Title: **Association Norms for German Noun Compounds and their Constituents**

Authors: Sabine Schulte im Walde and Susanne Borgwaldt

To appear in: Behaviour Research Methods.

The final publication is available at Springer via http://dx.doi.org/10.3758/s13428-014-0539-y.

Affiliations and addresses:

PD Dr. Sabine Schulte im Walde (corresponding author)
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5B
70569 Stuttgart
Germany
Phone: +49 (711) 685-84584
Email: schulte@ims.uni-stuttgart.de

Dr. Susanne Borgwaldt
Germanistisches Seminar
Universität Siegen
Adolf-Reichwein-Straße 2
57068 Siegen
Germany
Phone: +49 (271) 740-2705
Email: borgwaldt@germanistik.uni-siegen.de

Association Norms for German Noun Compounds and their Constituents

Sabine Schulte im Walde and Susanne Borgwaldt

Abstract

We present a collection of association norms to 246 German depictable compound nouns and their constituents, comprising 58,652 association tokens distributed over 26,004 stimulus-associate pair types. Analyses of the data reveal that participants mainly provided noun associations, followed by adjective and verb associations. In corpus analyses, co-occurrence values for compounds and their associations were below those for nouns in general and their associations. The semantic relations between compound stimuli and their associates were more often co-hyponymy and hypernymy and less often hyponymy than for associations to nouns in general. Finally, there was a moderate correlation between the overlap of associations to compounds and their constituents and the degree of semantic transparency.
These data represent a collection of associations to German compound nouns and their constituents which constitute a valuable resource concerning the lexical semantic properties of the compound stimuli, and the semantic relations between the stimuli and their associations. More specifically, they can be used for stimulus selection, hypotheses testing and further research on morphologically complex words. The norms are available in text format (utf-8 encoding) as supplemental material.

Keywords: associations, semantic relations, compound nouns, complex words

1. Introduction

In this article, we introduce a new data collection of association norms for German noun compounds and their constituents. Association norms have a long tradition in psycholinguistic research. They have been used for more than 30 years to investigate semantic memory, making use of the implicit notion that associates reflect meaning components of words. In experimental psychology, association norms have – for example – been extensively used to conduct studies of semantic priming to investigate (among other things) word recognition, knowledge representation and semantic processes (see McNamara (2005) for a review of methods, issues, and findings).

We collected associations to German noun compounds because we believe that the associations are a valuable resource for cognitive and computational linguistics research. Based on an existing collection of German noun compounds (von der Heide & Borgwaldt, 2009), we therefore gathered associations for the compounds and also for their constituents (e.g., *Ahornblatt/Ahorn/Blatt* `maple leaf/maple/leaf'). The data were collected via the crowdsourcing interface *Amazon Mechanical Turk (AMT)*. We perform detailed analyses of the collection, regarding the parts-of-speech of the associate responses, and the co-occurrence and syntactic patterns as well as the semantic relations between the stimulus-associate pairs. We also predict the degree of semantic transparency of the compounds, as based on a simple association overlap. The analyses are compared to those of an earlier collection (Schulte im Walde, Borgwaldt, & Jauch, 2012), where associations to a superset of our compound and constituent stimuli were gathered in a more controlled web experiment.

The association norms can be used as a lexical semantic resource concerning the target stimuli, i.e., the compound nouns and their constituents. The data should be relevant for research on the lexical semantic properties of the compound stimuli, e.g. the semantic relations between the stimuli and their associations, and the degree of semantic relatedness between the compound stimuli and their constituents, i.e., the degree of semantic transparency (alternatively *compositionality*).

In this article, we first provide an overview of association norms in general terms (Section 2) and introduce the German compound and constituent targets the norms rely on (Section 3), before we describe the collection and analyses of the noun compound association norms (Sections 4 and 5, respectively). The final part of the article summarises and discusses the results.

2. Previous work on association norms

2.1 Collections of association norms

One of the first collections of word association norms was done by Palermo & Jenkins (1964), comprising associations for 200 words. The *Edinburgh Association Thesaurus* (Kiss, Armstrong, Milroy, & Piper, 1973) was a first attempt to collect association norms on a larger scale, and also to create a network of stimuli and associates, starting from a small set of stimuli derived from the Palermo and Jenkins norms. A similar motivation underlay the association norms from the University of South Florida (Nelson, McEvoy, & Schreiber, 2004), who developed a stimulus-associate network for more than 20 years, starting in 1973. Their goal was to obtain the largest database of free associations ever collected in the United States available to interested researchers and scholars. More than 6,000 participants produced nearly three-quarters of a million responses to 5,019 stimulus words. Smaller sets of association norms have also been collected for example for German (Russell & Meseck, 1959; Russell, 1970; Melinger & Weber, 2006; Schulte im Walde, Melinger, Roth, & Weber, 2008), Dutch (Lauteslager, Schaap, & Schievels, 1986; de Deyne & Storms, 2008), French (Ferrand & Alario, 1998), Spanish (Macizo, Gómez-Ariza, & Bajo, 2000; Fernández, Diez, Alonso, & Beato, 2004), Portuguese (Comesaña, Fraga, Moreira, Frade, & Soares, 2014), and across languages (Kremer & Baroni, 2011), as well as for different populations of speakers, such as adults vs. children (Macizo et al., 2000; Hirsh & Tree, 2001), and for words with various degrees of emotion (John, 1988) or homographs (French & Richards, 1992). While some of the norms occasionally contain compounds, as far as we know, there is no specific collection focusing on associations for compounds and their constituents yet, other than our own previous collection for a superset of the compound and constituent stimuli (Schulte im Walde et al. 2012).

2.2 Analyses of association norms

In parallel to the interest in collecting association norms, researchers have analysed association data in order to get insight into semantic memory. The following paragraphs provide an overview of these analyses, starting with theoretical considerations on relationships between stimuli and responses in association norms, and progressing towards analyses of collected norms.

Clark (1971) identified relations between stimulus words and their associations on a theoretical basis, not with respect to collected association norms. He categorised stimulus-association relations into sub-categories of paradigmatic and syntagmatic relations, such as synonymy and antonymy, selectional preferences, etc. Heringer (1986) concentrated on syntagmatic associations to a small selection of 20 German verbs. He asked his subjects to provide question words as associations (e.g., *wer* `who', *warum* `why'), and used the responses to investigate the valency behaviour of the verbs. Spence & Owens (1990) showed that associative strength and word co-occurrence are correlated. Their investigation was based on 47 pairs of semantically related concrete nouns, as taken from the Palermo and Jenkins norms, and co-occurrence counts in a window of 250 characters in the 1-million-word Brown corpus. Church & Hanks (1990) were the first to apply information-theoretic measures to corpus data in order to predict word associations. However, they did not rely on or evaluate their findings against existing association data, but rather concentrated on the usage of the measure for lexicographic purposes. Rapp (2002) combined research questions and methods from the above previous work: He developed corpus-based approaches to predict paradigmatic and syntagmatic associations, relying on the 100-million word *BNC* corpus. Concerning paradigmatic associations, he computed word association as the similarity of context vectors, applying the *City block distance* (also known as *Manhattan distance*, or L_1 *norm*) as a similarity measure. A qualitative inspection revealed a strong overlap of very similar words with human associations, and applying the associations to solve the TOEFL test resulted in an accuracy of 69%. Concerning syntagmatic associations, he demonstrated that the word with the strongest co-occurrence to a target word (and filtered by a log-likelihood test) corresponded to the first human association of the respective target word in 27 out of 100 cases. Rapp's work used the *Edinburgh Association Thesaurus* as the association database. In addition to his above contributions, his paper also provided an illustration of how strongly the co-occurrence distance between target stimuli and their associations was related to the respective number of responses to the stimuli in the association norms.

Work by Fellbaum and colleagues in the 1990s focused on human judgements concerning the semantic relationships between verbs. Fellbaum & Chaffin (1990) asked participants in an experiment to provide associations to verbs. Their work concentrated on verb-verb relations and therefore explicitly required verb responses to the verb stimuli. The resulting verb-verb pairs were manually classified into five pre-defined semantic relations. Fellbaum (1995) investigated the relatedness between antonymous verbs and nouns and their co-occurrence behaviour. Within that work, she searched the Brown corpus for antonymous word pairs in the same sentence, and found that regardless of the syntactic category, antonyms occur in the same sentence with frequencies which are much higher than chance. Last but not least, the WordNet organisation of the various parts-of-speech largely relies on psycholinguistic evidence (Fellbaum, 1998).

Based on the associations to German nouns and verbs which were collected by Schulte im Walde et al. (2008), cf. Section 2.1, they performed detailed analyses at the syntax-semantics interface. Guida (2007) replicated most of their analyses on verb association norms for Italian verbs. Roth & Schulte im Walde (2008) extended the co-occurrence analysis for noun associations in Schulte im Walde et al. (2008) and explored whether dictionary and encyclopaedic information provided more world knowledge about associations than corpus co-occurrence. They found that the information in the three resource types complemented each other. Schulte im Walde & Melinger (2008) performed a more in-depth analysis of the co-occurrence distributions of the noun associations in Schulte im Walde et al. (2008). Schulte im Walde et al. (2012) performed a part-of-speech analysis on previously collected associations for compound nouns and their constituents.

From a more applied point of view in the field of Computational Linguistics, Melinger, Schulte im Walde, & Weber (2006) took the noun associations as input to a soft clustering approach, in order to predict noun ambiguity, and to discriminate the various noun senses of ambiguous stimulus nouns. Schulte im Walde (2008) relied on the associations to German verbs described by Schulte im Walde et al. (2008) to determine salient features for automatic semantic verb classification.

3. Noun Compounds

Compounds are morphologically complex words, coined by two or more simple words. Our focus of interest is on German noun compounds (see Fleischer & Barz (2012) for a detailed overview and Klos (2011) for a recent exploration), such as *Ahornblatt* `maple leaf', *Feuerwerk* `fireworks', *Nähmaschine* `sewing machine', *Obstkuchen* `fruit cake' and *Rotkohl* `red cabbage', where the grammatical head (in German, this is the rightmost constituent) is a noun, and the modifier can belong to various parts-of-speech.

More specifically, we are interested in the degrees of *semantic transparency* of German noun compounds, i.e., the relation between the meaning of the whole compound (e.g., *butterfly*) and the meaning of its parts (e.g., *butter, fly*), which has been studied intensively by psycholinguists, in order to find out how compound words are cognitively processed and represented in the mental lexicon. There is an ongoing debate about whether morphologically complex words are stored and processed as single units (Butterworth's full listing approach (1983)), whether they are decomposed into their morphemes (Taft, 2004; Taft & Forster, 1975), or whether they can be accessed both ways: as whole forms and componentially, via their constituent morphemes (*dual route models,* cf. e.g. Caramazza, Laudanna, & Romani, 1988; Baayen & Schreuder, 1999), and which variables predict processing behaviour. The majority of studies in this area investigate *morphological decomposition* during *compound comprehension,* but see e.g. Lüttmann, Zwitserlood, Böhl, & Bölte (2011) for evidence of *morphological composition* during *compound production*.

Factors that have been found to influence the cognitive processing and representation of compounds include orthographic variables like the number of letters (Bertram & Hyönä, 2003) or the presence of hyphens or interword spaces (Bertram, Kuperman, Baayen, & Hyönä, 2011), frequency-based measures as the frequencies of the compound and its constituents (e.g. Janssen, Bi & Caramazza (2008); van Jaarsveld & Rattink (1988)) and the morphological family size, i.e., the number of compounds that share a constituent (de Jong, Feldman, Schreuder, Pastizzo, & Baayen, 2002), variables relating to morphological complexity as e.g. the number of morphemes or the existence of linking elements (Krott, Schreuder, Baayen, & Dressler, 2007), and semantic variables as e.g. the relationship between compound modifier and head, i.e., a *teapot* is *a pot FOR tea,* and a *snowball* is *a ball MADE OF snow* (Gagné & Spalding, 2009).

With some researchers (e.g., Longtin, Segui, & Hallé, 2003; Marslen-Wilson, Tyler, Waksler, & Older, 1994) arguing that morphological decomposition happens only in semantically transparent polymorphemic words and not in semantically opaque ones, one variable that might be particularly important for the processing of compounds is their compositionality/semantic transparency. For example, studies by Sandra (1990) and Zwitserlood (1994) showed that the meanings of the constituents of semantically transparent compounds (e.g., *dog* and *house* in *doghouse*) were activated during processing, whereas the meanings of the constituents of opaque compounds (e.g. *butter* and *fly* in *butterfly*) were not activated.

Although inter-rater agreement about compounds' perceived semantic transparency is generally rather high (Reddy, McCarthy, & Manandhar, 2011; Roller, Schulte im Walde, & Scheible, 2013), there has been little research on how the semantic transparency is assessed. One assumption is that the degree of a compound's semantic transparency should also be reflected in its association patterns. If a compound is classified as opaque (e.g. *butterfly*), one would assume that the associations to the whole compound show less overlap with the associations to the components (*butter, fly*), than in the case of a transparent compound (e.g. *doghouse*), as during the processing of (partially) opaque compounds the opaque constituents might be less activated at the semantic level. This is what Libben's (1994; 1998) "automatic progressive parsing and lexical excitation (APPLE) model of morphological parsing" predicts: It assumes that compound words are represented at three levels: a stimulus level, a lexical level (purely morphological) and a conceptual level (semantic). A (partially) opaque compound such as *strawberry* is decomposed at the lexical level into *straw* and *berry*. At the conceptual level, however, only the semantically transparent constituent *berry* is represented and can accordingly generate associations. This representation difference might, for example, explain the observed dissociation between constituent repetition priming effects (Libben, Gibson, Yoon, & Sandra, 2003; Zwitserlood, 1994) and semantic priming effects (Sandra, 1990; Zwitserlood, 1994) for opaque compounds.

In sum, our collection of association norms for compounds and their constituents is mainly aimed to be of use for researchers on compound processing. It provides a resource for stimulus selection in experimental studies and also allows researchers to examine properties of associations to compounds and their constituents in detail, in order to gain more insight into the way meanings of compounds are processed and/or represented in the mental lexicon.

4. Experiment

Associations are commonly obtained by presenting *target stimuli* to the participants in an experiment, who then provide *associate responses*, i.e., words that are spontaneously called to mind by the stimulus words. The quantification of the resulting stimulus–association pairs (i.e., how often a certain association is provided for a certain stimulus) is called *association norm*. In the following, we describe the collection of our associations to German noun compounds.

Method

Material The target compounds and constituents were based on the selection of noun compounds by von der Heide & Borgwaldt (2009). They created a set of 450 concrete, depictable German noun compounds that they grouped into four transparency classes: compounds that are transparent with regard to both constituents (**TT**, for example, *Ahornblatt* `maple leaf'); compounds that are opaque with regard to both constituents (**OO**, for example, *Löwenzahn* `lion+tooth → dandelion'); compounds that are transparent with regard to the modifier but opaque with regard to the head (**TO**, for example, *Feuerzeug* `fire+stuff → lighter'); and compounds that are opaque with regard to the modifier but transparent with regard to the head

(**OT**, for example, *Fliegenpilz* `fly+mushroom → toadstool').[1] In total, the four classes contained 220 instances of TT, 126 instances of OT, 79 instances of TO, and 25 instances of OO.

The 450 noun compounds from von der Heide & Borgwaldt (2009) were categorized according to the morphological category of the modifier (AN: adjective-noun compound, NN: noun-noun compound, PN: preposition-noun compound, VN: verb-noun compound, MN: noun compound where the modifier is morphologically motivated by multiple classes, unique: noun compound with unique modifier). The categorization was performed by consensus decision of four computational linguists.

In our association collection, we only used 246 compounds of the 450 noun compounds from von der Heide & Borgwaldt (2009), i.e., 237 bi-morphemic noun-noun compounds and 9 noun compounds where the modifier is unique, a so-called cranberry morpheme (such as *him* in *Himbeere* `him+berry → raspberry'), with no meaning by itself. Each compound had exactly two simple constituents. The compound set comprised 106 instances of TT, 37 instances of TO, 87 instances of OT, and 16 instances of OO. We restricted the target set because the subset of the two-part noun-noun compounds was most relevant to our research. Appendix A provides the complete list of our noun compounds.

In total, our material comprised 571 targets. The total number of target stimuli was less than 3*246 because some compounds share constituents. We first divided the stimuli randomly into four separate packages, making sure that there was a similar number of compounds in each of the four packages, and that both constituents of each compound were in the same package as the compound. In this way, constituents that are shared by several compounds might appear in more than one part but we could collect the associations to the compounds and their constituents in the four packages independently of each other. Taking the multiple occurrences of some constituents into account, the four packages contained 173/169/173/172 targets, respectively. The four packages were then each randomly divided into eight parts containing 21-22 targets in a random order. To control for spammers and to identify non-native speakers of German, we also included three German fake compound nouns into each of the batches, in random positions of the lists. The list of fake nouns is (in alphabetical order): *Analigzerbruch, Armmoder, Brattlider, Bulkerzagen, Engschogen, Fennhoder, Harmweg, Luderschwege, Malligwohmer, Pillinrugen, Quetpfluge, Tropebuhle, Wierzverkuhr, Zogschucht.*

Procedure The experiment was performed via Amazon Mechanical Turk (AMT)[2]. When an AMT worker chose one of our batches, the worker was presented 24 or 25 noun compounds (including 21or 22 real and three fake compounds) in 24/25 *HIT*s, respectively. The setup of the experiment is shown by Figure 1 in Appendix B, with translations in red font. The actual collection of the associations was performed as shown in Figures 4 and 5 in Appendix B: Each trial (represented by a HIT) provided a brief description of the experiment, an example item with potential responses, and a single target (one of the noun compounds, or one of the constituents, or a fake noun). Below the target were three data input lines where participants could type their associations. They were instructed to type at most one word per line and, following German

1 The original classification into TT, TO, OT and OO was performed by the two authors (Claudia von der Heide and Susanne Borgwaldt). Transparency ratings were collected in a follow-up step, using a 1-7 scale.
2 www.mturk.com

grammar, to distinguish nouns from other parts-of-speech with capitalisation. Below the three input lines was a box that participants were asked to check if they did not know the word. The 4*8=32 batches were completed within 1 to 26 days each, and the whole collection was finished within 3 months.

Participants The participants were Amazon Mechanical Turk workers. We asked for 25 native German speakers per target, and paid 2 US Cent per trial (i.e., for up to three associations). We only accepted participants who identified the fake nouns correctly and who in addition had an overall approval rate of at least 95% (after the experiment was completed). These checks ensured that we received associations from native German speakers only. Over all trials, we had participants with 146 different worker IDs, who provided associations for 1 up to 683 stimuli.

Data

For each stimulus, we had between 2 and 120 participants. Because the participants could provide between zero and three associations per target, the actual number of associations per stimulus varies between 6 and 356. All but the unique constituents received ≥48 associations. In total, we collected 58,652 associations from 20,333 trials, an average of 2.88 associations per trial. The 58,652 association tokens are distributed over 26,004 association types.

5. Previous Experiment

This section reports on a previous experiment to collect associations to the compound nouns and their constituents from von der Heide & Borgwaldt (2009), as described by Schulte im Walde et al. (2012). This experiment was our first attempt to collect associations to compound nouns and their constituents, and was stopped after one year because the incoming data stagnated. The current experiment described in Section 4 was then set up to continue the collection with a more specific focus on noun-noun compounds, still using identical collection instructions and procedures.

The reasons why we report on the earlier experiment are the following.

- The previous associations were collected for a superset of the noun compounds that were used in the current study. The two collections can be exploited independently (taking into account that they were collected in different ways) or together (making use of a richer set of associations for the intersection of the noun-noun compounds and their constituents, as well as exploiting an overall larger stimulus set).
- Moreover, we were interested if we could find differences between the two sets of association norms, as they were collected in different ways.

Method

Material We used 442 compound nouns and constituents from the original selection by von der Heide & Borgwaldt (2009). In total, our material comprised 996 targets, i.e., 442 compounds and 554 constituents. The stimuli were divided randomly into 12 separate experimental lists of 83 nouns each.

Procedure The experiment was administered over the Internet and announced by emails to colleagues and friends. Participants were first provided a brief description of the experiment and asked for their biographical information, such as linguistic expertise, age and regional dialect (see Figure 4 in Appendix B, with a translated version in Figure 5). Next, the participant was presented with the written instructions for the experiment and an example item with potential responses. In the actual experiment, one of the 12 experimental lists was chosen randomly. Each of the 83 trials in the experimental list consisted of a single word presented in a box at the top of the screen. The word was either one of the noun compounds, or one of the constituents. If a compound constituent was not a noun, the base form was nominalised by starting it with a capital letter. For example, the verbal modifier *fahren* in *Fahrplan* `to ride+schedule → time table') was represented as *Fahren*, the adjectival modifier *blau* in *Blaubeere* `blue+berry → blueberry') was represented as *Blau*. The experiment ran for approx. one year.

The order of the target words was random for each data set and each participant. Below the target were three data input lines where participants could type their associations. They were instructed to type at most one word per line and, following German grammar, to distinguish nouns from other parts-of-speech with capitalisation. Below the three input lines was a box that participants were asked to check if they did not know the word.

Participants 268 participants took part in the experiment. 225 of them claimed their L1 to be German; for 19 of them, German was not their L1; 1 of the participants claimed to have grown up bi-lingual; and 23 did not provide information about their L1. 124 of the individuals identified themselves as having a linguistic background and 112 rated themselves as linguistic novices; 32 participants did not provide information about their linguistic background.

Data

For each experimental list, we had between 14 and 28 participants. Because the participants could provide between zero and three associations per target, the number of participants per stimulus varies between 10 and 36, and the number of associations per stimulus varies between 6 and 74. In total, we collected 47,249 associations from 17,128 trials, an average of 2.76 associations per trial. The 47,249 association tokens were distributed over 28,238 association types. In 861 trials, the participants did not provide any association, out of which 327 targets were explicitely checked as not known by the participants.

6. Overall Results and Analyses

Quantifying over all responses in the two experiments resulted in a total of 47,523/106,693 stimulus–association types/tokens. Table 1 summarises the sizes of the association norms with regard to the current study, the previous study and the union of the two.We distinguish between associations to all noun compound stimuli and their constituents (as in the previous experiment) and associations to the subset of 246 noun-noun compounds and their constituents (which were used in the current experiment and already part of the previous experiment). Even though our previous experiment ran for approximately one year and the current experiment for only three months, we collected more associations in the current experiment.

Experiment	Stimuli	Responses	
		Types	Tokens
Web Experiment	all noun compounds and their constituents	28,238	47,249
	noun-noun compounds and their constituents	15,600	26,397
AMT Experiment	*noun-noun compounds* and their constituents	26,004	58,652
Both Experiments	all noun compounds and their constituents	47,523	106,693
	noun-noun compounds and their constituents	34,560	85,049

Table 1. Sizes of association norms.

Tables 2 to 4 provide examples for the associations to three compounds and their constituents, in each case listing the 10 strongest (i.e., most frequently provided) associations. We selected three examples rather than a single one to observe (i) the effect of the degree of transparency of the compound on the associations to the compounds vs. constituents, as well as (ii) the effect of monosemous vs. polysemous stimuli with regard to the semantics of the associations. Note that *Fliegenpilz* is less transparent than *Ahornblatt* (at least with respect to its modifier), so that the associations of the compound and the modifier differ more strongly. Note also that two of the nouns are polysemous: *Fliege* and *Blatt*. For both nouns we received associations to two senses: in the case of *Blatt*, associations were given to the plant sense `leaf' as well as to the paper sense `sheet (of paper)'; *Fliege* evoked associations to the animal sense `fly' as well as to the clothes sense `bow tie'.

In the following subsections 6.1 to 6.4 we present a series of analyses of the association norms with regard to the following stimulus properties and stimulus-association relations:

- a *morpho-syntactic analysis*, looking into the parts-of-speech of the associations,
- a *distributional analysis*, looking into the co-occurrence of stimuli and associations,
- a *syntactic analysis*, looking into the dependency paths between stimuli and associations,
- a *semantic relation analysis*, looking into the stimuli-association semantic relations.

In all analyses, we will pay attention to the effects of our specific set of stimuli, compounds and their constituents, and explore which syntactic and semantic properties are specific to these stimuli. In subsection 6.5, we conclude our analyses with a study regarding the compounds' semantic transparency by investigating the correlation between association overlap (of compounds and constituents) and an existing set of compound–constituent ratings.

Ahornblatt `maple leaf'			*Ahorn* `maple'			*Blatt* `leaf/sheet (of paper)'		
Kanada	`Canada'	26	Baum	`tree'	39	Baum	`tree'	29
Baum	`tree'	23	Sirup	`syrup'	30	Papier	`paper'	27
Herbst	`autumn'	15	Kanada	`Canada'	22	grün	`green'	10
Sirup	`syrup'	12	Blatt	`leaf'	16	schreiben	`write'	7
Laub	`leaves'	7	Blätter	`leaves'	6	Pflanze	`plant'	4
Wald	`forest'	4	Herbst	`autumn'	4	Laub	`leaves'	3
rot	`red'	4	Wald	`forest'	3	Herbst	`autumn'	3
bunt	`colourful'	3	grün	`green'	2	weiß	`white'	3
grün	`green'	3	rot	`red'	2	Blume	`flower'	3
zackig	`jagged'	2	Ahornsirup	`maple syrup'	2	Wald	`forest'	2

Table 2. Most frequent responses to compound *Ahornblatt* `maple leaf' and its constituents.

Fliegenpilz `toadstool'			*Fliege* `fly/bow tie'			*Pilz* `mushroom'		
giftig	`poisenous'	34	nervig	`annoying'	11	Wald	`forest'	32
rot	`red'	26	Insekt	`insect'	9	giftig	`poisenous'	13
Wald	`forest'	18	summen	`buzz'	9	sammeln	`collect'	9
Gift	`poison'	6	Sommer	`summer'	6	Fliegenpilz	`toadstool'	6
Punkte	`dots'	5	Klatsche	`flap'	6	Champignon	`mushroom'	5
weiß	`white'	4	lästig	`annoying'	5	Suppe	`soup'	4
Märchen	`tale'	2	Fliegenklatsche	`fly flap'	5	essen	`eat'	3
rot-weiß	`red-white'	2	Tier	`animal'	5	essbar	`edible'	3
gepunktet	`dotted'	2	klein	`small'	5	Schimmel	`mould'	3
Vergiftung	`poisoning'	2	Krawatte	`tie'	4	Champignons	`mushrooms'	3

Table 3. Most frequent responses to compound *Fliegenpilz* `toadstool' and its constituents.

Schlittenhund `sledge dog'			*Schlitten* `sledge'			*Hund* `dog'		
Schnee	`snow'	24	Winter	`winter'	34	bellen	`bark'	46
Husky	`husky'	21	Schnee	`snow'	32	Katze	`cat'	33
Alaska	`Alaska'	11	Kinder	`children'	9	Haustier	`pet'	20
Winter	`winter'	10	kalt	`cold'	8	Tier	`animal'	18
kalt	`cold'	7	Spaß	`fun'	7	Freund	`friend'	17
Eskimo	`Eskimo'	5	rodeln	`sledge'	5	treu	`faithful'	13
Schlitten	`sledge'	4	fahren	`ride'	5	Leine	`leash'	10
rennen	`run'	4	Hügel	`hill'	4	Maus	`mouse'	9
Eis	`ice'	4	Kufen	`skids'	3	beißen	`bite'	8
ziehen	`pull'	3	Hund	`dog'	2	Hundehütte	`kennel'	6

Table 4. Most frequent responses to compound *Schlittenhund* `sledge dog' and its constituents.

6.1 Morpho-syntactic analysis

In the morpho-syntactic analysis, each response to the stimuli was assigned its – possibly ambiguous – part-of-speech (*pos*). The results provide insight into the relevance of predominant part-of-speech categories. Similar morpho-syntactic analyses have been performed by Guida (2007), Schulte im Walde et al. (2008) and Schulte im Walde et al. (2012).

As resource for the pos assignment we relied on a lemmatised and pos-tagged frequency list of the *SdeWaC* corpus (Faaß & Eckart, 2013), a cleaned version of the German web corpus *deWaC* created by the *WaCky* group (Baroni, Bernardini, Ferraresi, & Zanchetta, 2009). The *SdeWaC* contains approx. 880 million word tokens. We disregarded fine-grained distinctions such as case, number and gender features and considered only the major categories verb (V), noun (N), and adjective (ADJ). A fourth category `OTHER' comprises all other part-of-speech categories such as adverbs, prepositions, particles, interjections, conjunctions, etc. Ambiguities between the categories arose e.g. when the experiment participant could have been referring either to an adjective or a (non-capitalised) noun, such as *fett* `fat'.[3]

Having assigned part-of-speech tags to the associations, we were able to distinguish and quantify the morpho-syntactic categories of the associations. In non-ambiguous situations, the unique part-of-speech received the total stimulus-association strength. For example, *Herd* `cooker' was provided as association to the stimulus *Pfanne* `pan' by 11 participants. Our pos resource contained *Herd* in the corpus only as a noun. So *Pfanne* received a contribution of all 11 mentions for a noun pos. In ambiguous situations, the stimulus-association frequency was split over the possible part-of-speech tags according to the pos proportions in the frequency list. For example, *fett* `fat' was provided as association to the stimulus *Pfanne* by three participants. Our pos resource contained *Fett* 10,780 times in the corpus as a noun, and 493 times as an adjective. So, with regard to *Fett*, *Pfanne* received a contribution of 3*10,780/(10,780+493) = 2.87 nouns, and of 3*493/(10,780+493) = 0.13 adjectives.

The output of this analysis is frequency distributions of the part-of-speech tags for each stimulus individually, and also as a sum over all stimuli. Table 5 presents the total numbers over all associate response tokens, across the two experiments. Overall, the participants provided noun associations in the clear majority of token instances, 69-73%; adjectives were given in 12-16% of the associations, and verbs in 11-13%. There are slight differences across the two experiments and when considering all compounds vs. only the noun-noun compounds, but overall the proportions are very similar in this table and also in comparison to the same analyses in previous work regarding associations to noun stimuli (Schulte im Walde et al., 2008). This is the case even if we only consider the associations to the compound stimuli.

Experiment	Stimuli	Freq/ Prop	N	ADJ	V	OTHER
Web Experiment	all noun compounds and their constituents	Freq	33,322	6,835	5,264	1,827
		Prop	71%	14%	11%	4%

3 Despite our instructions, some participants failed to use capitalisation, leading to ambiguity.

	noun-noun compounds and their constituents	Freq	18,263	4,197	2,971	967
		Prop	69%	16%	11%	4%
AMT Experiment	*noun-noun compounds* and their constituents	Freq	43,104	7,162	7,365	1,021
		Prop	73%	12%	13%	2%
Both Experiments	*noun-noun compounds* and their constituents	Freq	61,367	11,359	10,337	1,986
		Prop	72%	14%	12%	2%
	noun-noun compounds only	Freq	23,002	5,246	3,972	845
		Prop	70%	16%	12%	3%

Table 5. Number and proportions of association parts-of-speech.

6.2 Co-occurrence analysis

This analysis examines whether the co-occurrence assumption holds for our association norms, i.e., what proportion of the associations is found in co-occurrence with the stimulus words in a corpus. The co-occurrence hypothesis assumes that associations are related to the textual co-occurrence of the stimulus-association pairs. The hypothesis has been confirmed in many previous studies, among them Miller (1969), Spence & Owens (1990), Fellbaum (1995), Schulte im Walde & Melinger (2008), Schulte im Walde et al. (2008), and Schulte im Walde & Müller (2013).

As association norms, we use all experiment data, and our study again relies on the *SdeWaC* corpus. This time, no parsing information is required, as we only check how often the associations co-occur within windows of 5 and 20 words (to the left and to the right).[4] The analysis is again token-based. Table 6 presents the results. The columns show the proportions of pairs that co-occur with a co-occurrence strength of 1 (i.e., they co-occur at least once), 2 (i.e., they co-occur at least twice), etc. The rows distinguish between windows of size 5 and 20. For example, the association *Wald* `forest', which was provided by 32 participants for the stimulus *Pilz* `mushroom', was found 76 times in a window of 5 words from the stimulus. The 32 tokens thus appear in all columns. In contrast, the association *Husky* `husky', which was provided by 21 participants for the stimulus *Schlittenhund* `sledge dog', was found 10 times in a window of 5 words from the stimulus. The 10 tokens thus appear in the columns with co-occurrence strength ≥ 1, 2, 5, 10.

For each co-occurrence range, we present two lines, one for all stimulus-association pairs, and the second when restricting the stimuli to the compounds. Note that the *all stimuli* condition is restricted to the 246 noun-noun compounds and their nominal constituents (i.e., disregarding non-nominal stimuli) so that we can compare the co-occurrence analyses to previous work on nominal stimuli and thus compare the co-occurrence information for (i) noun-noun compounds and their nominal constituents (all stimuli), (ii) the nominal compound stimuli only, and (iii) nouns in general (cf. Schulte im Walde et al., 2008).

4 Note that the sentences in the *SdeWaC* are sorted alphabetically, so the window co-occurrence refers to 5/20 words to the left and right BUT within the same sentence.

		Co-occurrence strength					
		1	2	5	10	20	50
window 5	*all stimuli*	73%	65%	53%	44%	34%	23%
	compound stimuli	58%	46%	30%	20%	11%	4%
window 20	*all stimuli*	79%	72%	62%	53%	43%	31%
	compound stimuli	67%	56%	41%	29%	18%	9%

Table 6. Corpus co-occurrence of stimulus-associate pairs.

We can see that the co-occurrence assumption clearly holds for our norms: 73% of the associations appeared within a 5-word window of the respective stimuli at least once, and more than half (53%) of the associations appeared at least 5 times in a 5-word window of the respective stimuli. However, we also notice that the co-occurrence values are below those for German nouns and their associations in general: Compared to earlier studies (Schulte im Walde et al., 2008), where we performed the co-occurrence check on associations of 409 depictable German nouns (only including few compound nouns) and found 84% of the stimulus-association pairs in a 20-word window at least once, the current analysis found only 79% (and with regard to compound-association pairs only 67%). The difference is stronger than it first appears, because the earlier analyses were performed on a 200-million word German newspaper corpus from the 1990s (so the co-occurrence source was much smaller than the 880-million word corpus *SdeWaC*). In order to make the two analyses more comparable, we repeated our co-occurrence check for the compound norms on the 200-million word corpus. For the general noun norm data in Schulte im Walde et al. (2008) we found 73/84% of the stimulus-association pairs at least once in windows 5 and 20, whereas we only found 57/68% of the stimulus-associate pairs in the current analysis (taking compound and constituent stimuli into account), and with regard to compound-association pairs we found only 38/53%.[5] This comparison clearly shows that while the co-occurrence assumption holds for our norms, the corpora offer less information on compounds than on simple nouns.

6.3 Syntactic dependency paths

This analysis goes beyond pure co-occurrence and investigates the syntactic dependency paths between the stimuli and their associations. The analysis tells us whether there are any specifically strong syntactic relationships as triggers for associations. For example, we do not know a priori whether responses refer to associations that are in a modifier relationship with the stimuli (such as *grün* `green' for *Salat* `salad'), or in conjunction (such as *Katze* `cat' for *Hund* `dog'), etc. To obtain insight into which syntactic dependencies might trigger associations to our compound and constituent stimuli, we checked for all stimulus-association pairs whether and which dependency paths exist between the stimuli and the associations.

The analysis relies on dependency parses of the *SdeWaC* corpus (see above) by Bohnet's MATE dependency parser (Bohnet, 2010). For each occurrence of a stimulus in the parsed corpus, we

5 Schulte im Walde & Müller (2013) performed a detailed comparison of stimulus- associate corpus co-occurrence across corpora, varying the corpus sizes and the corpus domains.

first checked for each association whether it appeared in the same sentence. In the positive case, we determined the shortest dependency path between the stimulus and the association. The analysis is again token-based: We took the strength between the stimulus-association pairs into account, i.e., how often the association was provided for a stimulus.

Table 7 presents the 10 most frequent paths that exist between the stimuli and the associations, accompanied by a stimulus-associate pair example and by the total frequency indicating how often the paths appeared over all stimulus-association pairs. The paths start with the part-of-speech of the stimulus and end with the part-of-speech of the associate response. The tag 'NN' refers to common nouns; 'ADJA' to attributive adjectives, 'APPR(ART)' to (portmanteau) prepositions, 'KON' to conjunctions, and any part-of-speech starting with 'VV' to a main verb.

Path	Example Stimulus–Associate Pair	Total Frequency
NN/NN	*Tasse* `cup' – *Kaffee* `coffee'	3,377
NN/APPR/NN	*Hütte* `shed' – *Holz* `wood'	2,357
NN/ADJA	*Salat* `salad' – *grün* `green'	3,343
NN/KON/NN	*Hund* `dog' – *Katze* `cat'	2,235
NN/VVFIN	*Buch* `book' – *lesen* `read'	1,163
NN/VVFIN/APPR/NN	*Kokosnuss* `coconut' – *Palme* `palm'	949
NN/VVFIN/NN	*Maulwurf* `mole' – *Hügel* `molehill'	941
NN/APPR/VVFIN/NN	*Papierkorb* `bin' – *Abfall* `trash'	888
NN/APPR/VVFIN	*Brille* `glasses' – *watch* `sehen'	661
NN/APPRART/NN	*Hut* `hat' – *Kopf* `head'	610

Table 7. Most frequent dependency paths between stimuli and assciates.

The semantics of the paths (at least of those that are similarly short as in the examples) is quite obvious. For example, 'NN/KON/NN' refers to the stimulus and the association appearing as two conjuncts (such as *Hund* `dog' and *Katze* `cat'); 'NN/ADJA' refers to an attributive adjective that directly modifies the noun stimulus (such as *grün* `green' for *Salat* `salad'), 'NN/APPR/NN' to a syntactic construction where either the stimulus or the association depends on a preposition that itself depends on the other noun (such as *Hütte*/*aus*/*Holz* `shed/made from/wood'). One could specify the paths in more detail (e.g., by adding the dependency direction, or by specifying the prepositional heads), or generalise more about some aspects (e.g., by summarising over the various forms of the verbs). However, the current version seems appropriate for a first impression of the syntactic dependencies between the stimuli and their associations.

Table 8 lists the five strongest examples for the five most prominent syntactic paths.[6] The path strength is that portion of the association strength for which the respective path between the stimulus and the association was found in the parsed corpus (as there might have been other paths between the stimulus-association pair). For example, 74.39% of the dependency paths between

6 Note that for ambiguous stimuli and associations, only the translation of the respective sense is given.

Tasse `cup' and *Kaffee* `coffee' were NN/NN, so given an association strength of 36, the path strength is 36 * 0.7439 = 26.78.

In order to look into the path instances, we checked on the actual corpus parses of the sentences where stimuli and associations appeared with the given paths. In most cases, our initial intuitions about the semantics of the paths were confirmed. For example, the NN/ADJA cases all refer to adjectival noun modifiers; the NN/KON/NN cases refer to conjuncts (often representing collocations) such as *Pfeil und Bogen* `bow and arrow', *Hund und Katze* `dog and cat', and *Kaffee und Kuchen* `coffee and cake'; and the NN/VVFIN cases refer to nominal complements of verbs (such as subjects as in *Hund-bellen* `dog barks' and *Telefon-klingeln* `phone rings', or direct objects as in *Brief-schreiben* `write letter'). The NN/NN cases, however, refer to measure constructions such as *Tasse Tee/Kaffee* `a cup of tea/coffee', or genitive constructions such as *Dach eines Hauses* `roof of a house', which we had not predicted in advance. Looking into actual sentences also helps to specify the NN/APPR/NN cases, by investigating the prepositional heads. For example, there are instances of *Zeit zwischen ... Uhr und ... Uhr* `time between ... o'clock and ... o'clock', and *Zeit von ... Uhr bis ... Uhr* `time from ... o'clock until ... o'clock'; *Thermoskanne mit Tee* `thermos with tea', and *Tee aus der Thermoskanne* `tea from the thermos'; *Lampe auf dem Nachttisch* `lamp on the bed table'; *Hütte aus Holz* `hut made out of wood'; *Wasser auf die Mühle* `water onto the mill' (note that the latter example is part of a collocation, meaning "grist to the mill", encouragement). In the NN/APPR/NN cases a refinement of the paths with prepositions types would obviously be a useful extension.

In sum, the dependency paths provide interesting and diverse insights into salient syntactic dependencies between the stimuli and the associations. To our knowledge, there is no previous related work that explored the syntactic paths in association norms. Rather, Schulte im Walde et al. (2008) explored parse tuples for noun-verb pairs to identify potential syntactic dependencies between stimulus nouns and associate verbs, and between stimulus verbs and associate nouns. The current analysis provides a broader spectrum of syntactic information, such that not only noun-verb pairs are involved. While the above information comprises both compound and constituent stimuli, the strongest paths for only compounds strongly overlap with this larger set.

Stimulus	Association	Association Strength	Path	Path Strength
Tasse `cup'	*Kaffee* `coffee'	36	NN/NN	26,78
Tasse `cup'	*Tee* `tea'	35		19,89
Dach `roof'	*Haus* `house'	67		18,50
Blatt `sheet'	*Papier* `paper'	27		17,78
Kanne `pot'	*Kaffee* `coffee'	40		17,20
Uhr `o'clock'	*Zeit* `time'	48	NN/APPR/NN	16,04
Thermoskanne `thermos'	*Tee* `tea'	34		10,66
Nachttisch `bed table'	*Lampe* `lamp'	21		10,50
Hütte `shed'	*Holz* `wood'	29		10,26
Mühle `mill'	*Wasser* `water'	16		9,29

Jeans `jeans'	*blau* `blue'	48	NN/ADJA	26,57
Licht `light'	*hell* `bright'	54		25,62
Salat `salad'	*grün* `green'	32		23,00
Ziegelstein `brick'	*rot* `red'	28		22,84
Messer `knife'	*scharf* `sharp'	28		21,15
Bogen `bow'	*Pfeil* `arrow'	52	NN/KON/NN	35,26
Hund `dog'	*Katze* `cat'	33		17,31
Kuchen `cake'	*Kaffee* `coffee'	21		14,98
Tisch `table'	*Stuhl* `chair'	31		13,60
Mann `man'	*Frau* `woman'	30		11,97
Hund `dog'	*bellen* `bark'	46	NN/VVFIN	20,52
Telefon `phone'	*klingeln* `ring'	22		18,76
Hahn `cock'	*krähen* `craw'	26		18,09
Brief `letter'	*schreiben* `write'	41		13,11
Buch `book'	*lesen* `read'	84		11,83

Table 8. Examples of stimulus-associate pairs and their dependency paths.

6.4 Semantic relations

The final analysis concerns types of semantic relationships (such as synonymy) that hold between the stimulus words and the associate responses. This analysis provides insight into which semantic relations the experiment participants might have had in mind when they provided associations. We rely on the German WordNet, *GermaNet (*Kunze, 2000), to explore the semantic information, and used GermaNet version 6.0, which was the latest version when the study started, comprising 69,594 synsets and 93,407 lexical units. We consider the paradigmatic relations *synonymy, antonymy, hypernymy, hyponymy*, and *co-hyponymy*. The synonymy information was based on the 69,594 synsets; in addition, GermaNet 6.0 contains a total of 74,945 hypernymy relations, and 1,587 antonymy relations.

The analysis is again token-based, i.e., incorporating the strengths between stimuli and associations. So if a relationship between a stimulus and an association is found, it is instantiated by the strength between them. For example, if the association *Obst* `fruit' was provided 12 times for the stimulus *Apfel* `apple', the strength of the semantic relation *hypernymy* that combined the two lexemes in GermaNet was instantiated by 12. In case another hypernym of *Apfel* (such as *Frucht* `fruit') was found among the associations, the stimulus-association strength was added to the existing hypernym strength. So on the one hand, we determine the strengths between individual stimulus-association pairs, and on the other hand we can summarise over all associations for a certain stimulus (and also for all stimuli) to identify the strengths of the various semantic relations with regard to a specific stimulus, and also over all stimuli. Table 9 presents the 20 strongest semantic relationships between individual compound stimuli and their associations.

In total, we found semantic relation information for 426 of our 571 stimuli (75%). If we only look at the compound stimuli, the proportion is much lower at 59%, covering associations to 145 out of the 246 compounds. It is very impressive, however, that all of the stimuli are included in GermaNet. That is, GermaNet covers not only the simple nouns in our data set but also all of the compounds. The missing proportions are due to non-existing relationships between the stimuli and the associations; for this, the coverage for the compounds is lower than for the whole set of stimuli, which is not surprising.

Stimulus	Association	Semantic Relation	Strength
Seezunge \`sole'	*Fisch* \`fish'	hypernym	31
Ohrring \`earring'	*Schmuck* \`jewellery'	hypernym	24
Mohrrübe \`carrot'	*Karotte* \`carrot'	synonym	22
Schlittenhund \`sledge dog'	*Husky* \`husky'	co-hyponym	21
Armbrust \`crossbow'	*Pfeil* \`arrow'	co-hyponym	19
Eisenbahn \`railway'	*Zug* \`train'	synonym	17
Wirbelsäule \`spine'	*Knochen* \`bone'	hypernym	16
Truthahn \`turkey cock'	*Pute* \`turkey hen'	co-hyponym	14
Meerschweinchen \`guinea pig'	*Haustier* \`pet'	hypernym	14
Federball \`badminton'	*Badminton* \`badminton'	co-hyponym	14
Taschentuch \`tissue'	*Tempo* \`tissue brand name'	hyponym	13
Bildschirm \`monitor'	*Fernseher* \`television'	co-hyponym	13
Kreditkarte \`credit card'	*VISA* \`VISA'	hyponym	11
Bullauge \`bull's eye'	*Fenster* \`window'	hypernym	11
Wäscheklammer \`clothespin'	*Wäscheleine* \`clothes line'	co-hyponym	10
Skistock \`ski stick'	*Ski* \`ski'	co-hyponym	10
Seemann \`seaman'	*Matrose* \`sailor'	co-hyponym	10
Fingerhut \`thimble'	*Nadel* \`needle'	co-hyponym	10
Brombeere \`blackberry'	*Himbeere* \`raspberry'	co-hyponym	10
Schornstein \`chimney'	*Kamin* \`chimney'	synonym	9

Table 9. Examples of semantic relations between stimulus-associate pairs.

Table 10 shows the proportions of GermaNet relations that have been identified for the stimulus-association pairs. The proportions in columns 2 and 3 correlate with the total proportions of semantic relation instances that are coded in GermaNet, so there are no specifically strong semantic relations among the stimulus-associate pairs. Looking at the associations to the compound stimuli in columns 4 and 5, however, demonstrates that with regard to the compound stimuli, it is more intuitive and thus easier to generate hypernyms and co-hyponyms than hyponyms, as the compounds are often very specific. There are only 26 types of hyponyms to the compound stimuli in total, with the most prominent examples being *Taschentuch-Tempo* and *Kreditkarte-VISA*, that appear among the top 20 pairs in Table 9. In sum, we found an interesting difference in the semantic relations of the stimulus-association pairs when looking at the compounds in comparison to all stimulus nouns.

	found in GermaNet			
semantic relation	*all stimuli*		*compound stimuli*	
co-hyponymy	2,410	43.00%	408	48,00%
hyponymy	1,324	23.00%	71	8,00%
hypernymy	1,249	22.00%	260	31,00%
synonymy	646	11.00%	113	13,00%
antonymy	38	1.00%	0	0,00%

Table 10. Proportions of semantic relations among stimulus-associate pairs according to GermaNet.

6.5 Associations and semantic transparency

Our main motivation to collect the association norms for the German noun compounds and their constituents was that the associations provide insight into the semantic properties of the compounds and their constituents and should therefore represent a valuable resource for cognitive and computational linguistics research on semantic transparency, and lexical semantics in general. More specifically, we are interested in the degree of semantic transparency of the compound nouns with regard to their constituents, and we are currently investigating whether the degree of overlap of associations is indeed an indicator of the semantic relatedness between the compounds and their constituents. The examples in Tables 2 to 4 gave a first impression that this might indeed be the case, and in the following we will provide more evidence for our hypothesis.

We apply a measure suggested by Schulte im Walde et al. (2012) that relies on a simple association overlap to predict the degree of semantic transparency of the experiment compound nouns: We use the proportion of shared associations of a compound and its constituent with respect to the total number of associations of the compound. The degree of semantic transparency of a compound noun is calculated with respect to each constituent of the compound. As an example of the calculation, when considering the 10 most frequent responses of the compound noun *Ahornblatt* \`maple leaf' and its constituents, as provided in Table 2, the compound noun received a total of 99 association tokens, out of which it shares 87 with the first constituent *Ahorn* \`maple', and 52 with the second constituent *Blatt* \`leaf'. Thus, the predicted degrees of semantic transparency are 87/99 = 0.88 for *Ahornblatt–Ahorn*, and 52/99 = 0.53 for *Ahornblatt–Blatt*. These predicted degrees of semantic transparency are compared against mean transparency judgements as collected by von der Heide & Borgwaldt (2009)[7], using the Spearman Rank-Order Correlation Coefficient ρ. Our hypothesis is that the larger the overlap of associations to a compound and a constituent, the stronger the degree of the compound-constituent transparency.

The resulting correlation values are ρ=.6227 (taking both constituent types into account), ρ=.6128 for the compound-modifier pairs and ρ=.6547 for the compound-head pairs. The simple association overlap measure therefore exceeds moderate correlation values, going up to .6547. Looking into the two experiments separately, the results are slightly better for the AMT norms than for the standard web data, and thus confirm the semantic usefulness of both experiment setups. In addition, the overlap measure is stronger for heads than for modifiers.

The case study shows that the associations provide insight into semantic properties of the compounds (and their constituents) that should be useful for models of compounds' semantic transparency. Specifically, the data seem to indicate that associations to compound nouns comprise associations to both head and modifier, i.e., they reflect meaning components of both constituents. As the overlap between constituents and compounds correlates with transparency ratings, we can conclude that associations to compound nouns contain fewer associations to their opaque constituents than to transparent constituents. That is, for a transparent compound like *Schlittenhund* `sledge dog' we expect to find associations to both components *Schlitten* `sledge' and *Hund* `dog', whereas in a partially opaque compound like *Fliegenpilz* `toadstool' associations might be more related to the transparent constituent *Pilz* `mushroom' than to the opaque constituent *Fliege* `fly/bow tie'. This result is in accordance with Libben's APPLE model (1994; 1998), that opaque constituents are not conceptually activated during processing.

7. Conclusion

The current study presented a collection of association norms for German noun compounds and their constituents. The collection has been conducted via crowdsourcing and was compared to an earlier web experiment. The norms are aimed to be of use for researchers on compound processing, containing association norms for not only compounds but also for their constituents, in order to provide a resource for stimulus selection in experimental studies (e.g., Sandra, 1990). Second, examining properties of associations to compounds and their constituents in detail, our data could provide more insight into the way in which the meanings of compounds are being processed and/or represented in the mental lexicon, i.e., complement the results of on-line experiments.

In total, the new association norms for 246 noun-noun compounds and their constituents comprise 58,652 associations (tokens), distributed over 26,004 stimulus-associate pairs (types). The quantitative contribution of the crowdsourcing experiment is slightly larger than that of the previous web experiment (58,652 vs. 47,249 stimulus-associate tokens), even though it only ran a quarter of the time.

---

7 von der Heide & Borgwaldt (2009) collected transparency ratings for their 450 compounds: 30 native German speakers were asked to rate the semantic transparency of the compounds with respect to each of their constituents on a scale from 1 (opaque) to 7 (transparent). For more details on the ratings see Schulte im Walde, Müller, & Roller (2013).

We analysed the stimulus-associate pairs of both the current and the previous experiments in four ways, performing (1) a morpho-syntactic analysis, (2) a distributional analysis, (3) a syntactic analysis, and (4) a semantic relation analysis. Concerning (1) a part-of-speech analysis of the associate responses and (3) a dependency path analysis of the stimulus-associate pairs, there were no noticable differences when comparing the associations to compounds and the associations to nouns in general. Concerning (2) the window co-occurrence of stimulus-associate pairs and (4) the semantic relations between stimuli and associations, we however demonstrated that

- while we could confirm the co-occurrence hypothesis for our compound nouns, the corpus coverage pointed to the caveat that compounds are typically covered worse by corpus data than simple nouns. This demonstrates the importance of a reasonable morphological annotation of compounds, and that we have considerably less corpus information for compound nouns in comparison to nouns in general;
- the semantic relations between associations and compound stimuli differ to those between associations and nouns in general. Not surprisingly, we find less hyponyms among the associations but more hypernyms and co-hyponyms.

Finally, the overlap between associations to compounds and associations to their constituents correlated moderately with semantic transparency ratings obtained by human raters (von der Heide & Borgwaldt, 2009), providing additional insight into how the meaning of complex words is related to the meaning of their parts. In summary, the associations could in general represent a valuable resource concerning the lexical semantic properties of the compound stimuli and the semantic relations between the stimuli and their associations, and – more specifically – they could be useful in future cognitive and computational linguistics research on semantic transparency.

Acknowledgements

This work was supported by the DFG Heisenberg Fellowship SCHU-2580/1 (Sabine Schulte im Walde). We would also like to thank Ronny Jauch, who performed parts of the analyses in Section 6 within his Diploma Thesis (Jauch, 2012), and the anonymous reviewers for their valuable comments.

Appendix A

In the following, we provide a list of our subset of bi-morphemic noun compounds from von der Heide & Borgwaldt (2009), including the compound constituents, and the categorisation label for „NN“ (noun-noun compound) vs. „unique“ (noun compound with unique modifier).

Ahornblatt	Ahorn	Blatt	NN
Armband	Arm	Band	NN
Armbrust	Arm	Brust	NN
Aschenbecher	Asche	Becher	NN
Bahnhof	Bahn	Hof	NN
Bananenschale	Banane	Schale	NN
Bärlauch	Bär	Lauch	NN
Basketball	Basket	Ball	NN
Baumhaus	Baum	Haus	NN
Bettwäsche	Bett	Wäsche	NN
Bierfass	Bier	Fass	NN
Bildschirm	Bild	Schirm	NN
Billiardtisch	Billiard	Tisch	NN
Bleistift	Blei	Stift	NN
Blockflöte	Block	Flöte	NN
Blumenkohl	Blume	Kohl	NN
Blumenstrauß	Blume	Strauß	NN
Blumenvase	Blume	Vase	NN
Brautkleid	Braut	Kleid	NN
Briefkasten	Brief	Kasten	NN
Briefmarke	Brief	Marke	NN
Briefpapier	Brief	Papier	NN
Briefträger	Brief	Träger	NN
Brillenetui	Brille	Etui	NN
Brombeere	Brom	Beere	unique
Bullauge	Bulle	Auge	NN
Büroklammer	Büro	Klammer	NN
Cowboyhut	Cowboy	Hut	NN
Cowboystiefel	Cowboy	Stiefel	NN
Dachboden	Dach	Boden	NN
Dachfenster	Dach	Fenster	NN
Dachstuhl	Dach	Stuhl	NN

Dominostein	Domino	Stein	NN
Eidechse	Eid	Echse	unique
Eisbär	Eis	Bär	NN
Eisberg	Eis	Berg	NN
Eisenbahn	Eisen	Bahn	NN
Eisstadion	Eis	Stadion	NN
Eiswürfel	Eis	Würfel	NN
Ellenbogen	Elle	Bogen	NN
Erdbeere	Erde	Beere	NN
Erdnuss	Erde	Nuss	NN
Espressomaschine	Espresso	Maschine	NN
Federball	Feder	Ball	NN
Federboa	Feder	Boa	NN
Feldsalat	Feld	Salat	NN
Feuerwerk	Feuer	Werk	NN
Feuerzeug	Feuer	Zeug	NN
Fieberthermometer	Fieber	Thermometer	NN
Filzstift	Filz	Stift	NN
Fingerhut	Finger	Hut	NN
Fleischwolf	Fleisch	Wolf	NN
Fliegenklatsche	Fliege	Klatsche	NN
Fliegenpilz	Fliege	Pilz	NN
Flohmarkt	Floh	Markt	NN
Fotoalbum	Foto	Album	NN
Fotoapparat	Foto	Apparat	NN
Fußball	Fuß	Ball	NN
Fußleiste	Fuß	Leiste	NN
Gesangsbuch	Gesang	Buch	NN
Gewächshaus	Gewächs	Haus	NN
Glockenspiel	Glocke	Spiel	NN
Golfball	Golf	Ball	NN
Gummibärchen	Gummi	Bärchen	NN
Gummiente	Gummi	Ente	NN

Gummistiefel	Gummi	Stiefel	NN
Gürteltier	Gürtel	Tier	NN
Haarreifen	Haar	Reifen	NN
Hahnenfuß	Hahn	Fuß	NN
Halskette	Hals	Kette	NN
Handcreme	Hand	Creme	NN
Handschuh	Hand	Schuh	NN
Handtasche	Hand	Tasche	NN
Handtuch	Hand	Tuch	NN
Haselnuss	Hasel	Nuss	NN
Hausboot	Haus	Boot	NN
Heidelbeere	Heide	Beere	NN
Heuhaufen	Heu	Haufen	NN
Himbeere	Him	Beere	unique
Hirschkäfer	Hirsch	Käfer	NN
Hollywoodschaukel	Hollywood	Schaukel	NN
Hornbrille	Horn	Brille	NN
Hufeisen	Huf	Eisen	NN
Hundehütte	Hund	Hütte	NN
Hüttenkäse	Hütte	Käse	NN
Jägerzaun	Jäger	Zaun	NN
Jeanshemd	Jeans	Hemd	NN
Jeansjacke	Jeans	Jacke	NN
Kaffeemaschine	Kaffee	Maschine	NN
Kaffeemühle	Kaffee	Mühle	NN
Kaffeepad	Kaffee	Pad	NN
Kettensäge	Kette	Säge	NN
Kleiderschrank	Kleider	Schrank	NN
Knoblauch	Knob	Lauch	unique
Kokosnuss	Kokos	Nuss	NN
Kopfkissen	Kopf	Kissen	NN
Kopfsalat	Kopf	Salat	NN
Kreditkarte	Kredit	Karte	NN

Kreissäge	Kreis	Säge	NN
Kronkorken	Krone	Korken	NN
Kronleuchter	Krone	Leuchter	NN
Kuckucksuhr	Kuckuck	Uhr	NN
Kuhfladen	Kuh	Fladen	NN
Kulturbeutel	Kultur	Beutel	NN
Kürbiskern	Kürbis	Kern	NN
Lachsschinken	Lachs	Schinken	NN
Latzhose	Latz	Hose	NN
Leberwurst	Leber	Wurst	NN
Lederhose	Leder	Hose	NN
Lichtschalter	Licht	Schalter	NN
Löwenzahn	Löwe	Zahn	NN
Luftmatratze	Luft	Matratze	NN
Luftpumpe	Luft	Pumpe	NN
Maisfeld	Mais	Feld	NN
Maiskolben	Mais	Kolben	NN
Marienkäfer	Maria	Käfer	NN
Maßstab	Maß	Stab	NN
Maulwurf	Maul	Wurf	NN
Mausefalle	Maus	Falle	NN
Meerschweinchen	Meer	Schweinchen	NN
Mettwurst	Mett	Wurst	NN
Mikadostäbchen	Mikado	Stäbchen	NN
Milchshake	Milch	Shake	NN
Mohrrübe	Mohr	Rübe	NN
Motorhaube	Motor	Haube	NN
Motorrad	Motor	Rad	NN
Mülleimer	Müll	Eimer	NN
Mülltonne	Müll	Tonne	NN
Mundharmonika	Mund	Harmonika	NN
Nachttisch	Nacht	Tisch	NN
Nadelkissen	Nadel	Kissen	NN

Nagelfeile	Nagel	Feile	NN
Nagellack	Nagel	Lack	NN
Nasenbär	Nase	Bär	NN
Nashorn	Nase	Horn	NN
Nilpferd	Nil	Pferd	NN
Nudelholz	Nudel	Holz	NN
Nummernschild	Nummer	Schild	NN
Obstkuchen	Obst	Kuchen	NN
Ohrring	Ohr	Ring	NN
Papierkorb	Papier	Korb	NN
Pfannkuchen	Pfanne	Kuchen	NN
Pfauenfeder	Pfau	Feder	NN
Pfeffermühle	Pfeffer	Mühle	NN
Postbote	Post	Bote	NN
Postkarte	Post	Karte	NN
Pudelmütze	Pudel	Mütze	NN
Radkappe	Rad	Kappe	NN
Reetdach	Reet	Dach	NN
Regenbogen	Regen	Bogen	NN
Regenmantel	Regen	Mantel	NN
Regenrinne	Regen	Rinne	NN
Regenschirm	Regen	Schirm	NN
Ringfinger	Ring	Finger	NN
Rittersporn	Ritter	Sporn	NN
Rucksack	Ruck	Sack	NN
Sandburg	Sand	Burg	NN
Sanduhr	Sand	Uhr	NN
Schachbrett	Schach	Brett	NN
Schildkröte	Schild	Kröte	NN
Schlauchboot	Schlauch	Boot	NN
Schlittenhund	Schlitten	Hund	NN
Schlüsselbund	Schlüssel	Bund	NN
Schneckenhaus	Schnecke	Haus	NN

Schneeball	Schnee	Ball	NN
Schneemann	Schnee	Mann	NN
Schnittlauch	Schnitt	Lauch	NN
Schornstein	Schorn	Stein	unique
Schulbuch	Schule	Buch	NN
Schwertfisch	Schwert	Fisch	NN
Seehund	See	Hund	NN
Seemann	See	Mann	NN
Seerose	See	Rose	NN
Seestern	See	Stern	NN
Seezunge	See	Zunge	NN
Seifenblase	Seife	Blase	NN
Seilbahn	Seil	Bahn	NN
Sessellift	Sessel	Lift	NN
Skistock	Ski	Stock	NN
Sonnenblume	Sonne	Blume	NN
Sonnenbrille	Sonne	Brille	NN
Sonnencreme	Sonne	Creme	NN
Sonnenschirm	Sonne	Schirm	NN
Sonnenuhr	Sonne	Uhr	NN
Spiegelei	Spiegel	Ei	NN
Spinnennetz	Spinne	Netz	NN
Stachelbeere	Stachel	Beere	NN
Stacheldraht	Stachel	Draht	NN
Stereoanlage	Stereo	Anlage	NN
Sternschnuppe	Stern	Schnuppe	NN
Strandkorb	Strand	Korb	NN
Straßenbahn	Straße	Bahn	NN
Strohhalm	Stroh	Halm	NN
Strumpfhose	Strumpf	Hose	NN
Suppenteller	Suppe	Teller	NN
Tannenzapfen	Tanne	Zapfen	NN
Taschenbuch	Tasche	Buch	NN

Taschenlampe	Tasche	Lampe	NN
Taschenmesser	Tasche	Messer	NN
Taschentuch	Tasche	Tuch	NN
Teddybär	Teddy	Bär	NN
Teebeutel	Tee	Beutel	NN
Teekanne	Tee	Kanne	NN
Teelicht	Tee	Licht	NN
Teelöffel	Tee	Löffel	NN
Teetasse	Tee	Tasse	NN
Telefonbuch	Telefon	Buch	NN
Telefonhörer	Telefon	Hörer	NN
Telefonzelle	Telefon	Zelle	NN
Tennisball	Tennis	Ball	NN
Tennisschläger	Tennis	Schläger	NN
Thermoskanne	Thermo	Kanne	unique
Tintenfisch	Tinte	Fisch	NN
Toilettenpapier	Toilette	Papier	NN
Truthahn	Trut	Hahn	unique
Türklinke	Tür	Klinke	NN
Vanilleeis	Vanille	Eis	NN
Visitenkarte	Visite	Karte	NN
Vogelhaus	Vogel	Haus	NN
Vogelkäfig	Vogel	Käfig	NN
Walnuss	Wal	Nuss	unique
Wäscheklammer	Wäsche	Klammer	NN
Wasserfall	Wasser	Fall	NN
Wasserhahn	Wasser	Hahn	NN
Wassermelone	Wasser	Melone	NN
Wasserwaage	Wasser	Waage	NN
Weihnachtsbaum	Weihnachten	Baum	NN
Werwolf	Wer	Wolf	unique
Windlicht	Wind	Licht	NN
Windmühle	Wind	Mühle	NN

Wintermantel	Winter	Mantel	NN
Wirbelsäule	Wirbel	Säule	NN
Wollschal	Wolle	Schal	NN
Würfelzucker	Würfel	Zucker	NN
Zahnbürste	Zahn	Bürste	NN
Zahnkrone	Zahn	Krone	NN
Zahnpasta	Zahn	Pasta	NN
Zahnrad	Zahn	Rad	NN
Zahnseide	Zahn	Seide	NN
Zahnspange	Zahn	Spange	NN
Zeitschrift	Zeit	Schrift	NN
Ziegelstein	Ziegel	Stein	NN
Zitronenpresse	Zitrone	Presse	NN
Zollstock	Zoll	Stock	NN
Zuckerhut	Zucker	Hut	NN
Zuckerwatte	Zucker	Watte	NN

Appendix B

In the following, we provide the experiment setups for the association collection via AMT in the current study, and for the association collection through a standard web experiment in the previous study (Schulte im Walde et al., 2012). The AMT setup of the current collection is shown by Figure 1, with translations in red font. The actual collection of the associations was performed as shown in Figure 2, translated by Figure 3. The web collection of the associations was performed as shown in Figure 4, translated by Figure 5.

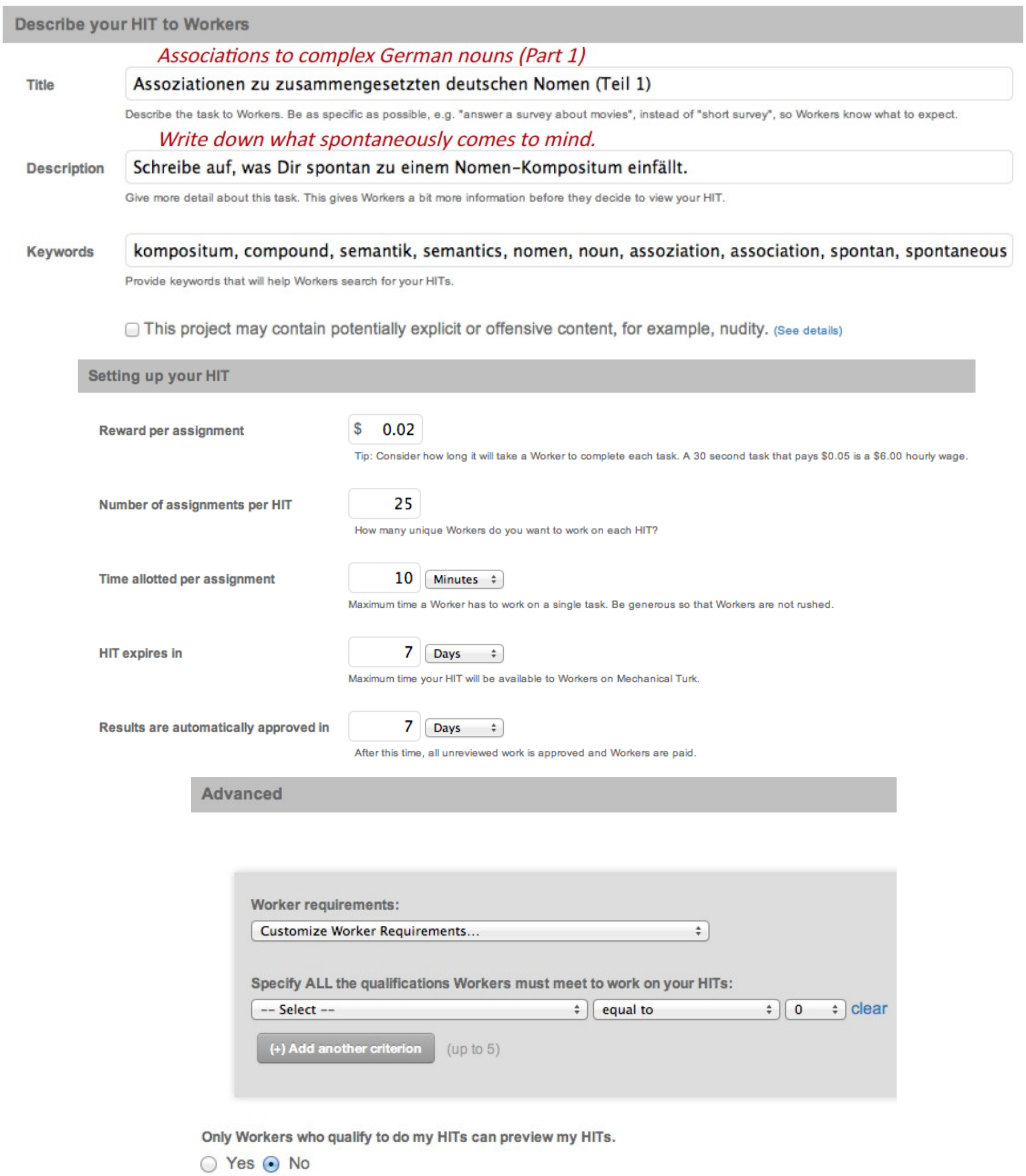


Figure 1. AMT setup.

Assoziationen zu deutschen Nomen

In diesem Experiment geht es um Nomen-Assoziationen.

Aufgabe: Wir präsentieren Nomen, und Du schreibst Wörter auf, die Dir **spontan** dazu einfallen. Pro Nomen stehen bis zu drei Felder zur Verfügung.

Beispiel: Woran denkst Du spontan bei dem Wort ***Schnee***?
Winter, kalt, Schlitten

Wichtige Hinweise:

- Nur für **deutsche Muttersprachler - Native speakers of German** only.
- Bitte nicht lange nachdenken, sondern **spontan** und **zügig** antworten.
- Bitte nur **ein Wort pro Zeile** schreiben.
- Bitte **alle 24 HITs** bearbeiten.
- Wir haben als Test einige Wörter in den Daten versteckt, die von Muttersprachlern als "erfunden" erkannt werden sollten. In einem solchen Fall bitte die Auswahl "Kenne ich nicht!" treffen.

Assoziationen zu ${word} :

Ich kenne das Nomen "*${word}*" nicht!

Figure 2. Background for AMT experiment.

<u>Associations</u> to German Nouns

This experiment is about associations to nouns.

Task: You will see nouns, and we ask you to list words that **spontaneously** come to mind. For each noun you will be provided three lines in which to list associations.

Example: What comes to mind when you read the word ***snow***?
winter, cold, sledge

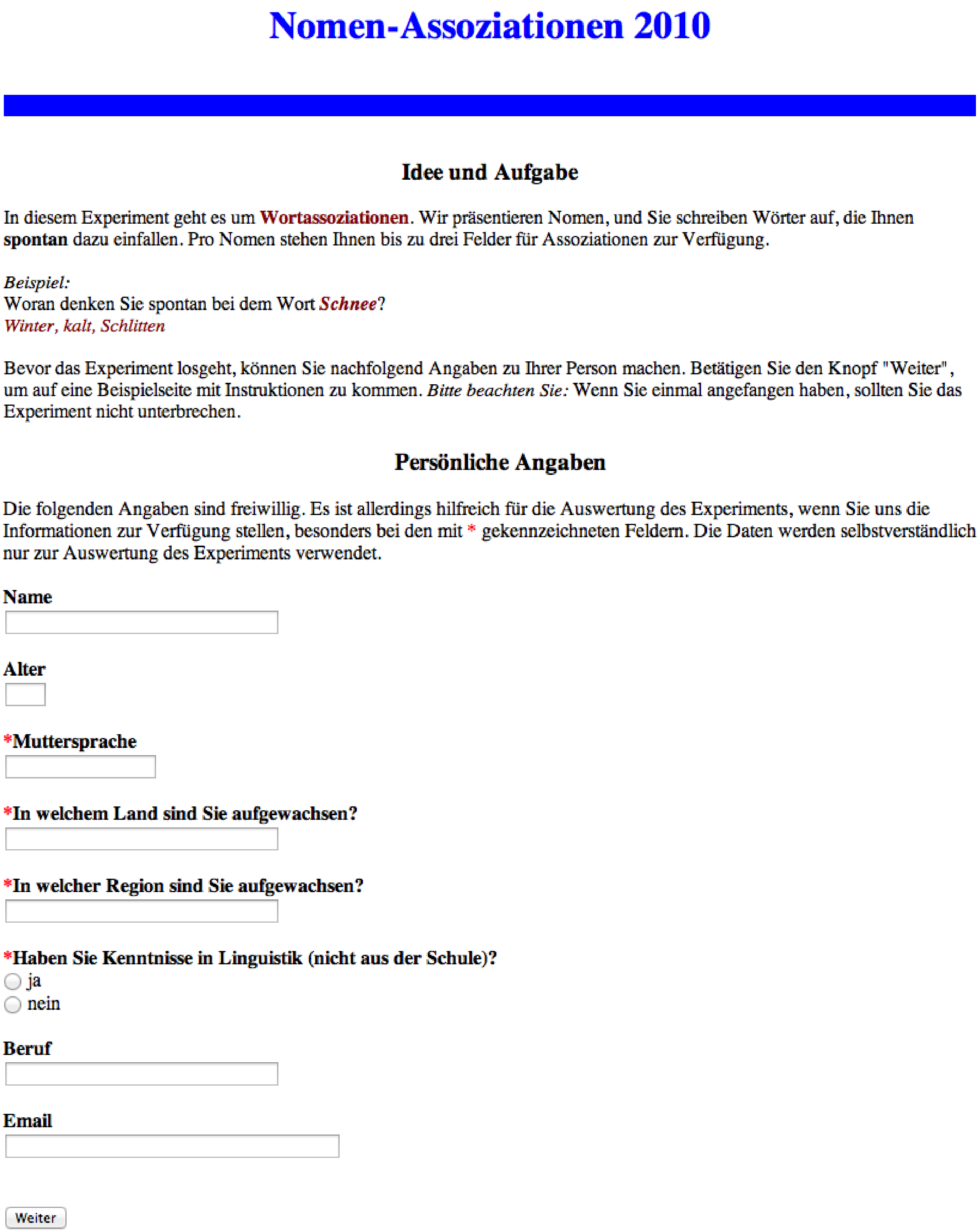
Important notes:

- The task is for **native speakers of German** only.
- Please respond **spontaneously** and **quickly**.
- Please only write **one word per line**.
- Please do **all 24 HITs** .
- As test cases to control for spammers, we hid some non-existing words in the data that you should identify as such. In these cases, please check the "I don't know!" button.

Associations to ${word} :

I don't know the noun "*${word}"!*

Figure 3. Background for AMT experiment, translated into English.



Nomen-Assoziationen 2010

Idee und Aufgabe

In diesem Experiment geht es um **Wortassoziationen**. Wir präsentieren Nomen, und Sie schreiben Wörter auf, die Ihnen **spontan** dazu einfallen. Pro Nomen stehen Ihnen bis zu drei Felder für Assoziationen zur Verfügung.

Beispiel:
Woran denken Sie spontan bei dem Wort ***Schnee***?
Winter, kalt, Schlitten

Bevor das Experiment losgeht, können Sie nachfolgend Angaben zu Ihrer Person machen. Betätigen Sie den Knopf "Weiter", um auf eine Beispielseite mit Instruktionen zu kommen. *Bitte beachten Sie:* Wenn Sie einmal angefangen haben, sollten Sie das Experiment nicht unterbrechen.

Persönliche Angaben

Die folgenden Angaben sind freiwillig. Es ist allerdings hilfreich für die Auswertung des Experiments, wenn Sie uns die Informationen zur Verfügung stellen, besonders bei den mit * gekennzeichneten Feldern. Die Daten werden selbstverständlich nur zur Auswertung des Experiments verwendet.

Name

Alter

***Muttersprache**

***In welchem Land sind Sie aufgewachsen?**

***In welcher Region sind Sie aufgewachsen?**

***Haben Sie Kenntnisse in Linguistik (nicht aus der Schule)?**
ja
nein

Beruf

Email

Weiter

Figure 4. Background to previous association experiment in Schulte im Walde et al. (2012).

Noun Associations 2010

Idea and Task

This experiment is about **word associations**. You will see nouns, and we ask you to list words that **spontaneously** come to mind. For each noun you will be provided three lines in which to list associations.

Example:
What comes to mind when you read the word ***snow***?
winter, cold, sledge

Before the experiment starts, we kindly ask you for some personal information. Then press the button "continue" to proceed to an example page with instructions. *Please note:* Once you have started the experiment, please do not pause until finished.

Personal Information

For evaluating the associations of the study, it would be very helpful for us to have as much background about the participants as possible, especially regarding fields marked by *. We will not use the data for anything other than evaluating the study.

Name

Age

***Native Language**

***In which country did you grow up?**

***In which region did you grow up?**

***Do you have a background in linguistics (professional or academic)?**

- yes
- no

Profession

Email

continue

Figure 5. Background to previous association experiment, translated into English.

References

Baayen, R.H. & Schreuder, R. (1999). War and peace: Morphemes and full forms in a noninteractive activation parallel dual-route model. *Brain and Language* 68, 27 - 32.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43 (3), 209-226.

Bertram, R., & Hyönä, J. (2003). The length of a complex word modifies the role of morphological structure: Evidence from reading short and long Finnish compounds. *Journal of Memory & Language* 48, 615-634.

Bertram, R., Kuperman, V., Baayen, R.H., & Hyönä, J. (2011). The hyphen as a segmentation cue in triconstituent compound processing: It's getting better all the time. *Scandinavian Journal of Psychology* 52, 530–544.

Bohnet, B. (2010). Top accuracy and fast dependency parsing is not a contradition. In: *Proceedings of the 23rd International Conference on Computational Linguistics.* Association for Computational Linguistics. 89-97.

Butterworth, B. (1983). Lexical Representation. In: Butterworth, B. (ed.) *Language Production*. vol 2. (pp. 257-294) London: Academic Press.

Caramazza, A., Laudanna, A., & Romani, C. (1988). Lexical access and inflectional morphology. *Cognition*, 28, 297-332.

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

Clark, H. (1971). Word associations and linguistic theory. In Lyons, J. (ed.). *New Horizons in Linguistics* (pp. 271-286). Baltimore: Penguin.

Comesaña, M., Fraga, I., Moreira, A., Frade, C., & Soares, A. (2014). Free associate norms for 139 European Portuguese words for children from different age groups. *Behavior Research Methods* 46, 564–574.

de Deyne, S., & Storms, G. (2008). Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods,* 40, 198-205.

de Jong, N., Feldman, L., Schreuder, R., Pastizzo, M., & Baayen, R.H. (2002). The processing and representation of Dutch and English compounds: Peripheral morphological and central orthographic effects. *Brain and Language* 81 (1), 555-567.

Faaß, G., & Eckart, K. (2013). SdeWaC – A corpus of parsable sentences from the web. *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology.* (pp. 61-68). Darmstadt: Gesellschaft für Sprachtechnologie und Computerlinguistik, 61-68.

Fellbaum, C. (1995). Co-occurrence and antonymy. *Lexicography* 8(4), 281-303.

Fellbaum, C. (1998). *WordNet – An Electronic Lexical Database*. Cambridge: MIT Press.

Fellbaum, C., & Chaffin, R. (1990). Some principles of the organization of verbs in the mental lexicon. In: *Proceedings of the 12th Annual Conference of the Cognitive Science Society of America.* Association for Computational Linguistics. 420-427.

Fernández, A., Diez, E., & Alonso, M., & Beato, M. S. (2004). Free-association norms for the Spanish names of the Snodgrass and Vanderwart pictures. *Behavior Research Methods, Instruments and Computers* 36(3), 577-583.

Ferrand, L., & Alario, F.-X. (1998). French word association norms for 366 names of objects. *L'Année Psychologique* 98(4), 659-709.

Fleischer, W. & Barz, I. (2012). *Wortbildung der deutschen Gegenwartssprache* (4th edition). Berlin, Boston: de Gruyter.

French, C., & Richards, A. (1992). Word association norms for a set of threat/neutral homographs. *Cognition & Emotion,* 6, 65-87.

Gagné, C., & Spalding, T. (2009). Constituent integration during the processing of compound words: Does it involve the use of relational structures? *Journal of Memory and Language* 60, 20-35.

Guida, A. (2007). *The representation of verb meaning within lexical semantic memory: Evidence from word associations*. Master's Thesis, Università degli studi di Pisa.

Heringer, H. J. (1986). The verb and its semantic power: Association as the basis for valence. *Journal of Semantics* 4, 79-99.

Hirsh, K., & Tree, J. (2001). Word association norms for two cohorts of British adults. *Journal of Neurolinguistics* 14(1), 1-44.

Janssen, N., Bi, Y., & Caramazza, A. (2008). A tale of two frequencies: Determining the speed of lexical access for Mandarin Chinese and English compounds. *Language and Cognitive Processes* 23 (7-8), 1191-1223.

Jauch, R. (2012). *Empirische Analysen von Assoziationen und distributionelle Modellierung der Kompositionalität von deutschen Nomen-Komposita.* M.Sc. Thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

John, C. H. (1988). Emotionality ratings and free-association norms of 240 emotional and non-emotional words. *Cognition & Emotion,* 2, 49-70.

Kiss, G., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In: Aitken, A., Bailey, R. & Hamilton-Smith, N. (eds.). *The Computer and Literary Studies*. (pp. 153–165). Edinburgh: Edinburgh University Press.

Klos, V. (2011). *Komposition und Kompositionalität.* Berlin: Walter de Gruyter.

Kremer, G., & Baroni, M. (2011). A set of semantic norms for German and Italian. *Behavior Research Methods,* 43, 97-109.

Krott, A., Schreuder, R., Baayen, R.H., & Dressler, W. (2007). Analogical effects on linking elements in German compound words. *Language and Cognitive Processes* 22 (1), 25-57.

Kunze, C. (2000). Extension and use of GermaNet, a lexical-semantic database. In: *Proceedings of the 2nd International Conference on Language Resources and Evaluation.* Association for Computational Linguistics. 999-1002.

Lauteslager, M., Schaap, T., & Schievels, D. (1986). *Schriftelijke Woordassociatienormen voor 549 Nederlandse Zelfstandige Naamworden.* Lisse: Swets and Zeitlinger.

Libben, G. (1994). How is morphological decomposition achieved? *Language and Cognitive Processes,* 9 (3), 369–391.

Libben, G. (1998). Semantic transparency in the processing of compounds: Consequences for representation, processing, and impairment. *Brain and Language* 61, 30–44.

Libben, G., Gibson, M., Yoon, Y. B., & Sandra, D. (2003). Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language* 84, 50–64.

Longtin, C.-M., Segui, J., & Hallé, P. (2003). Morphological priming without morphological relationship. *Language and Cognitive Processes* 18 (3), 313-334.

Lüttmann, H., Zwitserlood, P., Böhl, A., & Bölte, J. (2011). Evidence for morphological composition at the form level in speech production. *Journal of Cognitive Psychology* 23 (7), 818-836

Macizo, P., Gómez-Ariza, C., & Bajo, M. (2000). Associative norms of 58 Spanish words for children from 8 to 13 years old. *Psicológica,* 21, 287-300.

Marslen-Wilson, W., Tyler, L., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review* 101 (1), 3-33.

McNamara, T. (2005). *Semantic Priming: Perspectives from Memory and Word Recognition.* New York: Psychology Press, New York.

Melinger, A., Schulte im Walde, S. & Weber, A. (2006). Characterizing response types and revealing noun ambiguity in German association norms. In: *Proceedings of the EACL Workshop "Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics together"*. Association for Computational Linguistics. 41-48.

Melinger, A., & Weber, A. (2006). *Database of noun associations for German.* http://www.coli.uni-saarland.de/projects/nag/.

Miller, G. (1969). The organization of lexical memory: Are word associations sufficient? In Talland, G., & Waugh, N. (eds.). *The Pathology of Memory*. New York: Academic Press, 223-237.

Nelson, D., McEvoy, C., & Schreiber, T. (2004). The University of South Florida Free Association, Rhyme, and Word Fragment Norms. *Behavior Research Methods, Instruments, & Computers,* 36, 402-407.

Palermo, D., & Jenkins, J. (1964). *Word Association Norms: Grade School through College.* Minneapolis: University of Minnesota Press, Minneapolis.

Rapp, R. (2002). The computation of word associations: Comparing syntagmatic and paradigmatic approaches. In: *Proceedings of the 19th International Conference on Computational Linguistics*. Association for Computational Linguistics.

Reddy, S., McCarthy, D., & Manandhar, S. (2011). An empirical study on compositionality in compound nouns. In: *Proceedings of the 5th International Joint Conference on Natural Language Processing*, Chaing Mai, Thailand, 210-218.

Roller, S., Schulte im Walde, S., & Scheible, S. (2013). The (un)expected effects of applying standard cleansing models to human ratings on compositionality. In: *Proceedings of the 9th Workshop on Multiword Expressions.* Association for Computational Linguistics, 32-41.

Roth, M., & Schulte im Walde, S. (2008). Corpus co-occurrence, dictionary and wikipedia entries as resources for semantic relatedness information. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation.* Association for Computational Linguistics, 1852-1859.

Russell, W. (1970). The complete German language norms for responses to 100 words from the Kent-Rosanoff word association test. In: Postman, L., & Keppel, G. (eds.). *Norms of Word Association.* (pp. 53-94) New York: Academic Presss..

Russell, W., & Meseck, O. (1959). Der Einfluss der Assoziation auf das Erinnern von Worten in der deutschen, französischen und englischen Sprache. *Zeitschrift für Experimentelle und Angewandte Psychologie* 6, 191-211.

Sandra, D. (1990). On the representation and processing of compound words: Automatic access to constituent morphemes does not occur. *The Quarterly Journal of Experimental Psychology* 42, 529-567.

Schröder, A., Gemballa, T., Ruppin, S., & Wartenburger, I. (2012). German norms for semantic typicality, age of acquisition, and concept familiarity. *Behavior Research Methods,* 44, 380-394.

Schulte im Walde, S. (2008). Human associations and the choice of features for semantic verb classification. *Research on Language and Computation* 6(1), 79-111.

Schulte im Walde, S., & Melinger, A. (2008). An in-depth look into the co-occurrence distribution of semantic associates. In: Lenci, A. (guest editor). *Italian Journal of Linguistics. Special Issue on "From Context to Meaning: Distributional Models of the Lexicon in Linguistics and Cognitive Science"* 20(1), 89-128.

Schulte im Walde, S., Melinger, A., Roth, M., & Weber, A. (2008). An empirical characterisation of response types in German association norms. *Research on Language and Computation* 6(2), 205-238.

Schulte im Walde, S., Borgwaldt, S., & Jauch, R. (2012). Association norms of German noun compounds. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC).* Association for Computational Linguistics. 632-639

Schulte im Walde, S., & Müller, S. (2013). Using web corpora for the automatic acquisition of lexical-semantic knowledge. *Journal for Language Technology and Computational Linguistics* 28(2), 85-105. In: Schneider, R., Storrer, A., & Mehler, A. (guest editors). *Special Issue on "Web Corpora for Computational Linguistics and Linguistic Research".*

Schulte im Walde, S., Müller, S., & Roller, S. (2013). Exploring vector space models to predict the compositionality of German noun-noun compounds. In: *Proceedings of the 2nd Joint Conference on Lexical and Computational* Semantics. Association for Computational Linguistics. 255-265.

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

Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. *Quarterly Journal of Experimental Psychology Section A* 57 (4), 745-765.

Taft, M. & Forster, K. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior* 14 (6), 638-647.

Zwitserlood, P. (1994). Processing and representation of Dutch compounds: Effects of semantic transparency. *Language and Cognitive Processes 9* (3), 341-368.

van Jaarsveld, H., & Rattink, G. (1988). Frequency effects in the processing of lexicalized and novel nominal compounds. *Journal of Psycholinguistic Research 17*, 447-473.

von der Heide, C., & Borgwaldt, S. (2009). Assoziationen zu Unter-, Basis- und Oberbegriffen. Eine explorative Studie. In Vogel, R., & Sahel, S. (eds.). *Proceedings of the 9th Norddeutsches Linguistisches Kolloquium,* 51-74.