Building "Prototype Definitions" for the Monolingual Dictionary – Results of a Typicality Judgement Study

Online Proceedings of UK-CLA Meetings
http://uk-cla.org.uk/proceedings
Vol 4: 205 – 229

ANDREAS WIRAG

Graduate School "Teaching and Learning Processes",
Universität Koblenz-Landau

wirag@uni-landau.de

Abstract

The paper reports the outcome of a typicality judgement study used to create "prototype definitions" for the monolingual learner's dictionary (MLD). The study investigated whether a CL approach to word meaning, which relies on Prototype Semantics in the Roschian tradition (Rosch, 1975; Rosch & Mervis, 1975), might be employed to create innovative "prototype definitions" for the monolingual dictionary. Prototype Semantics predicts that different definition stimuli for the same headword, like other instantiations of lexical concepts, ought to exhibit effects of graded typicality (or "goodness-of-exemplar" effects) with reference to the headword. To investigate the assumption, a typicality judgement study was conducted using Likert-type scales to rate definition stimuli on their typicality towards a dictionary headword (N = 34). Non-parametric Kruskal–Wallis and post-hoc Wilcoxon tests demonstrate that (a) the investigated definition stimuli in fact exhibit typicality effects with reference to the headword, and that (b) ideal, or most typical, "prototype definitions" could be obtained through the procedure. The paper discusses implications of the study findings for Prototype Semantics and the potential impact of "prototype definitions" on a practical lexicography.

Key words: Applied Cognitive Linguistics, Monolingual Learner's Dictionary, Cognitive Lexicography, Prototype Semantics

1. Introduction

In recent decades, lexicographical research on the efficiency of the monolingual learner's dictionary (MLD) in second-language (L2) learning has produced a string of somewhat mixed results (for overviews, see, e.g., Béjoint, 2010; Nesi, 2014).[1] While a number of studies attests to the positive impact of the monolingual dictionary on L2 reading and L2 vocabulary learning (e.g., Knight, 1994; Tono, 2001, pp. 75–83; Nesi & Haill, 2002; Ronald, 2002; Szczepaniak, 2006), a sizable number of investigations fails to detect a positive influence of the MLD on these L2 competencies (e.g., Bensoussan, Sim, & Weiss, 1984; Nesi & Meara, 1991; McCreary & Dolezal, 1999; Nesi, 2000, pp. 61–64). In sum, as Béjoint (2010) qualifies the net results of these experimental efforts, "it is reasonable, and preferable for lexicography, to think that dictionaries are useful, in some cases, at least, and to some people" (p. 254).[2]

Qualitative studies that focus on the individual dictionary user suggest that students are prone to struggle with the definition "format" – or defining "style" – adopted by the monolingual learner's dictionary (e.g., McKeown, 1993; Nesi & Meara, 1994; Nesi & Haill, 2002; Beck, McKeown, & Kucan, 2013). In this manner, the dominant defining format of the MLD, the "traditional definition" style, used by major publishers such as the *Oxford Advanced Learner's Dictionary*, *Cambridge Advanced Learner's Dictionary*, or *Longman Dictionary of Contemporary English*, is based on an openly structuralist conception of word semantics. This definition format, therefore, rests on a feature-based approach to word meaning and defines a given dictionary headword by means of a listing (or enumeration) of the essential "semantic features", or atomistic sense components, that are considered to represent the headword sense (e.g., Svensén, 2009, ch. 13; Geeraerts, 2010).[3]

While this feature-based format of defining entry words to dictionary users represents a well-entrenched and time-honoured tradition in lexicography, researchers emphasise that it is often ineffective in conveying headword senses to L2 users (e.g., McKeown, 1993; Nesi & Meara, 1994; Nesi & Haill, 2002; Beck, McKeown, & Kucan, 2013). In fact, feature-based definitions are prone to present L2 users with an array of sense components, which – in the absence of further contextual or usage-based information – they have to incorporate into a coherent sense representation. Therefore, as Beck, McKeown and Kucan (2013) observe, "some definitions give multiple pieces of information but offer no guidance in how they should be integrated. For example, consider the definition for *exotic*: 'foreign; strange; not native.' A learner might wonder what relationship to draw among these parts" (p. 44). Furthermore, as Nesi and colleagues suggest, L2 dictionary users often fail to derive adequate word meanings from feature-based definitions, resulting in "serious errors of interpretation, which subjects were largely unaware of" (Nesi & Haill, 2002, p. 277), or a "total misunderstanding of word meaning" (Nesi & Meara, 1994, p.

7). In particular, with feature-based definitions, L2 dictionary users are found to resort to a "kidrule" strategy, in which they pick up on the first semantic feature (or sense component) that is intelligible to them. The initial sense feature, then, is substituted for the meaning of the entire entry term, resulting in an erroneous *pars pro toto* construal of its meaning. To illustrate the "kidrule" strategy, consider an L2 user of the learner's dictionary, who wrote the (attested) sentence "I will begin a new job that is *version* [instead of *different*]", based on a definition of *version* as "a slightly different form, copy or style of an article" (Nesi & Meara, 1994, p. 9).[4]

2. Prototype Semantics and Typicality Effects

As a response to the difficulties in the conception of an effective, intelligible, and user-friendly defining style for the monolingual dictionary, a Prototype Semantics – as one of several frameworks of a broader Cognitive Semantics – might represent a promising alternative to the older, feature-based defining format (see, for an extended discussion of an emerging *Cognitive Lexicography*, Ostermann, 2015). The prototype approach to word meaning, to reiterate from current outlines in Cognitive Linguistics, holds that the sense of a lexeme can be conceived of as a central, schematic representation, the eponymous "prototype", rather than as "checklist"-type collection of semantic features (e.g., Taylor, 1990; Geeraerts, 2002). As a mental construct, the prototype is argued to contain (or assemble) the redundant information that is shared by all instances of a given lexeme, as presumably extracted from a speaker's embodied and contextualised exposure to these instantiations. (Think of an "instance" or "instantiation" as, e.g., a subordinate for a category word, or as a Langackerian "usage-event" for a given lexeme). As a schema, the prototype integrates the shared similarities between the instances of a given word into a common structure; it contains, as Verspoor and Schmitt (2013) put it, "the perceived commonality that has emerged from exposure to distinct constructions" (p. 354).[5]

As a pervasive finding in Prototype Semantics, the different instances of a lexeme are observed to exhibit a distinctive effect of "graded typicality" in relation to a lexical prototype (also called "goodness-of-exemplar" effect; Evans, 2007). In this manner, the more representative an instantiation is of the schematic core sense of the lexeme – that is, the more it corresponds to the prototype –, the higher subjects are likely to rate (or judge) its typicality (e.g., Taylor, 2009).[6] As a result, studies that ask participants to rate a series of stimuli on their typicality towards a lexeme somewhat uniformly produce scales of graded typicality; on such a scale, highly typical instances correspond to the core sense of the lexeme more closely (e.g., an *orange* to FRUIT, a *2* to *even number*, "the blanket is on the bed" to the spatial sense of *on*), while less typical instances are perceived to conform to the prototype to a lesser extent (e.g., a

melon to FRUIT, a *4264* to *even number*, or "the house stood on the lake" to the spatial sense of *on*).[7]

As a rule, the notion of the prototype, its instantiations, and typicality effects are illustrated with reference to category words (i.e., superordinate nouns), which act as prototype schema (e.g., BIRD, VEHICLE, INSTRUMENT, etc.), and their associated subordinates, which serve as instances of the prototype (e.g., *blackbird, ostrich*, etc. for BIRD). In this manner, early prototype research, conducted in the laboratory of cognitive psychologist Eleanor Rosch (e.g., Rosch, 1975, 1999 [1978]; Rosch & Mervis, 1975; Rosch, Simpsons, & Miller, 1976), showed that participants reliably distinguish between subordinates in terms of their typicality towards a given category word (or, rather, towards the prototype of the category word). While a *plum* might receive a mid-low rating as an instance of FRUIT (say, a 3 out of 7), an *orange* might be judged as a highly typical instantiation of the prototype for FRUIT (say, a 6.5 out of 7). However, beyond this widely-cited research into the word semantics of category words, prototype effects are found to occur with a broad variety of parts-of-speech other than nouns for superordinate concepts. As several studies in Cognitive Psychology and Cognitive Linguistics indicate, effects of graded typicality that arguably derive from lexical prototypes are observed with the following parts-of-speech:

— superordinate nouns, such as TOY, VEHICLE, or INSTRUMENT, in relation to subordinate nouns, such as *doll*, *teddy bear*, *water pistol*, etc. for TOY (Rosch, 1975; Rosch & Mervis, 1975);

— generic nouns, such as *woman* or *odd number*, in relation to instantiations of these concepts, such as *housewife*, *waitress*, etc. for *woman* (Armstrong, Gleitman, & Gleitman, 1999 [1983]);

— superordinate verbs, such as *(to) kill* or *(to) speak*, in relation to subordinate verbs, such as *(to) assassinate*, *(to) murder*, etc. for *(to) kill* (Pulman, 1983);

— generic verbs, such as *(to) climb*, in relation to sentences that instantiate the verb concept, such as "a monkey climbing a flagpole" or "a snail climbing up a wall" (Fillmore, 1982);

— prepositions, such as *at, on* and *in*, in relation to utterances that instantiate the preposition concept, such as "I think he's at the supermarket" or "What are you looking at me for?" for *at* (Rice, 1996);

— finally, speech act verbs, such as *(to) lie*, in relation to short event narratives that instantiate the concept of *(to) lie* (Coleman & Kay, 1981).[8]

It is against this backdrop of experimental evidence, then, that the current paper hypothesises that effects of graded typicality might equally occur with definitions for the monolingual learner's dictionary. In this manner, if definitions are conceived of as instances of a given headword sense – which

might be termed their "job description" if they seek to convey the sense of an entry term to dictionary users – a given set of definitions, for the same headword, might in fact display effects of graded typicality in relation to the entry term. There might, in other words, be a scale of graded typicality with definitions for the learner's dictionary, in which "more typical" definitions correspond to the meaning of an entry term to a greater extent, while "less typical" definitions are little characteristic of the word sense they aim to elucidate.

While the assumption, if tenable, extends the range of lexical phenomena accounted for by Prototype Semantics, the question is of interest to a "practical lexicography" in particular. From an applied perspective, a definition that displays a high degree of typicality for a headword (say, the highest in a contrasting set) ought to correspond to its meaning to a greater extent. As a result, potential "most typical" definitions for a given entry term ought to represent efficient vehicles to convey the meaning of the L2 term to dictionary users, since they are, as per typicality judgement, highly representative of the headword concept itself. While the supposed benefits of a suggested "prototype definition" for the dictionary would evidently need to be explored in a separate empirical setting,[9] the current paper, as a first step towards the construction of prototype-based definitions for the dictionary, conducted a typicality judgement study to examine whether the anticipated effect of graded typicality in fact occurred.

3. A Typicality Judgement Study

3.1. Research Hypotheses

Based on experimental evidence from Prototype Semantics that attests to typicality effects in a broad range of lexical stimuli, the judgement study investigated two research hypotheses that relate to the assumption of a graded typicality in definitions for the monolingual learner's dictionary. As hypothesis H1, the study posited that dictionary definitions ought to exhibit effects of graded typicality in relation to an entry term they define. It stated, as an alternative hypothesis (or "hypothesis of difference"):

H1 In a set of different dictionary definitions, all of which define the same entry word, definitions are judged as different in typicality towards the same entry word.

In contrast, for the H0 (or "hypothesis of non-difference"), it was assumed that dictionary users might not perceive a set of definitions as different in typicality. This might occur if respondents perceived the definition stimuli as either "equal" in typicality, which should result in rating scores that are not

significantly different between definitions; or else, if subjects perceived the definitions to be "unmarked" with regard to typicality (say, if typicality was not a psychologically relevant attribute that applied to definitions at all), subjects ought to resort to random scoring, which would result in flat score distributions for all definitions, and an equally non-significant difference between stimuli.

As a second research hypothesis, if the expected difference in typicality was in fact significant within a set of different definitions for the same entry word, H2 stated, as a corollary of H1:

H2 In a set of different dictionary definitions, all of which define the same entry word, there are "most typical" definitions that exhibit a maximum in typicality for a given set.

As a result, while H1 pursued a theory-inclined agenda and aimed to complement previous experimental research on typicality effects in Prototype Semantics, H2 was more directly oriented towards an "applied" lexicographical agenda. In this manner, as pointed out, if "most typical" definitions could be obtained through the procedure, such "ideal" instantiations ought to correspond to a given word sense to a greater extent, and might therefore represent effective vehicles to convey the meaning of dictionary headwords to L2 users of the dictionary. To put the above hypotheses to the test, the following paragraphs present the method, outcome, and discussion of a typicality judgement study. As stated in H1 and H2, the study identified whether informants in fact distinguished between definition stimuli based on their typicality towards the headword, and whether "most typical" definitions could be obtained through the procedure.

3.2. Method

Participants

Participants in the typicality judgment study were thirty-four L1 German speakers that were acquired through the network of the researcher (20 female, 14 male, age M = 29.32, SD = 9.30, range = 20–62). All informants had completed secondary education, were L1 speakers of German, and at some distance beyond the critical period for L1 acquisition (see, for the "critical period hypothesis", e.g., Lenneberg, 1967; Muñoz, 2013). Therefore, all participants could be assumed to have a conventional understanding of the syntactic, morpho-semantic and lexico-semantic norms of German. Informants took part in the study voluntarily, gave consent to the use of data, and were not compensated for their participation.

Material: Definition Stimuli

To ensure that the comparison of definition stimuli between the entry terms

was standardised, a set of novel definitions was devised for the purpose of the study. In agreement with a usage-based commitment for CL (Evans, 2007), the definition stimuli were rendered as full sentences, in which the entry term was integrated into the defining sentence, so that "the definition should as far as possible resemble ordinary speech" (Svensén, 2009, p. 235).

The entry words for the typicality judgement study, whose definition stimuli were to be rated, were taken from the inventory of the Vocabulary Size Test / VST (Nation & Beglar, 2007; Beglar, 2010). The VST was used as source of entry terms, rather than a random draw from BNC frequency lists, so as to be able to use the VST to assess the efficiency of "prototype definitions" in a follow-up study (cf. footnote 9). Headwords were selected to be representative of the major parts-of-speech, i.e., noun, verb, and adjective, in an attempt to broaden the generalisability of the study. The twelve entry terms for the study sampled in this manner were: *awe*, *candid*, *crowbar*, *devious*, *erratic*, *haze*, *marrow*, *nozzle*, *pallor*, *ubiquitous*, *(to) veer*, and *whim*.

To create definition stimuli for each headword, the British National Corpus was searched for the most frequent collocates that co-occurred with each entry term (BNC, 2007). High-frequency collocates, as the kind of "lexical material" that is habitually used in conjunction with a given word, were thought to represent a viable basis for the creation of a set of plausible, but different definition stimuli for the judgement study (see, for the use of corpora as primary resource in lexicography, e.g., Cermák, 2003; Kilgarriff, 2015). For each entry word, the 30 most frequent collocates that were located within a range of ±2 items from the headword, were extracted from the BNC, and compiled into a table (Table 1).

***Table 1*. Example of BNC Collocates Used to Create Definition Stimuli**

marrow
bone (175), transplant (20), donor (11), trust (11), bones (9), fat (7), aspirate (6), spinal (6), biopsy (6), donors (5), unit (5), transplants (4), register (4), malignancy (3), spleen (3), transplantation (3), clinic (3), involvement (3), specialist (3), cells (3), becomes (3), showed (3), autologous (2), life-saving (2), donate (2), extracting (2), imaging (2), peas (2), ginger (2)
candid
recording (10), shots (4), remarkably (3), portrait (3), interview (3), assessment (3), discussion (3), brutally (2), extraordinarily (2), confession (2), unusually (2), gaze (2), moments (2), photographs (2), friendly (2), struck (2), expression (2), statement (2), makes (2), quite (2)

Note. Numbers indicate the BNC collocate frequency *f*. The search was capped at the 30 most frequent items, the threshold of inclusion was $f > 1$.

Based on the BNC collocate material, 36 full-sentence definition stimuli were created for the 12 entry words, so that three definitions related to the same entry term in a definition triplet (A, B, and C). For each headword, the stimuli were built using the collocate material in an order of decreasing BNC frequency (A > B > C), in order to ensure that the definition material was slightly different for each definition. The 12 definition triplets served as contrasting sets of definitions that referred to the same entry term, as required for the assessment of H1 and H2. To illustrate, consider the definition stimuli for the entry term *marrow*, as presented to the participants in the study (translated from German): "marrow is a component of the human bone" (A), "marrow has to be transplanted, as a last resort" (B), "marrow donors are constantly looked for" (C) (For the remaining definition stimuli, see the appendix).

Procedure

The typicality judgement study was implemented using *LimeSurvey* (LimeSurvey Project Team & Schmitz, 2015), a computer-based survey tool that presented instructions, definition stimuli and rating scales in a sequence of individual screens. For the Likert-type scales and rating procedure, the judgement study employed the rating method as exemplified by Rosch (1975) and Barsalou (1983). In the original typicality study, Rosch had instructed her participants to rate the typicality of subordinate nouns (e.g., *ball, doll, teddy bear*, etc.) in relation to a corresponding category word (e.g., TOY). In Rosch (1975), the original instructions read as follows:

> You are to rate how good an example of the category each member is on a 7-point scale. A 1 means that you feel the member is a very good example of your idea of what the category is. A 7 means you feel the member fits very poorly with your idea or image of the category (or is not a member at all). A 4 means you feel the member fits moderately well. For example, one of the members of the category *fruit* is *apple*. If *apple* fit well your idea or image of *fruit*, you would put a 1 after it; if *apple* fit your idea of *fruit* very poorly you would put a 7 after it; a 4 would indicate moderate fit. Use the other numbers of the 7-point scale to indicate intermediate judgments (p. 198; original emphasis).

In the same vein, for the present judgement study, a first screen introduced subjects to the rating task. In line with the nature of the stimuli, the study asked informants to judge how "fitting" or "typical" a given entry term was for a definition stimulus (A, B, and C). As in Barsalou (1983), the original coding was inverted to have low typicality correspond to a 1, while high typicality corresponded to a 7. Accordingly, the judgement instructions for the study read as follows (translated from German):

There are sentences in which a word is often used. There, the word sounds

particularly fitting, typical, or apt.

For instance, many people think that the word *gallant* is used in a typical and fitting manner in "He opened the door for his companion with a gallant gesture" – possibly more typically than in "The sonata clearly shows the gallant style of the time after 1730".

In this study, you are asked to judge how typical or fitting a word sounds in a given sentence. To this end, you will see a number of sentences with an underlined word (as above). It is your task to judge how typical or fitting the word sounds in the sentence.

Please indicate your judgment on a scale of *1* to *7*. A *1* stands for a "very untypical use", a *7* for a "very typical use" of the word. Use the numbers in between to indicate intermediate judgements.

There are no right or wrong answers. Your judgement should reflect what you think is right. Your first impression is usually the best.

After the instructions, the survey presented subjects with 12 consecutive screens, each of which showed one entry term (e.g., *marrow*, *candid*, etc.) and a corresponding definition triplet A, B, and C. For each definition A, B, and C, subjects were instructed to judge how "typical" or "fitting" the entry term was perceived to be, using an anchored 7-point Likert scale, which ranged from "very untypical use" (1) to "very typical use" (7). To neutralise a possible effect of stimulus sequence on the rating, the order of stimuli in each triplet was rotated (i.e., the order might have been A, B, C / A, C, B / B, A, C / etc. at random). The more useful rotation of the entry words themselves, complete with a corresponding triplet, was envisaged, but could not be implemented. Owing to the rigor of the survey tool, there were no missing responses.

3.3. Results

General Results

An initial descriptive analysis of judgement results, at the outset, provided a number of informative insights into the adequacy of the judgement task, rating instrument, and definition stimuli. On the whole, the set of definition stimuli used in the judgement study were perceived as rather typical of their respective entry terms. As is evident from the scatterplot (Fig. 1) and boxplot display (Fig. 2) below, in which mean typicality scores for all 36 definitions – individual (Fig. 1) and overall (Fig. 2) – are depicted, the typicality ratings were, in their majority, situated in the upper half of the 7-point scale. In fact, as the overall typicality mean for all definitions indicates (Fig. 2), the average typicality score across the 36 definitions was visibly shifted towards the upper limit of the 7-point scale (M = 5.04, SD = 0.95), 95% CI [4.73, 5.35].

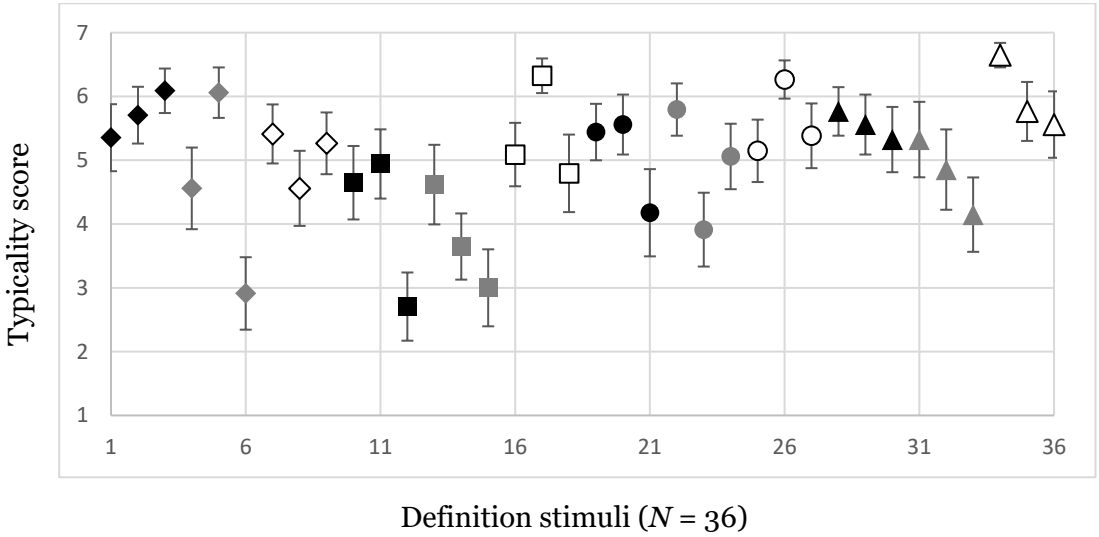

***Figure 1*. Scatterplot of mean typicality for individual definition stimuli (=dots) with 95% confidence intervals (=black bars). Each colour-shape dot type relates to a different entry word.**

Despite the more persistent trend towards the upper limit of the scale, the rating instrument did not appear to exhibit a ceiling effect that might account for the observed clustering (for instance, if inadequate instructions or scale anchors had prompted subjects to use top ratings only). This is apparent from the fact that subjects rated several definitions as mid-low or low in typicality, and from the boxplot display of typicality means across all 36 stimuli (Fig. 2). As is evident, the boxplot neither touches the scale ceiling, nor exhibits a marked compression towards the upper limit of the scale (consider skewness [0.82] and kurtosis [3.29], which are inconspicuous). As a result, the overall high typicality of definition stimuli is arguably not a function of the rating instrument, but can be taken to result from the stimuli themselves.

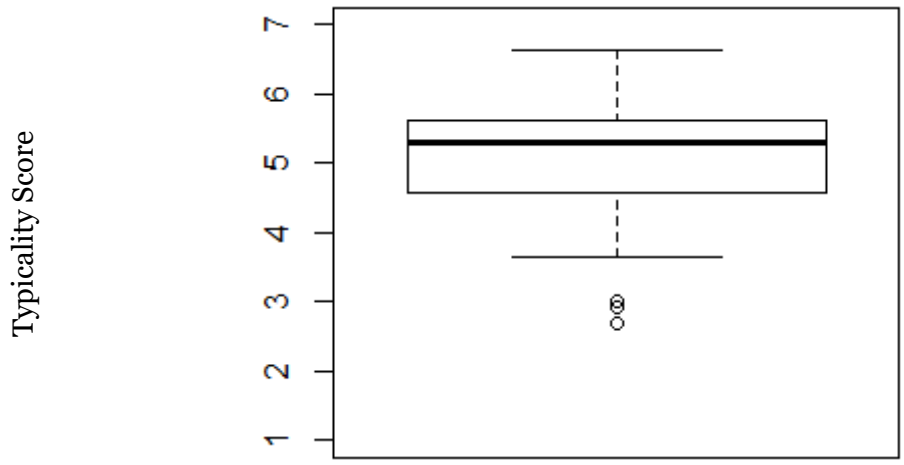

***Figure 2*. Boxplot of mean typicality for definition stimuli (overall); the black bar indicates the median, the box the upper and lower quartiles, dots are outliers.**

On the level of content, regarding the ecological validity of the definitions used for the judgement study, the overall high typicality scores appear to indicate that definitions had in fact corresponded to their respective entry words to a greater extent. Therefore, it could be argued that the majority of definition stimuli had functioned – or, had been perceived to function – as genuine monolingual definitions, which introduced or conveyed the headword concept to the study subjects. A small number of stimuli, however, had produced a more visible mismatch with the headword they were supposed to represent, as evident from their low typicality scores (consider the outliers in Fig. 2). As a result, we suspected that these stimuli had been inappropriate as monolingual definitions, and that they were implausible to occur in an actual dictionary. Therefore, to avoid a comparison of "apples-and-oranges", in which appropriate MLD definitions would be compared to more implausible (mis)definitions of the same entry word, stimuli that had received a score of more than two standard deviations below the overall typicality mean were omitted from further analysis (i.e., stimulus 6, 12 and 15 in Fig. 1; or, the outliers in Fig. 2).

Finally, further inferential analyses of the subject data and individual scoring behaviour shed light on minor, but insightful aspects of the judgement procedure. In this manner, two one-way ANOVAs were conducted using gender (male / female) and age (20–29 / 30–62) as factor and subject mean rating score as dependent variable, in order to detect a possible influence of gender or age on the typicality judgement of participants. The analyses revealed that there was no significant effect for gender [$F(1, 32) = 1.503$, $p = 0.229$] or age group [$F(1, 32) = 0.404$, $p = 0.529$] on the general rating behaviour. These results offer some provisional evidence that gender and age do not appear to influence lexical typicality judgements in the L1, and that judgement outcomes might be generalised across both genders and post-critical-period age groups.

Hypothesis H1 – Effects of Graded Typicality in MLD Definitions

Next, to identify whether respondents had in fact judged the typicality of definition stimuli to differ within a given set of definitions (i.e., to identify whether definitions in fact exhibited effects of graded typicality), the results for each definition triplet were subjected to further analyses. First, typicality ratings for each of the remaining 33 definition stimuli were tested for statistical normality using Shapiro-Wilks, which indicated that the distributions for all stimuli were non-parametric. Second, Levene's test for equality of variances was conducted for each definition set, which showed a number of triplets to exhibit unequal variances between stimuli. As a result, Kruskal–Wallis was adopted as a non-parametric alternative to the more conventional ANOVA, seeing that it tolerated violations of both normality and homogeneity of variance.

Kruskal–Wallis tests were calculated for all definition triplets, using the stimulus as a factor (A, B, C; or A, B in some cases) and typicality rating scores as the dependent variable. The following table displays Kruskal–Wallis results for each of the twelve triplets or sets (Table 2). To summarise results, there was a significant effect of definition stimulus on typicality scores in seven out of twelve cases at the significance level ($p < .05$), while the set with *devious* was close to significance. In three sets of definitions, the definition stimuli were not found to be distinct with regard to typicality. There is robust statistical evidence, therefore, that subjects perceived the different definition stimuli to be graded in typicality in seven cases, weak evidence in one case, and no significant evidence in three cases out of twelve definition sets in total (see discussion).

***Table 2*. Kruskal–Wallis Results for Graded Typicality in Definition Stimuli**

Headword		Definition stimulus			Kruskal–Wallis test
		A	B	C	
marrow	*M*	5.35	5.71	6.09	$H(2) = 3.614, p = 0.164$
	SD	1.63	1.38	1.08	
ubiquitous	*M*	4.59	6.06	–	$H(1) = 12.224, p < .001^{***}$
	SD	1.99	1.23	–	
devious	*M*	5.41	4.56	5.26	$H(2) = 4.217, p = 0.121^{+}$
	SD	1.44	1.83	1.50	
haze	*M*	4.65	4.94	–	$H(1) = 0.487, p = 0.485$
	SD	1.79	1.69	–	
(to) veer	*M*	4.62	3.65	–	$H(2) = 4.743, p < .05^{*}$
	SD	1.94	1.61	–	
erratic	*M*	5.09	6.32	4.79	$H(2) = 16.979, p < .001^{***}$
	SD	1.54	0.84	1.89	
candid	*M*	5.44	5.56	4.18	$H(2) = 9.337, p < .01^{**}$
	SD	1.37	1.46	2.12	
nozzle	*M*	5.79	3.91	5.06	$H(2) = 19.535, p < .001^{***}$
	SD	1.27	1.80	1.59	
awe	*M*	5.15	6.26	5.38	$H(2) = 11.784, p < .01^{**}$
	SD	1.52	.93	1.58	

whim	*M*	5.76	5.56	5.32	$H(2) = 1.030, p = 0.597$
	SD	1.18	1.46	1.59	
pallor	*M*	5.32	4.85	4.15	$H(2) = 7.812, p < .05^{*}$
	SD	1.84	1.96	1.81	
crowbar	*M*	6.65	5.76	5.56	$H(2) = 12.001, p < .01^{**}$
	SD	0.60	1.44	1.62	

Hypothesis H2 – Obtaining "most typical" or "ideal" definitions

After a significant effect of graded typicality was identified in seven definition sets, a post-hoc analysis was conducted to distinguish top rated – or "most typical" – definition stimuli within each set. As a non-parametric follow-up analysis to Kruskal–Wallis, a series of Mann–Whitney *U* tests was conducted for each set, applying Bonferroni correction in the case of multiple comparisons. In the definition sets, Mann–Whitney *U* permitted the identification of stimuli that were significantly different from definitions in the same set, that is, it established discreteness between definitions. On the basis of a significant difference between adjacent stimuli, a numerical comparison of stimulus scores identified those definitions (a) that were significantly different from next-lower stimuli, and (b) had received the highest typicality score in a given a set, yielding the "most typical" or "ideal" definitions in each set (Table 3). In one case, the top-rated stimulus was in fact non-different from the second-best definition, resulting in two "most typical" definitions for *candid*. In the case of *pallor*, the second-best definition was in fact non-different from both the top and third-best stimulus, which is why, in terms of statistical analysis, its status as a "most typical" definition needs to remain ambiguous. Table 3 displays Mann–Whitney *U* results for discreteness between definition stimuli and the score-based selection of "most typical" or "ideal" definitions for the judgement study (items in bold print).

***Table 3*. Mann–Whitney *U* Results for Typicality Differences between Stimuli**

Headword		"Most typical" definition stimulus			Mann–Whitney *U* test
		A	B	C	
ubiquitous	*M*	4.59	**6.06**	–	$p < 0.001$*** (A/B)
(to) veer	*M*	**4.62**	3.65	–	$p < 0.05$* (A/B)
erratic	*M*	5.09	**6.32**	4.79	$p < 0.001$*** (A/B), $p < 0.01$** (B/C)
candid	*M*	**5.44**	**5.56**	4.18	$p < 0.05$* (A/C), $p < 0.05$* (B/C)
nozzle	*M*	**5.79**	3.91	5.06	$p < 0.001$*** (A/B), $p < 0.05$* (B/C)
awe	*M*	5.15	**6.26**	5.38	$p < 0.05$* (A/B), $p < 0.05$* (B/C)
pallor	*M*	**5.32**	4.85	4.15	$p < 0.05$* (A/C)
crowbar	*M*	**6.65**	5.76	5.56	$p < 0.05$* (A/B), $p < 0.01$** (A/C)

Note. Items in bold print are top-rated stimuli in each set that were significantly different from next-lower definitions, as established per Mann–Whitney *U* test.

3.4. Discussion

As regards the assumption of graded typicality in definitions for the monolingual learner's dictionary (H1), Kruskal–Wallis testing appears, at first sight, to have produced a somewhat controversial outcome (Table 2). While the majority of definition stimuli exhibited graded typicality at the significance level – that is, the gradation of stimulus typicality could be demonstrated to exist in the population for these definition triplets –, the outcome in a number of sets appears to refute the previous hypothesis, as the anticipated difference could not be demonstrated to exist.

The seeming contradiction, however, can be resolved in less ambiguous terms if basic principles of null-hypothesis significance testing are called back to mind. As is well known, statistical analyses cannot demonstrate a null hypothesis to be correct, but merely reject – or fail to reject – the null assumption for a given set of data (e.g., Bortz & Schuster, 2010, chapter 1). In the present case, Kruskal–Wallis testing therefore in fact rejected the null assumption in 7 out of 12 cases – i.e., it demonstrated typicality grading to occur in these triplets –, while it failed to reject the null hypothesis for the remaining

5 cases – i.e., it did not demonstrate the effect to occur in these definition sets (as opposed to "it demonstrated the effect not to occur").

It is conceivable, therefore, that the adopted study setup was in fact unable to identify the anticipated effect in the remaining triplets due to more evident limitations of the study design. In this manner, the fact that numerous stimuli had been judged as highly typical of their respective headwords resulted, in the non-different triplets, in definition stimuli that were located very close to each other (compare, for *whim*, M_A = 5.76, M_B = 5.56, and M_C = 5.32; or Fig. 1). In statistical terms, however, the identification of small differences between stimuli (i.e., of small effect sizes) requires the use of an ever larger number of participants (e.g., Field, 2009, chapter 2). However, given the small sample size of 34 subjects, it is probable that the study setting was in fact unable to detect statistically significant typicality effects in definition sets whose differences between stimuli were in fact very small. To address this issue in future study designs, investigations might rely on a larger *a priori* sample, or – in an ideal case – calculate the required sample size minimum from the stimulus data obtained in the present judgement study (Field, 2009, chapter 2).[10]

Secondly, as regards the question of whether "most typical" definitions can be obtained through the adopted procedure (H2), Mann–Whitney *U* analyses of the judgement data demonstrated that such "ideal" definitions could be identified, or singled out. In this manner, for all sets of definitions that exhibited a difference in typicality, discrete, top-rated stimuli could be detected that corresponded to their respective headword concepts to a superior extent (Table 3).

While the study was therefore successful at producing a series of "prototype definitions", a number of intriguing follow-up concerns emerge from the current findings that might merit the attention of future investigations. In this manner, future studies might address the question of which linguistic (or extra-linguistic) criteria in fact increased the typicality of definitions for the monolingual dictionary. As is apparent from Table 3, it appears that definition stimuli had usually fared best that had been built from highly frequent BNC collocates (i.e., definition A or B). This suggests that *collocate frequency* might represent a vital means to increase the typicality of definitions for the monolingual dictionary, which in fact closely corresponds to the current defining practice in a corpus-based lexicography of the COBUILD tradition. Based on the knowledge of such typicality-enhancing criteria, an iterative process of (re)building definitions from previous "ideal" or "most typical" stimuli might be envisioned, in which a cycle of iterations ought to produce definitions whose typicality cannot easily be surpassed.

At the same time, as suggested above, an "applied" lexicographical perspective on the dictionary more evidently calls for a separate empirical assessment of the efficiency of the novel "prototype definition" for the learner's dictionary.

Such a follow-up study, then, ought to determine whether – in the long run – a Prototype Semantics approach to dictionary definitions might in fact enhance the capacity of the dictionary to support non-native, foreign-language users.

Acknowledgements

I would like to thank the two anonymous reviewers for their helpful comments on an earlier draft, which has helped to greatly improve the article.

References

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1999). What some concepts might not be [1983]. In E. Margolis & S. Laurence (Eds.), *A Bradford book. Concepts. Core readings* (pp. 225–259). Cambridge, Mass.: MIT Press.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, *11*(3), 211–227.

Beck, I. L., McKeown, M. G., & Kucan, L. (2013). *Bringing words to life: Robust vocabulary instruction* (2nd ed.). New York: The Guilford Press.

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing*, *27*(1), 101–118.

Béjoint, H. (2010). The lexicography of English: From origins to present. Oxford: Oxford Univ. Press.

Bensoussan, M., Sim, D., & Weiss, R. (1984). The effect of dictionary usage on EFL test performance compared with student and teacher attitudes and expectations. *Reading in a Foreign Language*, *2*(2), 262–276.

BNC Consortium. (2007). The British National Corpus, version 3 (BNC XML Edition). Retrieved from http://www.natcorp.ox.ac.uk/

Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7th ed.). Berlin, Heidelberg: Springer.

Cermák, F. (2003). Source materials for dictionaries. In Sterkenburg, P. G. J. van (Ed.), *Terminology and lexicography research and practice: Vol. 6. A practical guide to lexicography* (pp. 18–25). Amsterdam, Philadelphia: John Benjamins Pub.

Coleman, L., & Kay, P. (1981). Prototype semantics: The English word *lie*. *Language*, *57*(1), 26–44.

Coseriu, E. (2000). Structural semantics and 'cognitive' semantics. *Logos and*

Language: Journal of General Linguistics and Language Theory, *1*(1), 19–42.

Evans, V. (2007). A glossary of cognitive linguistics. Glossaries in linguistics. Edinburgh: Edinburgh Univ. Press.

Field, A. (2009). *Discovering Statistics Using SPSS* (2nd ed.). Los Angeles, Calif.: Sage.

Fillmore, C. J. (1982). Towards a desriptive framework for spatial deixis. In R. J. Jarvella & W. Klein (Eds.), *Speech, place, and action. Studies in deixis and related topics* (pp. 31–59). Chichester: Wiley.

Geeraerts, D. (2002). Conceptual approaches III: Prototype theory. In D. A. Cruse (Ed.), *Handbücher zur Sprach- und Kommunikations-wissenschaft: Vol. 21. Lexikologie* (pp. 284–291). Berlin: De Gruyter Mouton.

Geeraerts, D. (2007). Lexicography. In D. Geeraerts & H. Cuyckens (Eds.), *The Oxford handbook of cognitive linguistics* (pp. 1160–1174). Oxford: Oxford Univ. Press.

Geeraerts, D. (2010). *Theories of lexical semantics*. *Oxford linguistics*. Oxford: Oxford Univ. Press.

Hornby, A. S., Deuter, M., & Hey, L. (Eds.). (2015). *Oxford Advanced Learner's Dictionary of Current English* (9th ed.). Oxford: Oxford Univ. Press.

Kilgarriff, A. (2015). Using corpora as data sources for dictionaries. In H. Jackson (Ed.), *The Bloomsbury companion to lexicography* (pp. 77–96). London: Bloomsbury.

Knight, S. (1994). Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities. *The Modern Language Journal*, *78*(3), 285–299.

Lenneberg, E. H. (1967). *Biological foundations of language*. New York: Wiley.

LimeSurvey Project Team, & Schmitz, C. (2015). *LimeSurvey*. Hamburg.

McCreary, D. R., & Dolezal, F. T. (1999). A study of dictionary use by ESL students in an American university. *International Journal of Lexicography*, *12*(2), 107–146.

McKeown, M. G. (1993). Creating effective definitions for young word learners. *Reading Research Quarterly*, *28*(1), 16–31.

Muñoz, C. (2013). Age effects in SLA. In P. J. Robinson (Ed.), *The Routledge encyclopedia of second language acquisition* (pp. 12–16). New York, NY: Routledge.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13.

Nesi, H., & Meara, P. (1994). Patterns of misinterpretation in the productive use of EFL dictionary definitions. *System*, *22*(1), 1–15.

Nesi, H., & Meara, P. (1991). How using dictionaries affects performance in multiple-choice EFL tests. *Reading in a Foreign Language*, *8*(1), 631–643.

Nesi, H. (2000). The use and abuse of EFL dictionaries: How learners of English as a foreign language read and interpret dictionary entries. Lexicographica. Series Maior: Vol. 98. Berlin: de Gruyter.

Nesi, H. (2014). Research timeline: Dictionary use by English language learners. *Language Teaching*, *47*(1), 38–55.

Nesi, H., & Haill, R. (2002). A study of dictionary use by international students at a British university. *International Journal of Lexicography*, *15*(4), 277–305.

Ostermann, C. (2015). *Cognitive lexicography*. Lexicographica. Series Maior: Vol. 149. Berlin: de Gruyter.

Pulman, S. G. (1983). *Word meaning and belief*. Croom Helm Linguistics Series. London: Croom Helm.

Rice, S. (1996). Prepositional prototypes. In M. Pütz & R. Dirven (Eds.), *Cognitive linguistics research: Vol. 8. The construal of space in language and thought* (pp. 135–163). Berlin: de Gruyter.

Ronald, J. (2002). L2 lexical growth through extensive reading and dictionary use: A case study. In A. Braasch (Ed.), *Proceedings of the tenth EURALEX international congress, EURALEX 2002. Copenhagen, Denmark, August 13 - 17, 2002* (pp. 765–771). Copenhagen: Center for Sprogteknologi.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192–233.

Rosch, E. (1999). Chapter 8: Principles of categorization [1978]. In E. Margolis & S. Laurence (Eds.), *A Bradford book. Concepts. Core readings* (pp. 189–206). Cambridge, Mass.: MIT Press.

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

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(4), 491–502.

Sinclair, J. (Ed.). (2015). Collins Cobuild Advanced Learner's Dictionary: The source of authentic English (8th ed.). München: Langenscheidt.

Svensén, B. (2009). A handbook of lexicography: The theory and practice of dictionary-making. Cambridge: Cambridge Univ. Press.

Swanepol, P. (2003). Dictionary typologies: A pragmatic approach. In Sterkenburg, P. G. J. van (Ed.), *Terminology and lexicography research and practice: Vol. 6. A practical guide to lexicography* (pp. 44–69). Amsterdam, Philadelphia: John Benjamins Pub.

Szczepaniak, R. (2006). *The role of dictionary use in the comprehension of idiom variants*. Lexicographica. Series Maior: Vol. 131. Tübingen: Niemeyer.

Taylor, J. R. (1990). Schemas, prototypes, and models: In search of the unity of the sign. In S. L. Tsohatzidis (Ed.), *Meanings and prototypes. Studies in linguistic categorization* (pp. 521–534). London: Routledge.

Taylor, J. R. (2009). Prototype semantics. In K. Allan & K. Brown (Eds.), *Concise encyclopedia of semantics* (pp. 238–240). Amsterdam: Elsevier.

Tono, Y. (2001). Research on dictionary use in the context of foreign language learning: Focus on reading comprehension. Lexicographica. Series Maior: Vol. 106. Tübingen: Niemeyer.

Verspoor, M., & Schmitt, N. (2013). Language and the lexicon in SLA. In P. J. Robinson (Ed.), *The Routledge encyclopedia of second language acquisition* (pp. 353–360). New York, NY: Routledge.

Zgusta, L. (1971). *Manual of lexicography*. Prague: Academia.

Endnotes

[1] The monolingual learner's dictionary (also "EFL dictionary" or "advanced learner's dictionary") is a pedagogic L2-only dictionary that "take[s] cognisance of the linguistic competencies of non-native speakers" (Swanepol, 2003, p. 56). The monolingual dictionary is designed to assist foreign-language learners in a variety of L2 learning tasks (e.g., reading comprehension, vocabulary learning, text production, etc.); it typically uses a restricted defining lexicon, offers no diachronic information, foregrounds collocational patterns, and highlights syntagmatic sense relations of the headword (synonymy, antonymy, etc.).

[2] Consider that design issues might account for the fact that several studies identified no benefit of "monolingual" over "no-dictionary" conditions, rather than the genuine absence of L2-didactic effects through the use of the MLD. In this manner, sample sizes might be too small to identify small effect sizes in between-subject designs; low- or high-proficiency learners might not benefit

from the MLD at all; or, test instruments might not be sensitive to the specific type of L2 learning produced by the MLD.

[3] In his *Manual on Lexicography*, a standard reference work in dictionary-making, Zgusta (1971) observes that "the lexicographic definition enumerates only the most important semantic features of the defined lexical unit, which suffice to differentiate it from other units" (p. 253). To illustrate the structuralist tradition, compare the definitions for *measurement* or *frantic* in the latest, 9th edition of the *Oxford Advanced Learner's Dictionary* (Hornby, Deuter, & Hey, 2015): "the act or process of finding the size, quantity or degree of s.th." (p. 965); "done quickly and with a lot of activity, but in a way that is not very well organised" (p. 622).

[4] On the signifier side, a given headword (or entry word, entry term) in the dictionary might be mono- or multilexical (e.g., *cruel*, *teddy bear*); as a signified, it may refer to a single concept or several concepts in a polysemous sense structure (e.g., *mole* as [animal], [birthmark], or [informant]). To avoid the tiresome repetition of "*X* in the sense of X_1", the paper assumes that a given entry word refers to a single concept only.

[5] In an earlier conception of Prototype Semantics, the prototype was considered to be equivalent to the most typical, item-type instance of a given word (for discussions, see Coseriu, 2000; Taylor, 2009). In this manner, a *blackbird* or *sparrow* were conceived of as prototypes of the *bird* concept. In a current prototype theory, these exemplars are regarded as highly typical instantiations – presumably the most typical ones – of the *bird* prototype, which is a schematic representation of the sense of *bird*.

[6] In fact, various experimental measures other than subject ratings attest to the presence of typicality effects in lexical semantics, e.g., the order in which instances are learned, verification time for category membership, and probability of naming instances as output (e.g., Rosch, Simpson, & Miller, 1976; Taylor, 2009).

[7] For the above considerations, compare cognitive psychologists Coleman and Kay (1981): "Let us say, roughly, that a semantic prototype associates a word or phrase with a prelinguistic, cognitive schema or image; and that speakers are equipped with an ability to judge the degree to which an object (or, if you prefer, the internal representation thereof) matches this prototype schema or image" (p. 27).

[8] Fillmore's study (1982) is conceptual rather than experimental, but ties in with the greater research panorama.

[9] The didactic impact of a novel "prototype definition" for the learner's dictionary might be assessed through a pre-posttest-control-group design that

compares the efficiency of a "traditional definition" and "prototype definition" format. The influence of the respective definition formats, as independent variable, on, e.g., L2 reading comprehension or L2 vocabulary learning, as dependent variables, might be evaluated (see, for a study model, e.g., Knight, 1994).

[10] Of course, in a further scenario consistent with the evidence, effects of graded typicality might occur with a *specific type* of definition only, while they do not apply to *all definitions* of the monolingual learner's dictionary. Such an enforced partitioning of research phenomena, however, appears to violate the imperative of scientific parsimony.

Appendix

awe		
A	B	C
Sie blickte voll Ehrfurcht zu dem mächtigen Berg hinauf.	Sie blickte voll Ehrfurcht zu der Gottesfigur über dem Altar hinauf.	Sie blickte voll Ehrfurcht in die Weiten des Alls hinauf.
candid		
A	B	C
Sie gab erstaunlich offenherzige Antworten	Sie sprach in dem TV-Portrait erstaunlich	Sie zeigte sich auf den Kamerafotos erstaunlich

während des Interviews.	offenherzig über ihr Leben.	offenherzig.

crowbar		
A	B	C
Die Einbrecher hatten eine eiserne Brechstange mitgebracht.	Er benutzte die Brechstange als einen Hebel, um die Kiste zu öffnen.	Sie benutzte die Brechstange, um mit ihrer Hilfe die Türe zu öffnen.

devious		
A	B	C
Er benutzte eine hinterhältige List, um sie in die Irre zu führen.	Er war ein hinterhältiger Politiker.	Er verfolgte eine sorgfältig geplante, hinterhältige Taktik.

erratic		
A	B	C
Die Kurse stiegen und fielen in unberechenbarer Weise.	Das Wetter war unberechenbar und wechselte ständig von Sonnenschein zu Regen.	Er benahm sich wie ein Verrückter und änderte sein Verhalten in unberechenbarer Weise.

haze		
A	B	C
Mit der Morgenhitze kam Dunst über den Feldern auf.	Der Verkehr erzeugte eine dichte Glocke aus Dunst über der Stadt.	Seine Gedanken waren betäubt vom Alkoholdunst.

marrow		
A	B	C
Das Knochenmark ist Bestandteil des	Im schlimmsten Fall muss Knochenmark	Es werden ständig Spender für

menschlichen Knochens.	transplantiert werden.	Knochenmark gesucht.

nozzle		
A	B	C
Ein Wasserstrahl schoss aus der Düse heraus.	Die Flüssigkeit wird durch die Düse gepresst.	Das Wasser sprühte aus der Düse heraus.

pallor		
A	B	C
Auf ihrem Gesicht lag eine geisterhafte Blässe.	Auf ihrem Gesicht lag eine tödliche Blässe.	Ihr Gesicht zeigte Spuren von Blässe, Kälte und Schweiß.

ubiquitous		
A	B	C
Die Gegenwart Gottes ist für religiöse Menschen allgegenwärtig.	Die Angst vor dem Terror ist im Moment allgegenwärtig.	Der Jutebeutel ist im Moment allgegenwärtig.

(to) veer		
A	B	C
Der Wind schwenkte plötzlich um.	Sein Standpunkt in der Diskussion schwenkte plötzlich um.	Der Fahrer schwenkte das Auto plötzlich herum.

whim		
A	B	C
Er kaufte das funkelnde Feuerwehrauto aus einer kindlichen Laune heraus.	Sie fand ihn eigentlich unattraktiv, ging aber aus einer Laune heraus mit ihm aus.	Er war eine Person voller exzentrischer Launen und Neigungen.