm — which they should not be, anyways, as they were extracted from the WaCky corpus and not from the production experiment data set that defined the gold standard (this data set was prepared manually for the web search queries, whereas the word forms for all the additional modifiers from the WaCky corpus were generated automatically using the suffix pair sets).

A last query adaptation concerns the concept context span. The web search service facilitates looking for two words either co-occurring with a defined number of word tokens between them, or appearing within the same text of a website — but not within the context window of 20 sentences as it was defined in the analysis in the previous

Table 3.4.: Inflection suffix pairs used for generating all word forms for a given modifier–part (format: *modifier-suffix*|*part-suffix*). Where no suffix is specified (indicated with the symbol "␣"), the lemma form is sufficient. The German cases covered are nominative (Nom), genitive (Gen), dative (Dat). Using the modifier–part word form pairs in text without a definite article often requires different modifier-suffixes (Nom2, Dat2). The suffix pair sets 6 and 11 have no plural forms.

	Singular					Plural				
Set	Nom	Gen	Dat	Dat2	Nom2	Nom	Gen	Dat	Dat2	Nom2
1	e|␣	en|␣	en|␣	er|␣	e|␣	en|e	en|e	en|en	en|en	e|e
2	e|␣	en|␣	en|␣	er|␣	e|␣	en|en	en|en	en|en	en|en	e|en
3	e|␣	en|␣	en|␣	er|␣	e|␣	en|n	en|n	en|n	en|n	e|n
4	e|␣	en|es	en|␣	em|␣	er|␣	en|e	en|e	en|en	en|en	e|e
5	e|␣	en|es	en|␣	em|␣	es|␣	en|e	en|e	en|en	en|en	e|e
6	e|␣	en|s	en|␣	em|␣	er|␣	ø	ø	ø	ø	ø
7	e|␣	en|s	en|␣	em|␣	er|␣	en|e	en|e	en|en	en|en	e|e
8	e|␣	en|s	en|␣	em|␣	er|␣	en|␣	en|␣	en|␣	en|␣	e|␣
9	e|␣	en|s	en|␣	em|␣	er|␣	en|␣	en|␣	en|n	en|n	e|␣
10	e|␣	en|s	en|␣	em|␣	er|␣	en|n	en|n	en|n	en|n	e|n
11	e|␣	en|s	en|␣	em|␣	es|␣	ø	ø	ø	ø	ø
12	e|␣	en|s	en|␣	em|␣	es|␣	en|e	en|e	en|en	en|en	e|e
13	e|␣	en|s	en|␣	em|␣	es|␣	en|␣	en|␣	en|␣	en|␣	e|␣
14	e|␣	en|s	en|␣	em|␣	es|␣	en|en	en|en	en|en	en|en	e|en
15	e|␣	en|s	en|␣	em|␣	es|␣	en|␣	en|␣	en|n	en|n	e|␣
16	e|␣	en|s	en|␣	em|␣	es|␣	en|n	en|n	en|n	en|n	e|n

section between concept and *part* word. The adaptation here was to widen the context from 20 sentences to the whole website.

Procedure

The web search was conducted for the concept–modifier–part triples from the production experiment and from the set of additional triples described above both without considering the concept and in concept context. A script sent the queries and extracted from the query result pages the number of "deephits" (estimated number of web pages that match the query) and "totalhits" (estimated number of web pages excluding web page doubles).

In the out-of-context queries, each pair of modifier–part word forms was used with 0–3 wildcards (for non-specified words, symbol: `*`) in between to allow for the modifier to appear within the 4-word window that was the constraint in the previous study. For example, for *lange Arm* the separate queries `"lange Arm"`, `"lange * Arm"`, `"lange * * Arm"`, and `"lange * * * Arm"` were generated (quotes restrict the words between them to appear in that exact group and order on the web page to be counted). The result numbers were then summed up to have the totals for each modifier–part pair.

In the in-context queries, each query of the set just described was extended with all word forms of the concept corresponding to the respective *part*. For example, from the query `"lange * Arm"` and the concept *Affe*, the queries `"lange * Arm" Affe`, `"lange * Arm" Affen`, and `"lange * Arm" Affens` were used to search the web.

Evaluation

For the performance analysis, a rank list for modifier–part pairs in concept context, out of concept context, and a rank list combining these two productwise was created, analogously to how that was done in the previous study on the WaCky corpus. The rank lists were based on the deephits numbers; a second set of rank lists was based on the totalhits numbers.

Figure 3.6 on the next page displays the performance curves of the productwise-combination method for both the rank list based on deephits and the rank list based on totalhits. The two curves are very similar, and they are remarkably worse than performances based on the co-occurrence frequencies in the WaCky corpus.

Counterintuitively, running the selection algorithm on a much larger amount of data available from the web did not improve the performance results, but they were even worse. Besides the minor difference that estimated counts of web page hits were retrieved, rather than overall frequencies, various inconsistencies of the returned counts were discovered: Different points in time of the request for the same query, requesting the query via the browser instead of using a script, or querying the Yahoo Search site instead of using the BOSS API all resulted in different web page hit numbers. The arbitrariness of search engine counts was already addressed by Kilgarriff (2007).

Furthermore, the lower performance of this web search study as compared to the corpus-based study described in section 3.2.1 is similar to what Lapata and Keller

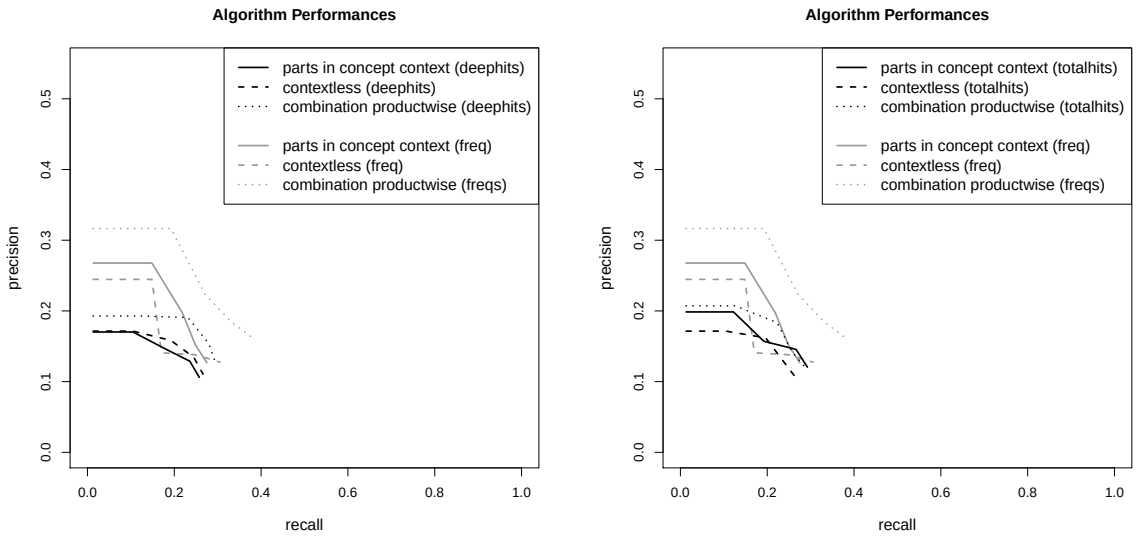

Figure 3.6.: Evaluation based on web page hits (left: "deephits", right: "totalhits") compared to performance of the same methods based on WaCky corpus co-occurrence frequencies

(2004) discovered. When performing various tasks using only counts, web-based models were significantly better than corpus-based models. However, the performance of web-based models (their advantage being vast amounts of available data) was surpassed by the performance of corpus-based models when incorporating linguistic information (available in corpora) into the model. As a conclusion, they proposed to use web-based models as an evaluation baseline. Following this suggestion, the corpus-based system described in section 3.2.1 performs significantly better than the web baseline determined in figure 3.6. Nevertheless, Lapata and Keller (2005), hypothesise that in general, at generation tasks (e. g., ordering of prenominal adjectives, in contrast to analysis tasks, such as compound bracketing), web-based models outperform corpus-based models. This was not the case in the current study discussed.

3.4. Summary

Extracting cognitively salient modifiers for given concept–part pairs is not a trivial task. Several approaches were investigated, where from corpus-frequency-based rank lists the top five modifiers were selected for each concept–part pair. The best method had a precision of 14 % at a recall of 43 %, and it simply combined the information of modifiers appearing together with the *part* noun in and out of concept context by multiplying the respective raw co-occurrence frequencies. More elaborated attempts to improve this performance were not successful. However, Barsalou (1993, p. 32) mentions a "surprising representational flexibility" of participants in a definition experiment. On the average, only 66 % of those features overlapped that were produced by the same

participant in two sessions with two weeks time in between. Considering this variability, the performance of the presented selection algorithm is better than at first thought. Furthermore, as a consequence of the general lack of data publicly available that could be used for the evaluation of the extraction method (conducting another experiment for that case would have been too costly), only an arguably small evaluation gold standard could be set. For this reason, achieved performance results might not be very accurate (i. e., the actual performance is possibly worse or better than expected from this small-scale evaluation).

4. Conclusion

This research empirically investigated the cognitive salience of semantic relations for a set of concrete basic-level concepts in a cross-lingual, parallel study of the two target languages German and Italian. As a result in the production experiment, features generated by native-speaking participants showed similar patterns across languages, despite the observed differences in verbalisations: Depending on the concept class, particular semantic relation types were more or less prominently produced for given concepts. When testing the perception of the production data of the above experiment by means of a feature verification experiment, results were not that consistent as to exactly confirm the first experiment. Nevertheless, concept classes and semantic relation types had a statistically significant influence on the mental processing of the semantic relations between concepts and features presented as word pairs.

The outcome of these two behavioural experiments suggests the following procedure for automatically acquiring cognitively salient semantic relations: Once the class of a given concept is known, the extraction should focus on those semantic relation types that were found to be prominently represented in the behavioural data for that concept class. However, the granularity of concept classes has to be defined. Moreover, the stimuli set will have to be expanded to include, e. g., abstract concepts — although we hope to mine some abstract concept classes on the basis of the properties of our concept set (colours, for example, could be characterised by the concrete objects of which they are typical).

The second investigation concerned the development of methods for the extraction of cognitively salient relations from corpora. The focus was on composite (adjective–noun) expressions for *part* relations, which are frequently used in concept descriptions and lexical resources but have not been addressed in other studies, yet. Assuming that *part* nouns had been already identified for a given concept by existing methods, the approach was to rank adjectival modifiers based on corpus frequencies and select the top five candidates. The best method (evaluated first on the German production experiment data) combined the information of modifiers co-occurring with the *part* noun on the one hand, and the information of this modifier–part pair co-occurring in the wide context of the concept word (by simply multiplying the respective occurrence frequencies).

The performance of about 14 % precision at around 43 % recall is remarkable considering the variability in concept descriptions produced by any speaker for a specific concept and the fact that the baseline of randomly selecting modifiers for part–concept pairs approximates 0. This performance was robust when evaluating the method on a different language (Italian), on formerly unseen concepts, and on extracted modifiers rated by participants as plausible or implausible. Interestingly, a qualitative analysis

showed that modifiers produced and modifiers rated as plausible (i. e., perceived) did not have a large overlap. This means that the algorithm is capable of collecting both prominently produced and saliently perceived modifiers (and with the same performance).

As only a small data set was appropriate and available for the method evaluation, the performance values do not represent a reliably accurate measure estimation. To improve this situation, apart from collecting a larger set of semantic norms, the currently best method could be adapted to extract also numerals as permissible modifiers (so far, the target only comprised adjectives), as in "*four* wheels". A further extension for the generation of human-like concept properties would be to train an algorithm to decide if the *part* relation should include a modifier at all — or if the *part* noun alone is sufficient as the *part* property of a specific concept (cf. *big ears* vs. *trunk* for the concept *elephant*).

The work presented here provides, on the one hand, new findings on the cross-lingual nature of feature-based concept representations. On the other hand, the empirical study is complemented with first approximations towards the automatic extraction of cognitively salient relations from corpora. In particular, we focused on the extraction of composite expressions for constitutive *parts* of concepts, which is a new topic and worth exploring. Applications for which such automatic extraction methods for cognitively salient semantic relations are profitable include pedagogical lexicographic projects (extending language learner dictionaries or generating vocabulary word lists) and research on cognitive concept processing (computational generation of larger feature norms and possibly building models of human-like behaviour).

Finally, this study once more promotes interdisciplinary co-operations in the general research fields of psychology, linguistics, and information science, with the alluring prospect of improving mutual understanding and fruitful joint projects advancing science in every single one of these fields.

A. Semantic Relation Types

Table A.1.: 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 five groups: entity properties (e), taxonomic categories (c), situational properties (s), introspective properties (i), and miscellaneous (m).

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

... continued on next page.

Code	Definition	Example	Lang	No.	%
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
			IT	129	1.51
io	(cognitive) operation	sword — like a long knife	DE	195	1.95
			IT	64	0.75
ic	contingency	aeroplane — 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

... continued on next page.

Code	Definition	Example	Lang	No.	%
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	aeroplane — flies upwards	DE	1	0.01
			IT	0	0.00

B. Perception Experiment Stimuli

Table B.1.: The set of stimuli word pairs (translated into English) used in the perception experiment — 50 concepts from 10 concept classes and the corresponding (valid or invalid) semantically related words of six different relation types. Note that in several cases the meanings could not be captured accurately in the translation.

class	concept	(in-) valid	semantic relation type					
			category	part	quality	behaviour	function	location
bird	duck	v	water bird	beak	small	swims	roasted	pond
		iv	vegetable	arm	yellow	washes	doused	ocean
bird	eagle	v	raptor	claws	brown	flies	hunt	mountain
		iv	game	toe	golden	converts	carried	hotel
bird	goose	v	poultry	feather	fat	quacks	slaughtered	shed
		iv	songbird	horn	short	claps	guards	table
bird	owl	v	bird of prey	eye	calm	nocturnal	hatches	forest
		iv	insect	ditch	high	bakes	protection	hospital
bird	pecker	v	bird	beak	colourful	knocks	nests	tree
		iv	power animal	hand	big	calls	stores	court
bodypart	arm	v	bodypart	muscle	warm	swings	hold	torso
		iv	clothing	eye	white	jumps	drive	floor

... continued on next page.

class	concept	(in-) valid	semantic relation type					
			category	part	quality	behaviour	function	location
bodypart	eye	v	sense organ	lens	oval	waters	see	face
		iv	item	tooth	dark	falls	hold	sky
bodypart	hand	v	limb	finger	flexible	trembles	grasp	arm
		iv	instrument	lip	invisible	lies	stand	head
bodypart	leg	v	extremity	hairy	bent	move	stand	lower body
		iv	machine	ear	black	screams	see	torso
bodypart	nose	v	organ	hair	moist	running	smell	face
		iv	fruit	tongue	bright	rubs	bite	neck
building	bridge	v	construction	rail	arc-shaped	swings	cross	river
		iv	room	chimney	long	runs	sleep	church
building	cottage	v	shelter	clay	tiny	hosts	sleep over	mountains
		iv	cattle	sheet	gloss	crawls	cooked	branch
building	garage	v	construction	wall	dark	protects	park	underground
		iv	gardening tool	corridor	high	drives	count	kitchen
building	house	v	building	mural	solid	decays	live	village
		iv	animal	flame	rough	flaps	feel	field
building	tower	v	structure	concrete	huge	protrudes	work	city
		iv	nutrient	clay	wide	grows	shoot	mountain
clothing	belt	v	accessory	leather	black	decorates	fasten	trousers
		iv	material	hood	deep	clacks	cut	street

... continued on next page.

class	concept	(in-) valid	semantic relation type					
			category	part	quality	behaviour	function	location
clothing	dress	v	garment	cloth	elegant	hangs	put on	body
		iv	furniture	hair	wet	steals	eaten	house
clothing	pullover	v	apparel	sleeve	thick	itches	slip on	wardrobe
		iv	luggage	cushion	narrow	hangs	sleep	face
clothing	scarf	v	accessory	wool	fleecy	flaps	warms	neck
		iv	candy	collar	wide	greens	wrap	hand
clothing	shoe	v	clothes	seam	flexible	pressures	put on	foot
		iv	furniture	blood	slim	sticks	feel	face
fruit	apple	v	fruit	core	green	falls	bitten	box
		iv	vegetable	nose	silver	pushes	drink	cask
fruit	banana	v	subtropical fruit	peel	curved	grows	slip	palm tree
		iv	citrus fruit	tuft	red	flies	walk	cage
fruit	cherry	v	berry fruit	pit	red	ripens	garnish	garden
		iv	flower	marzipan	cool	blooms	hoe	refrigerator
fruit	peach	v	plant	vitamins	ball	smells	dessert	bush
		iv	berry	rail	savoury	bakes	lift	nose
fruit	pear	v	stone fruit	stem	brown	hangs	picked off	tree
		iv	vessel	wood	warm	climbs	cover	pan
furniture	bed	v	furniture	blanket	cosy	creaks	sleep	bedroom
		iv	toy	hand	round	rings	throw	bathroom

. . . continued on next page.

class	concept	(in-) valid	semantic relation type					
			category	part	quality	behaviour	function	location
furniture	chair	v	furniture	backrest	hard	shaky	sit	room
		iv	sports equipment	knee	calm	flounces	saw	sky
furniture	cupboard	v	piece of furniture	drawer	solid	cramps	store	corridor
		iv	clothing	window	straight	looks	iron	parking
furniture	sofa	v	seat	bolster	cushy	inviting	rest	apartment
		iv	drink	cardboard	quiet	narrates	drink	stage
furniture	table	v	item	board	flat	shakes	breakfast	restaurant
		iv	drink	elbow	ball	dances	thrown	chest
implement	fridge	v	electric appliance	door	spacious	cools	store	house
		iv	life form	gelatin	soft	kneads	cover	automobile
implement	knife	v	object	handle	sharp	injures	cut	kitchen
		iv	spice	pulp	heavy	eats	drill	tree
implement	mug	v	vessel	bottom	hollow	tips over	drink	table
		iv	household utensil	leaf	hot	steams	talk	sea
implement	paintbrush	v	implement	wood	wide	drips	paint	bucket
		iv	material	cork	pointed	cleans	pour	oven
implement	pencil	v	utensil	lead	thin	writes	draw	pencil case
		iv	tool	feather	flat	falls	cut out	bottle
mammal	cat	v	mammal	paws	soft	sneaks	petting	backyard
		iv	small animal	hand	hot	barks	write	aquarium

... continued on next page.

class	concept	(in-) valid	semantic relation type					
			category	part	quality	behaviour	function	location
mammal	cow	v	cattle	horn	smooth	chews	milked	alp
		iv	poultry	ice	lilac	stops	paint	stove
mammal	dog	v	vertebrate	muzzle	muscular	sniffs	defends	meadow
		iv	object	door	hard	observed	think	oven
mammal	donkey	v	hoofed animal	legs	grey	stubborn	load	shed
		iv	vehicle	hose	green	blows	stay overnight	barrel
mammal	monkey	v	mammal	fur	human-like	climbs	presented	jungle
		iv	bird	stone	blue	quacks	sit	hell
vegetable	bean	v	groceries	husk	green	overgrows	cultivated	can
		iv	cereal	pit	thick	itches	load	train
vegetable	carrot	v	root vegetable	carotene	long	cracks	cooked	patch
		iv	fruit	fat	skewed	melts	pray	sieve
vegetable	olive	v	crop plant	leaf	purple	decays	flavour	pizza
		iv	seafood	meat	cold	cleans	stew	mill
vegetable	potato	v	nourishment	sprout	yellow	rolls	baked	cellar
		iv	pasta	grain	thin	fades	put off	moon
vegetable	pumpkin	v	twine plant	pulp	smooth	thrives	harvested	ground
		iv	legume	lid	sharp	glows	stand	pot
vehicle	aeroplane	v	machine	jet engine	loud	flies	travel	sky
		iv	tool	belly	small	walks	walk	tower

. . . continued on next page.

class	concept	(in-) valid	semantic relation type					
			category	part	quality	behaviour	function	location
vehicle	bus	v	utility vehicle	wheel	heavy	stops	transport	city
		iv	building	gate	quiet	climbs	cross	air
vehicle	motorbike	v	two-wheel vehicle	tank	light	roars	drive	roadway
		iv	car	trunk	fat	swings	live	market
vehicle	ship	v	means of traffic	engine	big	swims	transport	sea
		iv	animal	wheel	light	lands	dive	mountain
vehicle	truck	v	vehicle	brake	stinks	rolls	carry	highway
		iv	groceries	sand	rickety	injures	run	can

C. Mixed Effects Models Results

Table C.1.: Model fit for all correct responses

factor interactions			
relation type	concept class	$\|t\| > 2$	t-value
category	mammal	*	2.80
	bird		-0.27
	fruit	*	2.47
	vegetable	*	3.67
	bodypart		0.07
	clothing		0.63
	implement		1.13
	vehicle	*	2.08
	furniture	*	2.24
part	mammal		-1.94
	bird	*	-5.35
	fruit		-0.21
	vegetable		-0.35
	bodypart	*	-2.44
	clothing		-0.64
	implement		0.50
	vehicle		0.35
	furniture		-0.24
location	mammal		-1.72
	bird	*	-5.25
	fruit		0.23
	vegetable		-1.50
	bodypart		-0.47
	clothing	*	-2.20
	implement	*	-3.34
	vehicle		-0.67
	furniture		-0.80

Table C.2.: Model fit for all correct responses to valid relations

factor interactions			
relation type	concept class	$\|t\| > 2$	t-value
category	mammal		-0.25
	bird		-1.92
	fruit		0.30
	vegetable		1.79
	bodypart		0.35
	clothing	*	2.54
	implement	*	2.07
	vehicle		1.52
	furniture	*	4.18
part	mammal	*	-5.33
	bird	*	-5.23
	fruit		-0.02
	vegetable		-1.64
	bodypart	*	-4.62
	clothing		-1.12
	implement		-0.64
	vehicle		-1.89
	furniture		-0.37
location	mammal	*	-3.03
	bird	*	-3.05
	fruit		1.44
	vegetable		0.87
	bodypart	*	-3.16
	clothing		-0.03
	implement		-0.09
	vehicle		-1.67
	furniture		0.01

Table C.3.: Model fit for all correct responses to invalid relations

factor interactions			
relation type	concept class	$\|t\| > 2$	t-value
category	mammal	*	2.32
	bird		1.15
	fruit	*	3.47
	vegetable	*	2.83
	bodypart		-1.22
	clothing		-1.68
	implement		-0.63
	vehicle		0.58
	furniture		-1.41
part	mammal		0.65
	bird	*	-2.91
	fruit		-0.30
	vegetable		0.32
	bodypart		0.34
	clothing		0.38
	implement		1.39
	vehicle		1.50
	furniture		-0.45
location	mammal		-0.38
	bird	*	-3.97
	fruit		0.39
	vegetable	*	-2.28
	bodypart	*	2.09
	clothing	*	-2.79
	implement	*	-3.44
	vehicle		0.63
	furniture		-1.03

Table C.4.: Model fit with concept super classes as an independent variable

factor interactions			
relation type	concept super class	$\|t\| > 2$	t-value
category	animal		0.53
	plant	$*$	3.35
part	animal	$*$	-5.49
	plant		0.27
location	animal	$*$	-3.83
	plant		1.07

Table C.5.: Fit for binary model

factor interactions				
relation type	concept class	$p < 0.05$	z-score	p-value
category	mammal		-0.710	0.477988
	bird	$*$	3.760	0.000170
	fruit		-1.355	0.175452
	vegetable		-0.890	0.373331
	bodypart		-1.194	0.232367
	clothing	$*$	3.329	0.000870
	implement		-0.552	0.580630
	vehicle	$*$	2.885	0.003920
	furniture	$*$	-3.053	0.002267
part	mammal		1.392	0.163777
	bird	$*$	6.423	$1.33 \cdot 10^{-10}$
	fruit		-0.058	0.953369
	vegetable		-0.150	0.881012
	bodypart	$*$	3.058	0.002230
	clothing	$*$	2.871	0.004094
	implement		-1.394	0.163453
	vehicle	$*$	4.731	$2.24 \cdot 10^{-6}$
	furniture		0.141	0.887841
location	mammal		-1.084	0.278464
	bird	$*$	3.073	0.002119
	fruit	$*$	-5.619	$1.93 \cdot 10^{-8}$
	vegetable	$*$	-2.075	0.037980
	bodypart		0.939	0.347774
	clothing	$*$	3.447	0.000567
	implement		-1.342	0.179544
	vehicle	$*$	4.782	$1.74 \cdot 10^{-6}$
	furniture	$*$	-2.439	0.014711

Table C.6.: Fit for binary model, only for responses to valid relations

factor interactions				
relation type	concept class	$p < 0.05$	z-score	p-value
category	mammal		1.166	0.243654
	bird	*	3.479	0.000504
	fruit		-0.003	0.997219
	vegetable	*	2.212	0.026986
	bodypart	*	-2.007	0.044776
	clothing	*	2.092	0.036425
	implement		0.578	0.563289
	vehicle	*	4.118	$3.82 \cdot 10^{-5}$
	furniture	*	-3.839	0.000123
part	mammal	*	3.450	0.000561
	bird	*	6.776	$1.24 \cdot 10^{-11}$
	fruit		0.989	0.322500
	vegetable		1.569	0.116637
	bodypart	*	6.011	$1.84 \cdot 10^{-9}$
	clothing	*	3.163	0.001561
	implement		-0.745	0.456038
	vehicle	*	5.611	$2.01 \cdot 10^{-8}$
	furniture		0.019	0.985019
location	mammal		0.079	0.937350
	bird	*	2.629	0.008570
	fruit	*	-6.527	$6.73 \cdot 10^{-11}$
	vegetable		-1.902	0.057174
	bodypart	*	4.441	$8.94 \cdot 10^{-6}$
	clothing	*	2.420	0.015505
	implement	*	-2.809	0.004971
	vehicle	*	5.735	$9.74 \cdot 10^{-9}$
	furniture	*	-3.847	0.000120

Table C.7.: Fit for binary model, only for responses to invalid relations

factor interactions				
relation type	concept class	$p < 0.05$	z-score	p-value
category	mammal	*	-4.591	$4.40 \cdot 10^{-6}$
	bird		1.355	0.175569
	fruit	*	-3.077	0.002093
	vegetable	*	-4.393	$1.12 \cdot 10^{-5}$
	bodypart		-0.402	0.687595
	clothing	*	2.759	0.005798
	implement		-1.696	0.089863
	vehicle		-0.983	0.325365
	furniture		-1.015	0.310234
part	mammal	*	-1.999	0.045644
	bird		1.888	0.059088
	fruit		-1.798	0.072199
	vegetable	*	-2.638	0.008340
	bodypart	*	-3.463	0.000533
	clothing		-0.104	0.916924
	implement	*	-2.390	0.016867
	vehicle		0.483	0.628885
	furniture		-0.612	0.540390
location	mammal		-0.667	0.504693
	bird		1.879	0.060206
	fruit		0.032	0.974821
	vegetable		0.171	0.864278
	bodypart	*	-4.436	$9.18 \cdot 10^{-6}$
	clothing	*	2.382	0.017198
	implement	*	2.017	0.043665
	vehicle		0.198	0.843203
	furniture		0.520	0.602753

D. Programming Scripts

The sections below shortly describe a selection of scripts I wrote that were useful during my research and might be of use for other researchers. A basic knowledge of the scripting languages perl[1] and python will help to further adapt them to individual needs.

The scripts are freely available for download from URL `http://clic.cimec.unitn.it/Files/PublicData/diss-scripts_GKremer.zip`

D.1. Stimuli Order Randomisation

File name: `rep-pick-random-lines.perl`

Many experiments require presenting a set of stimuli in a random order; at the same time, it is desirable to have a frequency-balanced distribution of stimuli across participants. This issue is getting more complex when the number of participants is not defined from the start and when there are more stimuli in the full set than aimed to be presented to one participant in an experiment run.

This script facilitates to repeatedly select a defined number of lines (representing separate stimuli or stimulus codes) randomly from a stimuli set file for creating a stimuli file that is to be used in a single experiment run. At any time, the frequency of any stimulus used in total in the combined single stimuli files will differ at most by 1 from the frequency of any other stimulus, i. e., the number of observations will be almost the same for all stimuli.

In more detail, each call to the script generates a stimuli file for a new participant with random order of the stimuli. Additionally, the script writes a summary file containing the given participant code and associated stimuli file name (that is generated partly by the script itself using a formerly unused character combination; only a base name for consistency of the file names for the specific experiment has to be defined). In case there are more stimuli than experiment trials for a participant, the whole stimuli set is randomised and only the required number of stimuli is printed. The remaining stimuli are stored in a separate file and will be used at the next call first, before introducing another full, original stimuli set into the process (thus ensuring all stimuli are used for an almost equal number of times across all participants).

Parameters to be specified for each call to the script include:

[1] see URL `http://www.perl.org`

- the name for the file containing the complete stimuli set (in text format), where stimuli (or, alternatively, stimulus codes) are given one per line
- the number of stimuli to be selected randomly from the stimuli file
- the name for the summary file (storing participant code and the name of the corresponding stimuli file for that experiment run)
- the participant code or name
- the template prefix for the stimuli file to be created for the respective participant

These parameters can be hard-coded in the script (not recommended) or specified as command line options. Calling the script with the option `-h` displays usage information comprising the options' names for the parameters described above.

D.2. Serial Print of Experiment Sheets

File name: `print-exp.perl`

This perl script is from the preparation phase of the production experiment described in section 2.1. Each participant had to be provided with experiment sheets (each showing a word stimulus near the top), an instructions sheet and a questionnaire. There were more stimuli in the whole set than presented to a participant; nevertheless, for the analysis, the number of observations for stimuli should be equally distributed across the data from the final set of participants. To keep the participant in the experiment anonymous, a number code was printed on the questionnaire and every experiment sheet (actually, also on the instructions sheet that were left to the participant after the experiment to be able to ask for removal of her/his data). Additionally, this number code included the position of the respective stimulus sheet in the set of sheets presented to that participant.

In more detail, the script generates and prints a set of paper sheets with a stimulus word per page (subset of the whole stimuli set for the experiment, in a randomised order) taken from a file in text format containing the full set of stimuli for the experiment, one per line. Along with the stimuli sheets, an instruction sheet and a questionnaire sheet are printed for each participant. All sheets get a participant–page code stamp printed in the upper left corner. The participant code is stored in a separate summary file to keep track of the participant codes for the next call (the participant code is incremented for each new participant). Participant–page code and stimulus are stored line by line in a file where the handwritten responses of participants can be added conveniently after the experiment.

Options to be specified at the call of the script comprise (call the script with the option `-h` to find out about the exact usage of these):

- the number of the participant code to start with,

- the number of sets of the experiment sheets to print (i. e., number of participants),
- the number of participant for which sets of experiment sheets should be printed, and
- the language (necessary for the choice of stimuli words).

Regarding requirements for the script, the typesetting system LaTeX should be installed; printing is executed via the (Unix operating system) command *lpr*. Additional files are necessary, which are also provided for download with the script to have a functioning framework that can be adapted then to your own needs:

- `instructionsDE.tex`, `instructionsIT.tex` (instructions in the respective language);
- `questionnaireDE.tex`, `questionnaireIT.tex` (questionnaires in the respective language);
- `StimuliProdEx.txt` (stimuli file);
- `stimiprint.tex` (template for the experiment sheets).

D.3. Feature Verification Experiment Run

File name: `vpropverify.py`

The feature verification experiment described in section 2.2.1 used this script for displaying word stimuli and recording participant responses. In such an experiment, the accuracy of stimuli presentation time and the accuracy of measuring of the reaction time (requiring to check frequently for response buttons pressed in parallel to presentation) is critical.

The script uses the Vision Egg[2] library (for python) to display (word) stimuli for accurate time spans, to record exact reaction times, mouse buttons clicked, an auto-incrementing participant code, and various PC-performance data.

The stimuli word pairs are randomised by a function in the script, intermixing stimuli of the conditions *valid* and *invalid* at the ratio of 50 %. Again, number of observations were required to be equal across stimuli, which was achieved by storing the unused stimuli in a temporary file that is read at the start of the next experiment run, before providing the full set of stimuli to the randomisation function.

The only parameter option that can be specified when calling the script is the language of the participant group (German or Italian); see the usage message by specifying the option `-h` with the call to the script. All other settings are hard-coded and have to be modified directly in the script.

[2] available from URL `http://www.visionegg.org`

The experiment data of all participants is stored in a common file. A separate file stores configuration and PC-performance for each experiment run. Additional files that the script currently needs for execution are provided along with the script:

- `stimuliDE.txt`, `stimuliIT.txt` (stimuli for the *valid* condition),
- `false-stimuliDE.txt`, `false-stimuliIT.txt` (stimuli for the *invalid* condition), and
- `testrun-stimuliDE.txt`, `testrun-stimuliIT.txt` (stimuli for the test runs before the real experiment).

References

Abel, A., Gamper, J., Knapp, J., and Weber, V. (2003). New answers to old questions about lexicon acquisition and dictionary use. In Lassner, D. and McNaught, C., editors, *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2003*, pages 1218–1224, Honolulu, Hawaii, USA. AACE.

Almuhareb, A. (2006). *Attributes in Lexical Acquisition*. PhD thesis, Department Of Computing and Electronic Systems, University of Essex.

Almuhareb, A. and Poesio, M. (2004). Attribute-based and value-based clustering: An evaluation. In *Proceedings of EMNLP*, pages 158–165.

Andrews, M., Vigliocco, G., and Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116(3):463–498.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Ashcraft, M. H. (1978a). Property dominance and typicality effects in property statement verification. *Journal of Verbal Learning and Verbal Behavior*, 17:155–164.

Ashcraft, M. H. (1978b). Property norms for typical and atypical items from 17 categories: A description and discussion. *Memory & Cognition*, 6(3):227–232.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59:390–412.

Baroni, M., Barbu, E., Murphy, B., and Poesio, M. (2010). Strudel: A distributional semantic model based on properties and types. *Cognitive Science*, 34(2):222–254.

Baroni, M. and Bernardini, S. (2004). Bootcat: Bootstrapping corpora and terms from the web. In *Proceedings of the Fourth Conference on International Language Resources and Evaluation (LREC)*, pages 1313–1316.

Baroni, M. and Lenci, A. (2008). Concepts and properties in word spaces. *Italian Journal of Linguistics*, 20(1):55–88.

Baroni, M. and Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*. To appear.

Barsalou, L. W. (1993). Flexibility, structure, and linguistic vagary in concepts: Manifestations of a compositional system of perceptual symbols. In Collins, A. F., Gathercole, S. E., Conway, M. A., and Morris, P. E., editors, *Theories of Memory*, pages 29–101. Lawrence Erlbaum Associates, London.

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

Caramazza, A. and Shelton, J. R. (1998). Domain-specific knowledge systems in the brain: The animate–inanimate distinction. *Journal of Cognitive Neuroscience*, 10:1–34.

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

Chanier, T. and Selva, T. (1998). The ALEXIA system: The use of visual representations to enhance vocabulary learning. In *Computer Assisted Language Learning*, volume 11, pages 489–522.

Church, K. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

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

Cree, G. S., McNorgan, C., and McRae, K. (2006). Distinctive features hold a privileged status in the computation of word meaning: Implications for theories of semantic memory. *Journal of Experimental Psychology: Learning, Meaning, and Cognition*, 32(4):643–658.

Cree, G. S., McRae, K., and McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science: A Multidisciplinary Journal*, 23(3):371–414.

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.

De Deyne, S. and Storms, G. (2008). Word associations: Norms for 1,424 dutch words in a continuous task. *Behavior Research Methods*, 40(1):198–205.

Devereux, B., Pilkington, N., Poibeau, T., and Korhonen, A. (2010). Large-scale acquisition of feature-based conceptual representations from textual corpora. In Ohlsson, S. and Catrambone, R., editors, *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, pages 49–54, Austin, TX, USA. Cognitive Science Society.

Evert, S. (2008). Corpora and collocations. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics: An International Handbook*, pages 1212–1248. Mouton de Gruyter, Berlin, Germany.

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

Forer, D., Paladino, M. P., Vettori, C., and Abel, A. (2008). 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.

Garera, N. and Yarowsky, D. (2009). Structural, transitive and latent models for biographic fact extraction. In *Proceedings of EACL*, pages 300–308, Athens, Greece.

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

Geckeler, H. (2002). Anfänge und Ausbau des Wortfeldgedankens . In Cruse, D. A., Hundsnurscher, F., Job, M., and Lutzeier, P. R., editors, *Lexikologie. Ein internationales Handbuch zur Natur und Struktur von Wörtern und Wortschätzen*, volume 21 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, pages 713–728. de Gruyter, Berlin – New York.

Girju, R., Badulescu, A., and Moldovan, D. (2006). Automatic discovery of part–whole relations. *Computational Linguistics*, 32(1):83–135.

Guagnano, D. (2010). *Bilingualism and Cognitive Development: A Study on the Acquisition of Number Skills*. PhD thesis, Centro Interdipartimentale Mente/Cervello (CIMeC), Università degli Studi di Trento.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 18(4):441–461.

Harris, Z. S. (1985). Distributional structure. In Katz, J. J., editor, *The Philosophy of Linguistics*, pages 26–47, New York. Oxford University Press.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of* COLING, pages 539–545, Nantes, France.

Hindle, D. (1990). Noun classification from predicate–argument structures. In *Proceedings of the Annual Meeting of the ACL*, pages 268–275.

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

Kilgarriff, A. (2007). Googleology is bad science. *Computational Linguistics*, 33(1):147–151.

Lapata, M. and Keller, F. (2004). The Web as a baseline: Evaluating the performance of unsupervised web-based models for a range of NLP tasks. In Dumais, S., Marcu, D., and Roukos, S., editors, *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 121–128, Boston, Massachusetts, USA. Association for Computational Linguistics.

Lapata, M. and Keller, F. (2005). Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing (TSLP)*, 2(1):1–31.

Liu, H. and Singh, P. (2004). ConceptNet: A practical commonsense reasoning toolkit. *BT Technology Journal*, pages 211–226.

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. *Behavior Research Methods*, 37(4):547–559.

Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(3):207–238.

Meyer, D., Zeileis, A., and Hornik, K. (2006). The strucplot framework: Visualizing multi-way contingency tables with vcd. *Journal of Statistical Software*, 17(3):1–48.

Moss, H. E., Tyler, L. K., and Devlin, J. T. (2002). The emergence of category-specific deficits in a distributed semantic system. In Forde, E. M. E. and Humphreys, G. W., editors, *Category Specificity in Brain and Mind*, chapter 5, pages 115–147. Psychology Press, East Sussex, UK.

Müller, W., editor (1997). *Die sinn- und sachverwandten Wörter. Synonymwörterbuch der deutschen Sprache*. Dudenverlag, Mannheim.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, 89(6):609–626.

Murphy, G. L. (2002). *The Big Book of Concepts*. MIT Press.

Pantel, P. and Pennacchiotti, M. (2006). Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of* COLING, pages 113–120, Sydney, Australia.

Rosch, E. and Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4):573–605.

Sahlgren, M. (2005). An introduction to random indexing. Retrieved from URL `http://www.sics.se/~mange/papers/RI_intro.pdf`.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3,4):379–423,623–656.

Shaoul, C. and Westbury, C. (2008). Performance of HAL-like word space models on semantic clustering. In *Proceedings of the* ESSLLI *Workshop on Distributional Lexical Semantics*, pages 42–46, Hamburg, Germany.

Smith, E. E., Shoben, E. J., and Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, 81(3):214–241.

Solomon, K. O. and Barsalou, L. W. (2001). Representing properties locally. *Cognitive Psychology*, 43(2):129–169.

Spence, D. P. and Owens, K. C. (1990). Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research*, 19(5):317–330.

Stopelli, P., editor (1999). *Dizionario sinonimi e contrari*. Garzanti, Milano.

Straw, A. D. (2008). Vision egg: An open-source library for realtime visual stimulus generation. *Frontiers in Neuroinformatics*, 2(4).

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

Vinson, D. P. and 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.

Vinson, D. P. and Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, 40(1):183–190.

Vinson, D. P., Vigliocco, G., Cappa, S. F., and Siri, S. (2003). The breakdown of semantic knowledge along semantic field boundaries: Insights from an empirically–driven statistical model of meaning representation. *Brain and Language*, 86:347–365.

Wu, L. and Barsalou, L. W. (2009). Perceptual simulation in conceptual combination: Evidence from property generation. *Acta Psychologica*, 132(2):173–189.