Seth Mehl*

What we talk about when we talk about corpus frequency: The example of polysemous verbs with light and concrete senses

https://doi.org/10.1515/cllt-2017-0039

Abstract: Gilquin (2008, What you think ain't what you get: Highly polysemous verbs in mind and language. In Jean-Remi Lapaire, Guillaume Desagulier & Jean-Baptiste Guignard (eds.), *From gram to mind: Grammar as cognition*, 235–255. Bordeaux: Presse Universitaires de Bordeaux) reported that light uses of verbs (e.g. *make use*) tend to outnumber concrete uses of the same verbs (e.g. *make furniture*) in corpora, whereas concrete senses tend to outnumber light senses in responses to elicitation tests. The differences between corpus frequency and cognitive salience remain an important and much-discussed question (cf. Arppe et al. 2010, Cognitive corpus linguistics: Five points of debate on current theory and methodology. *Corpora* 5(1). 1–27). The question is particularly complicated because both *corpus frequency* and *cognitive salience* are difficult to define, and are often left undefined. Operationalising and defining corpus frequencies are the issues at the heart of the present paper, which includes a close, manual semantic analysis of nearly 6,000 instances of three polysemous verbs with light and concrete uses, *make, take*, and *give*, in the British component of the International Corpus of English. The paper compares semasiological frequencies like those measured by Gilquin (2008) to onomasiological frequency measurements (cf. Geeraerts 1997, *Diachronic prototype semantics: A contribution to historical lexicology*. Oxford: Clarendon Press). Methodologically, the paper demonstrates that these approaches address fundamentally different research questions, and offer dramatically different results. Findings indicate that corpus frequencies in speech may correlate with elicitation test results, if the corpus frequencies are measured onomasiologically rather than semasiologically; I refer to Geeraerts's (2010, *Theories of lexical semantics*. Oxford: Oxford University Press) hypothesis of *onomasiological salience* in explaining this observation.

Keywords: semantics, prototypes, onomasiology

***Corresponding author: Seth Mehl,** School of English, University of Sheffield, Jessop West, 1 Upper Hanover Street, S3 7RA Sheffield, UK, E-mail: s.mehl@sheffield.ac.uk

1 Introduction

In an influential study, Gilquin (2008) observed that light uses of verbs (e.g. *make a decision, take action*) tend to occur far more frequently in corpora than concrete senses (e.g. *make furniture, take the books*), whereas in elicitation tests, concrete senses of verbs tend to be generated by respondents far more frequently than light uses. Prior to Gilquin, Sinclair (1991: 112–113) had made similar observations, but Gilquin (2008) develops the point via experimentation, not only measuring frequencies of each sense of *take* and *give* in the Switchboard and FROWN corpora, but also performing elicitation tests with native speakers, in which respondents were asked to generate the first sentence that came to mind with the target verbs *give* and *take*. The resulting difference between observed corpus frequencies and elicitation tests is important insofar as it relates to theories of cognitive salience and prototypicality: elicitation tests are one acknowledged method of identifying the most cognitively salient meaning, or the prototypical meaning of a word, and corpus frequencies are another. The difference between corpus frequencies and elicitation tests is also important as it relates to fundamental methods of measuring corpus frequencies, a question at the heart of the present paper.

Affirming the importance of Gilquin's (2008) study, Werner and Mukherjee (2012) replicated the corpus portion of the study using selected written texts from three components of the International Corpus of English (ICE), i.e. Great Britain (ICE-GB), India (ICE-India), and Sri Lanka (ICE-SL), in order to determine whether Gilquin's (2008) conclusion about corpus frequencies could be maintained across varieties of World Englishes. Werner and Mukherjee (2012) measured corpus frequencies similarly to Gilquin (2008), and affirmed that frequencies of light *give* and *take* are higher than frequencies for concrete *give* and *take* across all data sets, but that considerable variation arises between the varieties vis-à-vis other senses of each verb. Werner and Mukherjee (2012) call for further elicitation tests in each of the three regions to develop the research questions further.

In cognitive corpus semantics, the relationship between corpus frequencies and cognitive salience remains an important and much-discussed question (cf. Gilquin 2006; Nordquist 2004; Arppe et al. 2010; Glynn 2014). The question is complicated by the fact that both *corpus frequency* and *cognitive salience* are difficult to define. Corpus linguists do not often explicitly explore definitions of *frequency* at all, and when they do, they tend to recommend normalising per million words (cf. McEnery and Wilson 2001: 83; McEnery and Gabrielatos 2006: 52–53; Lindqusit 2009: 41–42; Evison 2010: 126). Cognitive linguists have defined *cognitive salience* in many different ways (cf. Rosch 1973, 1975a, 1975b; Geeraerts 1997; Taylor 2003; Geeraerts 2006 [1989]; Gries 2006), including Geeraerts's (2010: 201) innovative and valuable notion of onomasiological salience.

Indeed, the discord between existing definitions is the motivation for Gilquin's (2008) study. One of the aims of Gilquin's (2008: 3) paper is to highlight the different ways that linguists employ the term *prototypicality*, and to underline that the term tends to be used "loosely," with linguists "not always making it clear what they have in mind." Gilquin notes "the unspoken assumption ... that the most salient exemplar in the mind is also the most frequent one in language," even though "the two criteria provide divergent results and can therefore not be seen as different indicators of the same phenomenon" (ibid: 3). One of the goals of the present paper is to highlight the different ways that linguists employ the term *corpus frequency*, and to underline that relative frequency measures should be defined and operationalised explicitly, in relation to research questions.

Both Gilquin (2008) and Werner and Mukherjee's (2012) studies measure corpus frequencies semasiologically. In corpus semantics, semasiology is an approach which begins with a word form and identifies all the meanings that can be expressed by that word form. The converse of semasiology, onomasiology, begins with a meaning and identifies all the word forms that can be used to express that meaning. The two approaches to measuring corpus frequency are fundamentally different, as I discuss in greater detail in Section 2. In the present paper, I present a new corpus study which innovates on Gilquin's (2008) and Werner and Mukherjee's (2012) work by measuring not only semasiological frequencies of light and concrete verb senses, but also onomasiological frequencies of those senses in relation to their respective semantic (onomasiological) alternates.

This study is occupied primarily with the nature of frequency in corpus semantics, and secondarily with how various definitions of corpus frequency might theoretically relate to issues in cognitive linguistics, specifically cognitive salience and prototypicality of multiple senses of polysemous words. The study poses two research questions, the first semasiological and the second onomasiological:

i. Are Gilquin's (2008) and Werner and Mukherjee's (2012) observations on the semasiological corpus frequencies of light uses and concrete senses corroborated by new observations of *make, take*, and *give* in speech and writing in ICE-GB?
ii. How do onomasiological measures of light and concrete senses in ICE-GB compare to Gilquin's (2008) and Werner and Mukherjee's (2012) semasiological observations?

First, I review the theoretical discussion behind the nature of frequency measurements in corpora, including semasiological and onomasiological frequencies in corpus semantics. Then, I briefly summarise key work related to cognitive salience and prototypicality, including foundational work by Rosch (1973, 1975a, 1975b), ongoing debates on the nature of salience and prototypes, and Geeraerts's (2010) hypothesis of *onomasiological salience*. I discuss a number of explanations for the previously observed discrepancy between corpus frequencies and cognitive salience. Finally, I present a semasiological and onomasiological analysis of the concrete senses and light uses of *make, take*, and *give* in ICE-GB, with a careful and transparent definition of *light use* and *concrete sense*, in order to address the research questions above. I perform a close, manual semantic analysis of nearly 6,000 instances of *make, take*, and *give*, along with thousands of instances of their onomasiological alternates. First, a semasiological analysis is performed, comparable to Gilquin (2008) and Werner and Mukherjee (2012). Senses of each verb are identified manually by reading each instance of each verb and cataloguing it as concrete or light. Frequencies of the senses of each verb are compared in speech and in writing. Then, an onomasiological analysis is performed. Onomasiological alternates are identified in a data-driven way for each sense, as described below. Closely reading of each example in context is absolutely necessary in the identification of onomasiological alternates. Frequencies of each onomasiological alternate are then compared, presented as probabilities, and tested statistically. Finally, I demonstrate that the previously observed discrepancy between cognitive salience tests and corpus frequency may be in part an epiphenomenon of an underspecified approach to corpus frequency. Specifically, I demonstrate that corpus frequencies may not differ so considerably from elicitation tests for cognitive salience – if the corpus frequencies are measured onomasiologically, in spoken language. Most importantly, I conclude that it is absolutely necessary for corpus linguists to define relative frequencies carefully and explicitly with a meaningful relationship to research questions.

2 Frequency in corpus semantics

Arppe et al. (2010: 7) succinctly note the unique advantage of corpus linguistics for addressing "questions that can be answered through the observation of (relative) frequencies of occurrence. Such data can then yield generalizations about questions of natural language use." Relative frequencies, or normalised frequencies, can be derived in many different ways, by relating a raw number of occurrences of some linguistic feature in a data set to some baseline comparator. There are many types of normalised frequencies that can be derived: measuring relative frequencies per million words is a standard that is reproduced in numerous corpus linguistics textbooks (cf. McEnery and Wilson 2001: 83; McEnery et al. 2006: 52–53; Lindqusit 2009: 41–42), as well as in *The Routledge handbook of corpus linguistics* (Evison 2010: 126). Other viable options are discussed less in the literature: frequencies can be normalised per part of speech (e.g. per thousand nouns), per phrase (e.g. per thousand noun phrases), per clause (e.g. per hundred subordinate clauses), or per morpheme (e.g. per million morphemes), among many other options. It is crucial that a normalisation procedure is not only explicitly stated, but justified methodologically in relation to a research question. Specifically, researchers should generally consider a normalisation procedure that relates to the feature being examined. For example, a study on restrictive relative clauses may benefit from normalisation related to restrictive relative clauses, measuring restrictive relative clauses per thousand relative clauses, per thousand dependent clauses, or per thousand clauses, among other options. Each of those options will differ from each other, potentially in significant ways, and each will differ from normalisation per million words. Each option addresses a slightly different research question as well.

Corpus semantics can be seen as investigating form and meaning: specifically, the relative frequency of association of form and meaning in natural language use (Glynn 2014: 14). Normalising corpus frequencies in relation to form and meaning is therefore desirable. To do so, it is helpful to categorise relationships between form and meaning as semasiological or onomasiological. Semasiological normalisation counts each instance of each sense of a word, and normalises by the total number of instances of the word. That is, semasiological research can be seen as asking the following precise research question: given that word form *a* occurs in the corpus, what is the probability that it is expressing meaning *x*, *y*, *z*, etc.? Onomasiological normalisation can be seen as the converse: given that meaning *x* is expressed in the corpus, what is the probability that it is expressed via word *a*, *b*, *c*, etc. (cf. Geeraerts 1988)? Onomasiological normalisation thus counts each instance of each word expressing a given meaning, and normalises by the total number of expressions of that meaning.

As Wallis (2012) argues, a semasiological normalisation represents an exposure rate: given that a listener or reader encounters word *a*, what is the probability that he or she is encountering meaning *x*, *y*, *z*, etc.? An onomasiological normalisation, on the other hand, represents a selection preference (ibid.): given that a speaker or writer is expressing meaning *x*, what is the probability that it is expressed via word *a, b, c*, etc.? Semasiological approaches are therefore useful for research questions related to exposure rates, while onomasiological approaches are useful for research questions related to selection preferences. For example, a semasiological normalisation is useful for dictionary design: dictionaries sometimes present a given word form and then list the sense with the highest relative semasiological frequency first (cf. Collins COBUILD 1995). A semasiological normalisation can also facilitate research questions relating to exposure rates and cognition (cf. Schmid 2007; see also Section 3 below). An onomasiological normalisation might be useful for style guides that advise which word is the standard or most common choice for a given meaning in a given context or genre. An onomasiological normalisation also relates to cognitive research, particular vis-à-vis Geeraerts's (2010) hypothesis of onomasiological salience, which I discuss in Section 4.

Examples in English abound of both semasiological corpus semantics (cf. Lee and Ziegeler 2006; Lange 2007; Hundt 2009; Fuchs 2012; Fuchs et al. 2013) and onomasiological corpus semantics (cf. Haase 1994; Schneider 1994; Balasubramanian 2009), but explicit discussion or justification of either method is uncommon.[1] For example, the two key studies examined here, by Gilquin (2008) and Werner and Mukherjee (2012), are both semasiological, but neither explicitly states its methodological approach or normalisation practices as semasiological. Although it has been argued that semasiological observations indicate exposure rates and onomasiological observations indicate selection preferences, those facts are often underexamined in research conclusions. The present study takes an important step forward by clearly distinguishing the two methods and comparing them, in relation to established and important research on elicitation test results and corpus frequencies for polysemous words.

3 Cognitive salience and elicitation tests

The concept of *prototypes* was introduced in psychology research by Rosch (1973), referring to the "clearest cases [of category members], best examples ... [which] serve as reference points in relation to which other category members are judged" (Rosch 1975a: 544–545). *Prototypicality* can be described as the cognitive organisational principle by which dissimilar examples can be deemed

[1] Hundt (2009), importantly, identifies her study as semasiological, and notes that the research would have been more robust if it had been conducted onomasiologically instead.

members of a single category. Rosch (1973) demonstrated that concrete examples of a category exhibited prototypicality: for example, a particular shade of red was prototypical for a given language community, insofar as it was deemed the best example of *red*, against which other, different shades were judged. Similarly, particular species of birds could be seen as prototypical birds by a given language community, such that one species is the best example against which other less good examples are judged, and so on. Rosch (1973, 1975a, 1975b) identified prototypes experimentally using various methods. For example, subjects' first or fastest responses to elicitation tests were seen as indicative of a prototype: when asked to name a bird, the first bird that comes to mind is the prototypical member of the category. Alternatively, subjects' intuitive sense of the best example of a category could be seen as the prototype. In still other instances, the use of a reference example was seen as indicative of prototypes: subjects describe a non-prototypical example in relation to a prototypical example as a reference.

In cognitive linguistics, prototypicality has become a standard for describing semantic fuzziness, both within semantic categories (such as RED or BIRD) and within polysemous words (such as with meanings of *take* or *give*). In applying the notion of prototypicality to polysemous words, the word itself is viewed as a category, and the different related senses are the members of the category (cf. Geeraerts 2006 [1989]). It is this definition of prototypicality that interests Gilquin (2008) and Werner and Mukherjee (2012).

Recently, an alternative means of identifying prototypicality has arisen. Researchers have begun to identify prototypes as the most frequently occurring example of a category in corpus data (cf. Gries 2006; Gilquin 2006; Geeraerts 2006 [1989]; Gilquin 2008; Heylen et al. 2008; Arppe et al. 2010). According to Geeraerts (1988: 222), corpus frequency can be a "heuristic tool in the pinpointing of prototypes." But exactly which relative frequency is thought to represent the prototype? Is it frequency per million words, per thousand nouns, per hundred phrases or clauses, or per million morphemes? Or is it semasiological or onomasiological frequency? Geeraerts (1997) employs both semasiological and onomasiological frequencies as indicators of prototypicality, and discusses the relationships between the two measures. Taylor (2003: 54) explicitly asserts that it is semasiological frequency that represents prototypicality. Schmid (2007: 119–120) explains the theoretical mechanism for this proposed correlation, based on the acknowledgement that semasiological frequencies represent exposure rates: according to Schmid, frequency of exposure to a word sense results in the routinisation or entrenchment of the cognitive activation of that sense. Further, "deeply entrenched cognitive units are more likely to become cognitively salient than less well entrenched ones” (ibid. 119–120). Thus, a high exposure rate, as indicated by semasiological frequency, ought to result in high cognitive salience or prototypicality. Taylor (2012: 148) summarises: “[semasiological] frequency influences performance on all manner of experimental tasks” related to the psycholinguistics of language production and reception. General corroboration of this observation can be found in an array of psycholinguistic literature (cf. Bybee and Hopper 2001; Gries and Divjak 2012; Divjak and Gries 2012). Alternatively, Geeraerts (2010: 201) has proposed *onomasiological salience* as a measure of prototypicality, defined as the preference for a given word form over its semantic alternates. These selection preferences can be measured as onomasiological frequencies: the relative preference for one form over a semantic alternate for expressing a given meaning, across a population, is an indicator of cognitive salience or prototypicality. Geeraerts (ibid.) has asserted that onomasiological salience “can be equated with the notion of “entrenchment.” The difference between Geeraerts’s (2010) assertion and Schmid’s (2007) is huge, but it can be addressed empirically: which corpus frequency (semasiological or onomasiological) correlates with entrenchment, salience, or prototypicality? Moreover, which proposed measures of prototypicality (including elicitation tests, intuition, use of reference examples, semasiological frequency, and onomasiological frequency) correlate with each other? The second question is a much larger one than can be addressed in the present work. Instead, the present study builds on Gilquin (2008), to compare semasiological and onomasiological frequencies to previously published findings from elicitation tests. Gilquin (2008) affirmed that semasiological frequencies, as per Schmid’s (2007) claim, do not correlate with elicitation test results for salience. The findings of the present study further corroborate the lack of correlation between semasiological frequencies and elicitation test results. However, the present findings demonstrate that onomasiological frequencies in speech tend to correlate with elicitation tests for salience much more closely than semasiological frequencies.

4 Corpus frequency and cognitive salience: The case of verbs with concrete senses and light uses

Gilquin (2008) concludes that in corpus data, light uses of verbs are more frequent than concrete senses of the same verbs, whereas concrete senses are generated more frequently than light uses in elicitation tests for cognitive salience. Gilquin (2008) defines 15 senses for *give* and 18 senses for *take*, derived via consultation with five learners' dictionaries. She identifies multiple concrete senses, in which the direct object of the verb has a concrete referent. She also identifies a light use, in which the verb and direct object represent an action whose semantic content is expressed primarily via the direct object; for example, *take action* is equivalent to *act* (v.) (cf. Jespersen 1954; see below for further discussion of light verbs). Participants in an elicitation test were asked to generate a sentence using example words (including *give* and *take*); the first sentence generated by each participant using *give* and *take* was then manually analysed and categorised into one of the defined senses. Concrete senses dominate this data; i.e. concrete senses are far more common than other senses. Gilquin then extracted 500 instances of *take* and 500 instances of *give* from each of two corpora, the FROWN corpus of written American English and the Switchboard corpus of spoken American English, and found that light uses dominate this data; i.e. light uses are more common than other senses.

Gilquin's (2008) findings can be further analysed in multiple ways. For example, it is straightforward to hypothesise that light uses are more frequent than concrete senses in use because concrete senses are more pragmatically restricted: a concrete sense of *take* or *give* is restricted to a narrow range of real-world contexts, in which a concrete object is being transferred. Light uses such as *take action, take a decision*, or *give support* can be employed in an extremely wide array of real-world contexts, not limited to discussion of transferring concrete things. It is also straightforward to hypothesise, based on fundamental principles of cognitive linguistics, that concrete senses will be most salient because of embodied experience: because our experience of the world is first and foremost concrete, via embodied sensory experience, concrete senses of verbs may be primary in the mind (cf. Lakoff and Johnson 1980; Johnson 2007).

However, Gilquin's (2008) findings can also be analysed methodologically. How might we measure salience differently such that it might correlate with corpus frequency? It is certainly conceivable that another established method for measuring salience might correlate with Gilquin's frequency measurements: perhaps speaker intuition or the use of reference examples. Or, crucially, it might be that onomasiological elicitation tests would correlate with corpus frequencies: for example, an elicitation test might ask a participant to fill in the blank in a sentence, such that concrete *take* would be one possible response, as would semantic alternates of concrete *take*. Alternatively, we can reflect on Gilquin's study by asking how we might measure corpus frequency differently such that it might correlate with her salience measurements, as I demonstrate here.

5 Corpus study

5.1 Data and methods

The data set for the present study, ICE-GB, is designed to represent speech and writing in Great Britain during the early 1990s. The corpus consists of approximately 1 million words, in 500 texts (300 spoken, 200 written) of 2,000 words each. Speakers and writers in the corpus are from the UK, over 18 years of age, and have completed school in English (Greenbaum 1996: 6). The corpus is not controlled for numerous other variables, including topic or content,[2] or formality, nor are speakers and writers controlled for sociolinguistic features such as gender identification, age, education, or racial identification, and so on.[3]

The present study builds on previous work by Gilquin (2008) and Werner and Mukherjee (2012) by analysing corpus frequencies of various senses of *take* and *give*. In addition, the present study also analyses *make*, another high-frequency, polysemous verb with both light and concrete senses.[4]

5.2 Data analysis

First, a semasiological comparison was performed, similar to that in Gilquin (2008) and Werner and Mukherjee (2012). All instances of all forms of all three verbs were identified in ICE-GB using the ICECUP interface (Nelson et al. 2002). Senses of each verb were identified manually by reading each instance of each verb and cataloguing it as either concrete or light. Concrete senses are indicated by the presence of a direct object that is directly perceptible by the five senses. Building on Gilquin's (2008) and Werner and Mukherjee's (2012) definitions, the concrete senses of each verb can be glossed as follows:

i. *make*: produce; create a concrete thing
ii. *take*: transfer a concrete thing towards an agent or to a destination
iii. *give*: provide; transfer a concrete thing from an agent to a recipient

[2] While most corpora are not controlled for topic, some are. For example, Baker et al. (2013) compiled a corpus of articles about Islam published in the British press. Also, the *People, Products, Pets, and Pests* project has compiled a corpus of texts on topics related to animals (https://animaldiscourse.wordpress.com/).

[3] Data on gender, age, and education are available for ICE-GB via ICECUP (Nelson et al. 2002), but ICE-GB was not sampled in order to balance those features (Greenbaum 1996).

[4] This research constitutes a portion of a much larger research project examining semasiology and onomasiology of *make, take*, and *give* in ICE components representing Great Britain, Singapore, and Hong Kong (cf. Mehl 2017; Mehl In press).

As I discuss further below in this section, the concrete sense of *take*, as stated here, raises unique problems in the onomasiological analysis.

Light uses of each verb are identified as those instances occurring with a direct object that has a related verb, where the related verb is semantically equivalent to the light verb construction (cf. Poutsma 1926; Jespersen 1954; Dixon 1991; Algeo 1995; Huddleston and Pullum 2002: 290–294; Dixon 2005; Karimi 2013; Ronan and Schneider 2015). For example, the direct object in *make a decision* is *decision*; the related verb is *decide*, and *decide* is roughly equivalent in meaning to *make a decision*. No restrictions are placed on the related verb's form (e.g. whether it is isomorphic with the direct object's form); nor are restrictions placed on other grammatical alteration (such as passivisation of the light verb construction or related verb), or grammatical modification (such as adjective modifiers of the light verb's direct object or adverb modifiers of the related verb).

A semasiological analysis was then performed, mirroring Gilquin (2008) and Werner and Mukherjee (2012), and measuring exposure rates to the concrete senses and light uses of each verb in speech and writing in ICE-GB. This measure indicates the rate at which a reader or listener will encounter the concrete sense or the light use of the given verb in the sample. Table 1 shows the raw numbers in the written portion of ICE-GB.

Table 1: Raw frequency of occurrence of concrete senses and light uses of *make, take*, and *give* in the written portion of ICE-GB.

	make	***take***	***give***
Concrete	68	62	52
Light	321	85	167

Figure 1 shows that, in the written portion of ICE-GB, the light use of each verb is more common than the concrete sense. For example, out of the total number of instances of *make* in all concrete and light uses, just over 80% of instances are the light use, and just under 20% are the concrete use.

The written corpus data in Figure 1 corroborates Gilquin's (2008) and Werner and Mukherjee's (2012) observations that light uses occur more frequently than concrete senses. I interpret this as an exposure rate: readers of British English (as represented by the text types in ICE-GB) can expect to encounter light uses of each verb more frequently than concrete senses. This in turn may relate to theories of entrenchment via exposure (cf. Schmid 2007).

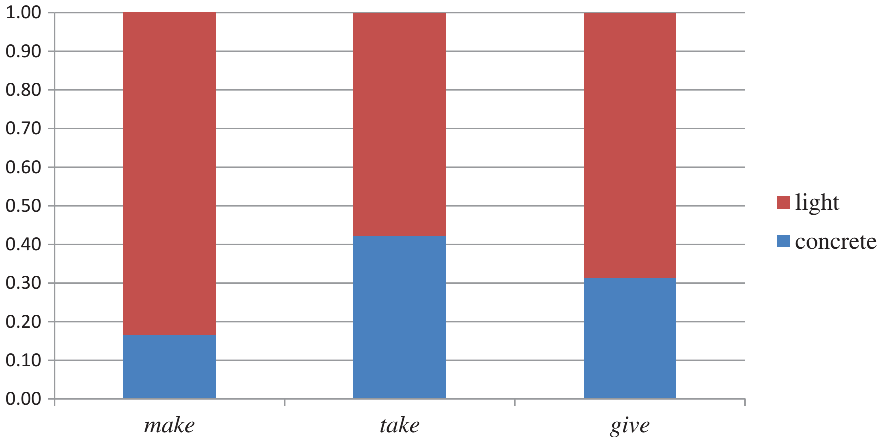

Figure 1: Exposure rates for *make, take*, or *give* (in all inflectional forms) with the concrete sense or the light use in the written portion of ICE-GB. The *y*-axis represents exposure rates for each sense in relation to the other, from 0 to 1.0.

Table 2 presents the raw numbers for the concrete sense and light use of each verb in the spoken portion of ICE-GB, for comparison to the raw numbers in Table 1. Table 2 shows that in the spoken data, the light use is more common than the concrete sense for *make* and *give*, whereas for *take*, the concrete sense occurs more than the light use. Figure 2 then displays as probabilities the raw numbers from Table 2.

Table 2: Raw frequency of occurrence of concrete and light uses of *make, take*, and *give* in the spoken portion of ICE-GB.

	make	***take***	***give***
Concrete	96	131	105
Light	353	79	227

Unlike in the written data, in the spoken data, Gilquin's (2008) and Werner and Mukherjee's (2012) observations are not corroborated entirely: it is not the case that the light use is consistently more common than the concrete use. For *take* in the spoken portion of ICE-GB, the concrete sense is more common than the light use. It is clear that this may occur in corpora that are not controlled for topic or real-world context; the written texts in ICE-GB may simply contain a large number of discussions involving the transferral of concrete objects.

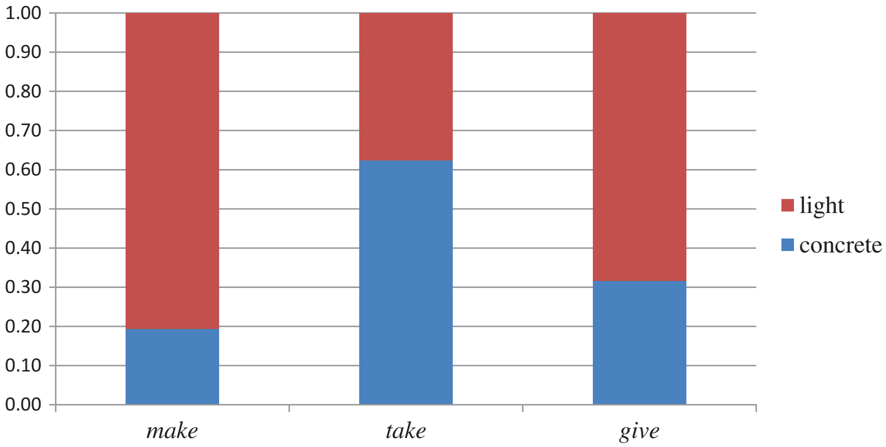

Figure 2: Exposure rates for *make, take*, and *give* (in all inflectional forms) with the concrete sense or the light use in the spoken portion of ICE-GB. The *y*-axis represents exposure rates from 0 to 1.0.

Following the semasiological analysis, onomasiological alternates were then identified in a data-driven way for each sense. No pre-existing candidate list of alternates was employed. Instead, for the concrete sense of each verb, onomasiological alternates in the corpus are defined as those verbs that occur in the corpus with the same concrete direct objects as *make, take*, and *give*, and with a roughly equivalent meaning. *Roughly equivalent meaning* is generally straightforward; in nearly all cases, recognising equivalent meaning is identified through manual, close reading of each of thousands of examples of each verb, with each concrete direct object, in corpus context. For example, *make compost* and *produce compost* both occur in the corpora, and close reading of contexts suggests that they occur with roughly equivalent meaning. *Make compost* and *carry compost* both occur in the corpora as well, but close reading indicates that the meanings are not roughly equivalent. A close reading of each example of each verb in context is absolutely necessary for this sort of semantic analysis; automation simply would not suffice.

After all concrete alternates in ICE-GB were identified for each verb, a Newcombe–Wilson test with continuity correction was performed across the full set of alternates,[5] for a rough picture of significant differences in selection

[5] Results of a Newcombe–Wilson test with continuity correction will differ only rarely from a comparable Chi-square test (Wallis 2009). One advantage of the Newcombe–Wilson test is that it does not allow confidence intervals to extend below 0 or above 1, which would be a logical

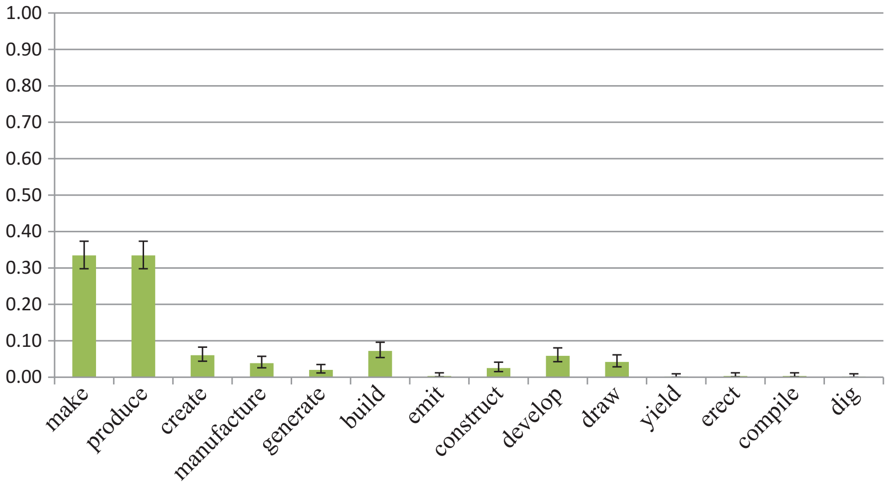

Figure 3: Probability of occurrence of concrete *make* and 13 semantic alternates in ICE-GB. Error bars represent Wilson intervals.

preferences across the full set of alternates. As displayed in Figure 3, it was determined that most concrete alternates occur so rarely in the data that no distinguishable preference for or against them is apparent.

For example, *construct*, *manufacture*, and *create* are all alternates for concrete *make*, but raw frequencies for each are very low, and no statistically significant difference is discernible. Upon further inspection of the semantics of each alternate, it was further determined that most concrete alternates are also so semantically specific that they do not truly alternate with each other: for example, language users in the corpora *make holes* and *dig holes*, but do not *manufacture holes*; likewise, they *make products* and *manufacture products*, but they do not *dig products*. Thus, while *dig* and *manufacture* can both alternate with *make*, they do not alternate with all instances of *make*, and they do not alternate with each other. Because of this lack of universal alternation, as well as the low frequency of occurrence of most alternates, the most highly frequent, semantically general alternate was chosen for a pairwise comparison with each verb. In fact, the most high-frequent alternate was in each case also the most semantically general. The most highly frequent, semantically general alternate in the corpus for *make* is

impossibility. While other statistical tests could be legitimately applied, this test is well justified, and it is not standard procedure to compare various tests against each other unless the tests themselves are the object of the investigation.

produce[6]; and for *give* is *provide*. For *take*, two highly frequent, semantically general alternates can be identified: *collect* and *carry*. *Collect* can be glossed as "transfer a concrete thing towards an Agent" and *carry* as "transfer a concrete thing, by an Agent, to a destination." These two alternates are aggregated for a pairwise comparison, such that *take* conveying either meaning is compared to the aggregated instances of *collect* + *carry*. Data on these alternations appear below.

An onomasiological analysis of the concrete sense of each verb appears below.

A single-sample Chi-square test is performed on each alternation, comparing actual selection preferences to expected selection preferences for each alternate if each was selected randomly: the expected frequency for each alternate is thus 50% of the total number of instances of both alternates. The null hypothesis for this test is that the underlying selection preference for each verb or its alternate is random. The single-sample Chi-square test shows that in speech, each concrete verb (*make, take*, and *give*) is preferred over its alternate beyond what is expected by chance ($p < 0.05$). Put differently, each concrete verb is significantly preferred over its alternate in speech. The raw data in Table 3 are displayed as probabilities in Figure 4.

Table 3: Raw frequency of occurrence of concrete *make, take*, and *give* and their respective semantic alternates in the spoken portion of ICE-GB.

Concrete verb	Instances in speech	Concrete alternate	Instances in speech
make	96	*produce*	64
take	131	*carry* + *collect*	38
give	105	*provide*	11

Table 4: Raw frequency of occurrence of concrete *make, take*, and *give* and their respective semantic alternates in the written portion of ICE-GB.

Concrete verb	Instances in writing	Concrete alternate	Instances in writing
make	68	*produce*	135
take	62	*carry* + *collect*	49
give	52	*provide*	57

Table 4 presents raw frequency of occurrence of concrete *make*, *take*, and *give* and their respective semantic alternates in the written portion of ICE-GB. Figure 5 displays the probability that the given verb or its alternate appears in the written portion of ICE-GB. The data indicate that in writing, preferences for the alternate

[6] Senses in which *make* relates to the production of food are removed from this data, as *produce* does not occur in the corpus with direct objects representing food.

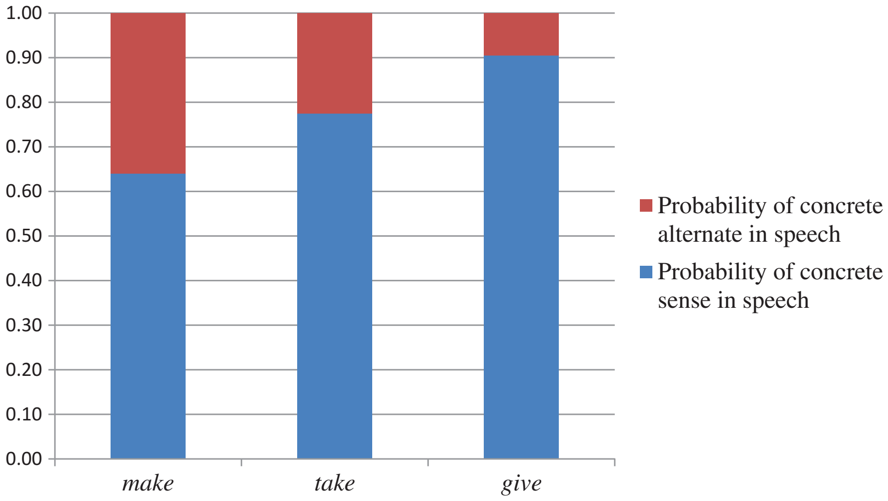

Figure 4: Selection preferences for concrete *make, take*, and *give* and their respective semantic alternates in the spoken portion of ICE-GB. The *y*-axis represents selection probabilities from 0 to 1.0.

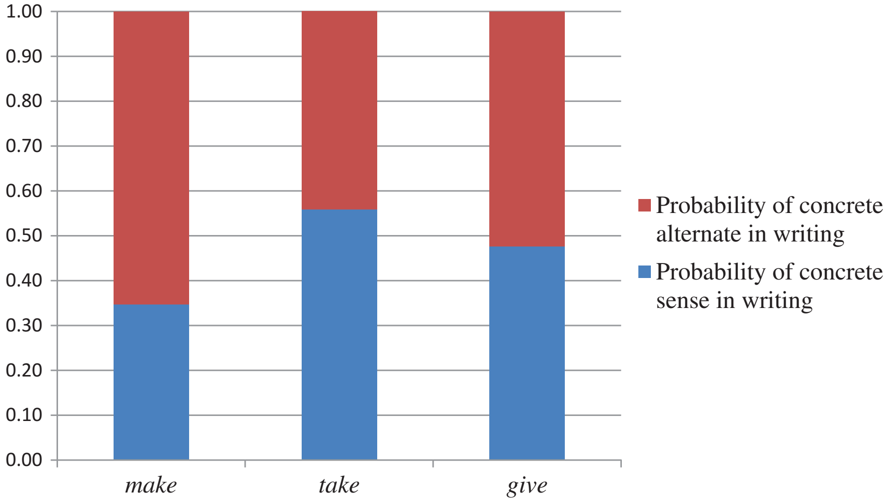

Figure 5: Selection preferences for concrete *make, take*, and *give* and their respective semantic alternates in the written portion of ICE-GB. The *y*-axis represents selection probabilities from 0 to 1.0.

verbs rise considerably. *Produce* is preferred over *make* at a probability of around 65% to 35%. A single-sample Chi-square test shows that this preference is significantly stronger than would be expected by chance ($p < 0.05$). A single-sample Chi-square test fails to refute the null hypothesis: that is, preferences for *take* in relation to *collect* + *carry* and for *provide* in relation to *give* are indistinguishable from chance ($p > 0.05$). Thus, selection preferences differ according to register, such that monosyllabic, Germanic alternate is strongly preferred in speech, while the polysyllabic, Latinate alternate increases in probability in writing.

An onomasiological analysis was then conducted on the light uses of each verb. For the light use of the verb, onomasiological alternates arise from the definitional nature of light verb constructions: the direct object of the light verb construction has a related verb whose meaning is equivalent to the light verb construction. Thus, the onomasiological alternate of each light verb construction is its related verb. For example, *make a decision* alternates with *decide*. Analysed light verb constructions are displayed in Table 5 and Table 6.

Table 5: Raw frequency of occurrence of light *make, take*, and *give* constructions and their respective semantic alternates in the spoken portion of ICE-GB.

Light verb construction	Instances in speech	Related verb (alternate)	Instances in speech
make a decision	28	*decide*	156
make use	5	*use* (v.)	488
make a change	12	*change* (v.)	66
make a contribution	14	*contribute*	19
take a decision	12	*decide*	156
take a look	3	*look* (v.)	288
take action	22	*act* (v.)	12
give support	7	*support* (v.)	59
give information	7	*inform*	24

Table 6: Raw frequency of occurrence of light *make, take*, and *give* constructions and their respective semantic alternates in the written portion of ICE-GB.

Light verb construction	Instances in writing	Related verb (alternate)	Instances in writing
make a decision	31	*decide*	106
make use	6	*use* (v.)	697
make a change	8	*change* (v.)	25
make a contribution	5	*contribute*	29
take a decision	9	*decide*	106
take a look	3	*look* (v.)	56
take action	12	*act* (v.)	9
give support	4	*support* (v.)	74
give information	9	*inform*	20

Light verb constructions that occur at least three times in the corpus, and whose onomasiological alternates occur at least five times in the corpus, were identified and analysed, via a single-sample Chi-square on each alternation, comparing actual selection preferences to expected selection preferences of 0.5 for each pair of alternates, or 0.33 for each trio of alternates, if each was selected randomly. The null hypothesis for this test is that the selection preference for each verb or its alternate is random.

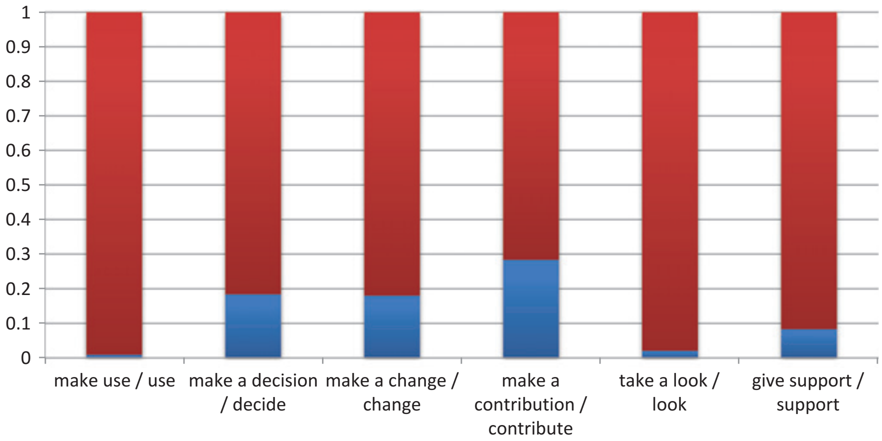

Figure 6: Selection preferences for light *make, take*, and *give* constructions and their respective semantic alternates in the aggregated spoken and written portions of ICE-GB. The light verb construction appears in blue, the related verb alternate in red. The *y*-axis represents selection probabilities from 0 to 1.0.

With most light verb constructions, the related verb alternate is preferred over the light verb construction in both speech and writing. Figure 6 displays the probability for selecting either the light verb construction or its alternate related verb for all five pairs or trios that show this strong preference in both speech and writing. For example, speakers and writers select *make use* in only 1% of opportunities and *use* (v.) in 99% of opportunities.[7]

[7] I am grateful to an anonymous reviewer for suggesting the possibility that onomasiological selection preferences for or against light verb constructions might relate to mutual information (MI) scores for the verb–DO pairing in the construction. MI is a measure of information in a system (cf. Fano 1961), and it can generally be conceptualised in linguistics as a measure of how non-random a sequence of linguistic features appears (cf. Church and Hanks 1990). It is conceivable that a very low or negative MI score for a light verb construction might relate to a strong selection preference against that light verb construction, favouring instead the related verb alternate. MI is generally measured against a baseline of all words in a sample, but it can

There are, however, exceptions to the general trend presented in Figure 6. Two pairs do not show the same significant preference for the related alternate verb over the light verb construction in both speech and writing: *take action*/*act* (v.) and *give information*/*inform*. There is no significant preference for either *take action* or *act* (v.) in speech or writing.[8] *Inform* is significantly preferred over *give information* and *provide information* in speech, but preferences for the three forms are indistinguishable from random probabilities in writing.

In sum, the onomasiological trends for light verb constructions are complex – certainly more complex than the trend in concrete senses for the same verbs. Nonetheless, in most cases, the related alternate verb tends to be preferred over its light verb construction counterpart in speech and writing.

5.3 Discussion

First, the data illustrate the extremely consequential differences between a semasiological analysis and an onomasiological analysis. In sum, and to simplify slightly: semasiologically, concrete senses exhibit low relative frequency (i.e. lower than other senses of the same verb) while light uses exhibit high relative frequency. Conversely, onomasiologically, concrete senses exhibit high relative frequency (i.e. higher than their semantic alternates) while light uses exhibit low relative frequency.

Semasiologically, in the corpora, light *make, take*, and *give* tend to be most frequent, and concrete *make, take*, and *give* least frequent (with the exception of *take* in writing). This generally corroborates Gilquin's (2008) and Werner and Mukherjee's (2012) observations, but the exception of *take* in writing also underlines the difficulties in confidently measuring exposure rates. How do we interpret these semasiological findings? As discussed in Section 2, a semasiological analysis indicates an exposure rate. A reader or listener encountering British English in the text types sampled for ICE-GB might expect to encounter light

be measured against a grammatical baseline (cf. Fitzmaurice et al. 2017), or an onomasiological baseline. Multiple baselines would need to be employed in testing this hypothesis in relation to the present work, which is promising ground for future research, but is beyond the scope of the present paper.

[8] *Act* (v.) is highly polysemous. Instances that do not alternate with *take action*, such as examples of the sense "perform," e.g. *to act in a play*, have been manually identified; such instances are not counted as alternates in the present study. Again, this process underlines the absolute necessity of close reading and manual analysis of every single example in the corpus.

make, take, and *give* more often than concrete *make, take*, and *give*, with the exception of *take* in writing. This exposure rate would be important to lexicographers designing dictionaries, who might choose to list the most semasiologically frequent sense first. Indeed, the Collins COBUILD (2006) dictionary does list the light uses of each verb first. Semasiological frequencies may also reflect facts about topic or real-world context in the corpus – neither of which is systematically controlled. The occurrence of concrete senses depends upon a discussion of a particular range of concrete situations in the real world – generally related to the transfer or movement of a concrete object. That is, occurrence of concrete senses relates to real-world topics of conversation. The occurrence of light uses is much more flexible, given the range of light constructions that can occur in a range of real-world scenarios (such as *take a look* or *make use*, which can be used in concrete or non-concrete contexts, in an array of situations). We had reason to hypothesise above a general trend in which light uses are more common than concrete uses. However, we cannot deny the possibility that a given text or context might sometimes require concrete uses even more than light ones, given the topic or context, and the communicative needs. In particular, a text or set of texts about the transfer of concrete things might be expected to affect these results. Because ICE-GB is not controlled for topic, a systematic re-sampling of ICE-GB might result in a very different array of topics, and a different set of exposure rates. So, the observation that concrete *take* occurs more than light *take* may simply be an epiphenomenon of uncontrolled variables – the array of topics and real-world contexts represented in the ICE corpora.[9]

Onomasiological findings are nearly the converse of the semasiological findings: concrete senses tend to exhibit high relative frequency, and light uses tend to exhibit low relative frequency. How do we interpret onomasiological findings? An onomasiological analysis indicates selection preferences. Speakers and writers of British English tend to select concrete *make, take*, and *give* more than their semantic alternates in speech (see Figure 4), with an increased preference for their semantic alternates in writing (see Figure 5). Speakers and writers of British English tend to prefer related verbs over light

[9] Expanding the present study to larger, less-curated corpora would in turn expand the problems of the present study. Typical very large corpora such as the British National Corpus (BNC) are no more controlled for real-world context or topic than ICE-GB, and the arbitrary or erratic nature of topics in ICE-GB is only magnified in the arbitrary, erratic nature of topics in the much larger BNC. In addition, the semantic analysis here depends entirely on close human reading of each example in its utterance and discourse context. This is a considerable undertaking for the thousands of examples here, but it becomes prohibitive with much larger corpora.

verb constructions in both speech and writing, with important exceptions. Onomasiological observations could be relevant for style guides or for language instructors who teach students to follow established norms related to speech and writing, or, perhaps, formal and informal language: for example, it might be useful to encourage students to consider selecting *produce* rather than *make* in written language.

I would like to present a working hypothesis, based on the present data, that corpus frequencies may in fact correlate with the sorts of semasiological elicitation tests for cognitive salience conducted by Gilquin (2008), if corpus frequencies are measured onomasiologically in spoken data.[10] To summarise:

i. In elicitation tests, concrete senses are generated most frequently (Gilquin 2008).
ii. Onomasiologically, concrete senses exhibit high relative frequency (i.e. higher frequency than their semantic alternates) in speech (but not in writing).

It is clear from this evidence that elicitation tests correlate with onomasiological relative frequencies in speech, but not in writing.

iii. In elicitation tests, light uses are generated least frequently (Gilquin 2008).
iv. Onomasiologically, light uses tend to exhibit low relative frequency (i.e. lower than their semantic alternates) in speech and writing, with exceptions for some light verb constructions.

From this evidence, it is clear that elicitation tests tend to correlate with onomasiological relative frequencies in speech, but not with semasiological frequencies. All of these observations align with Geeraerts's (2010) hypothesis of onomasiological salience, which states that onomasiological corpus frequencies ought to indicate cognitive salience and prototypicality. Onomasiological salience may therefore be a more useful notion in considering prototypicality, salience, and corpus frequencies than more traditional notions of entrenchment through high exposure rates. The theoretical mechanism for this observation is a process of entrenchment not through exposure rates but through selection preferences: the process of selecting a word form over its alternate results in routinisation and entrenchment of that word form to express that meaning, which in turn results in higher cognitive salience for the form–meaning relationship.

[10] Gilquin (2008: 8) also observes that " … although as a rule the spoken data come slightly closer to the elicitation data than the written data, considering the spoken data only still results in a discrepancy between frequency and elicitation."

6 Conclusions

Gilquin (2008) asserted that concrete senses tend to be most commonly generated in elicitation tests for cognitive salience, while light uses tend to occur most frequently in corpus data. The present study takes a step further by distinguishing between two types of relative frequency in corpus semantics: semasiological and onomasiological. In doing so, the present study demonstrates the value of operationalising relative corpus frequencies carefully and explicitly in relation to existing theoretical frameworks and the given research questions. In addition, the present study moves towards resolving the apparent contradiction that Gilquin observes. Elicitation test results may correlate with relative corpus frequencies if corpus frequencies are measured onomasiologically, and in speech.

An onomasiological analysis reflects, quite simply, a different research approach from a semasiological analysis. In corpus semantics, therefore, it is absolutely necessary that the frequency measure (semasiological, onomasiological, or otherwise) be explicitly stated and justified in relation to the research question. Research questions regarding exposure rates are best addressed using a semasiological analysis, while research questions regarding selection preferences are best addressed using an onomasiological analysis.[11] Most importantly for the present study, onomasiological relative frequency, in speech, seems to correlate most closely to results of elicitation tests for cognitive salience. This would seem to affirm a theoretical framework of onomasiological salience, rather than a framework in which exposure rates lead to entrenchment. Geeraerts's (2010) notion of onomasiological salience might therefore be modified such that it is spoken language in particular in which onomasiological salience ought to be measured.

Future research can address the specific questions of corpus frequencies and prototypicality via additional elicitation tests, including both written and spoken elicitation tests, and both semasiological and onomasiological elicitation tests. In addition, more polysemous lexical items ought to be investigated onomasiologically, in relation to elicitation tests. Further investigations should also be conducted along the lines of Werner and Mukherjee (2012), into varieties of English worldwide, asking whether elicitation test results or onomasiological relative frequencies vary by geographic region. For example, Mehl (2017) presents broad similarities, with some complex differences as well, between Singapore English, Hong Kong English, and British English in onomasiological frequencies for light verb constructions.

[11] Semasiological and onomasiological analyses can of course complement each other, as in Geeraerts's (1997) study on semantic variation and change in contemporary Dutch.

If onomasiological frequency measurements do indeed correlate with elicitation tests, potential impact would be immense. Researchers would be able to examine onomasiological frequencies in spoken corpora rather than performing elicitation tests. That possibility would facilitate cognitive research into languages and varieties around the world, without the necessity of *in situ* psycholinguistic testing, and would also encourage the creation of more spoken corpora. This would represent a dramatic shift in data collection methods in linguistics.

References

Algeo, John. 1995. Having a look at the expanded predicate. In Bas Aarts & Charles Meyer (eds.), *The verb in contemporary English: Theory and description*, Cambridge: Cambridge University Press.

Arppe, Antti, Gaetanelle Gilquin, Dylan Glynn, Martin Hilpert & Arne Zeschel. 2010. Cognitive corpus linguistics: Five points of debate on current theory and methodology. *Corpora* 5(1). 1–27.

Baker, Paul, Costas Gabrielatos & Tony McEnery. 2013. Sketching Muslims: A corpus driven analysis of representations around the word 'Muslim' in the British press 1998–2009. *Applied Linguistics* 34(3). 255–278.

Balasubramanian, Chandrika. 2009. Circumstance adverbials in registers of Indian English. *World Englishes* 28(4). 485–508.

Bybee, Joan & Paul Hopper (eds.). 2001. *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins Publishing Company.

Church, Kenneth Ward & Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1). 22–29.

Collins CoBUILD English Dictionary. 1995. 2nd edn. Glasgow: Harper Collins.

Divjak, Dagmar & Stefan Th Gries. 2012. *Frequency effects in language representation*. Berlin: De Gruyter Mouton.

Dixon, Robert M. W. 1991. *A New approach to English Grammar, on semantic principles*. Oxford: Oxford University Press.

Dixon, Robert M. W. 2005. She gave him a look, they both had a laugh and then took a stroll: GIVE A VERB, HAVE A VERB and TAKE A VERB constructions. In Robert M. W. Dixon (ed.), *A Semantic approach to English Grammar*, 459–483. Oxford: Oxford University Press.

Evison, Jane. 2010. What are the basics of analysing a corpus?. In Anne O'Keefe & Michael McCarthy (eds.), *The Routledge handbook of corpus linguistics*, 122–135. London: Routledge.

Fano, Robert M. 1961. *Transmission of information: A statistical theory of communications*. Boston: MIT Press.

Fitzmaurice, Susan, Justyna A. Robinson, Marc Alexander, Iona C. Hine, Seth Mehl & Fraser Dallachy. 2017. Linguistic DNA: Investigating conceptual change in Early Modern English discourse. *Studia Neophilologica* 89. 21–38.

Fuchs, Robert. 2012. Focus marking and semantic transfer in Indian English. *English World-Wide* 33(1). 27–52.

Fuchs, Robert, Ulrike Gut & Taiwo Soneye. 2013. "We just don't even know": The usage of the pragmatic focus particles *even* and *still* in Nigerian English. *English World-Wide* 34(2). 123–145.

Geeraerts, Dirk. 1988. Where does prototypicality come from? In Brygida Rudzka-Ostyn (ed.), *Topics in Cognitive Linguistics*, 207–229. Amsterdam: John Benjamins.

Geeraerts, Dirk. 1997. *Diachronic prototype semantics: A contribution to historical lexicology*. Oxford: Clarendon Press.

Geeraerts, Dirk. 2006 [1989]. Prospects and problems of prototype theory. In Dirk Geeraerts, *Words and other wonders*, 3–26. Berlin: Mouton de Gruyter.

Geeraerts, Dirk. 2010. *Theories of lexical semantics*. Oxford: Oxford University Press.

Gilquin, Gaetanelle. 2006. The place of prototypicality in corpus linguistics: Causation in the hot seat. In Stefan Gries & Anatol Stefanowitsch (eds.), *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis*, 159–191. Berlin: Mouton de Gruyter.

Gilquin, Gaetanelle. 2008. What you think ain't what you get: Highly polysemous verbs in mind and language. In Jean-Remi Lapaire, Guillaume Desagulier & Jean-Baptiste Guignard (eds.), *From gram to mind: Grammar as cognition*, 235–255. Bordeaux: Presse Universitaires de Bordeaux.

Glynn, Dylan. 2014. Polysemy and synonymy: Cognitive theory and corpus method. In Dylan Glynn & Justyna A. Robinson (eds.), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy*, 7–38. Amsterdam: Benjamins.

Greenbaum, Sidney. 1996. Introducing ICE. In Sidney Greenbaum (ed.), *Comparing English worldwide: The International Corpus of English*, 3–12. Oxford: Clarendon Press.

Gries, Stefan Th. 2006. Corpus-based methods and cognitive semantics: The many senses of *to run*. In Stefan Gries & Anatol Stefanowitsch (eds.), *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis*, 57–99. Berlin: Mouton de Gruyter.

Gries, Stefan Th. & Dagmar Divjak. 2012. *Frequency effects in language learning and processing*. Berlin: De Gruyter Mouton.

Haase, Christoph. 1994. Conceptual specifics in East African English: Quantitative arguments from the ICE-East Africa corpus. *World Englishes* 23(2). 261–268.

Heylen, Kris, Jose Tummers & Dirk Geeraerts. 2008. Methodological issues in corpus-based Cognitive Linguistics. In Gitte Kristiansen & René Dirven (ed.), *Cognitive Sociolinguistics: Language variation, cultural models, social systems*, 91–128. Berlin: Mouton de Gruyter.

Huddleston, Rodney & Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Hundt, Marianne. 2009. How often to things *get V-ed* in Philippine and Singapore English? A case study of the *get* – passive in two outer-circle varieties of English. In Rhonwen Bowen, Mats Mobarg & Solve Ohlander (eds.), *Corpora and discourse – and stuff: Papers in honor of Karin Aijmer*. Gothenburg Studies in English 96, 121–131. Gothenburg: Gothenburg University Press.

Jespersen, Otto. 1954. *A modern English grammar on historical principles, Part VI: Morphology*. London: Bradford and Dickens.

Johnson, Mark. 2007. *The meaning of the body: Aesthetics of human understanding*. Chicago: University of Chicago Press.

Karimi, Simin. 2013. Introduction. *Lingua* 135. 1–6.

Lakoff, George & Mark Johnson. 1980. *Metaphors we live by*. Chicago: University of Chicago Press.

Lange, Claudia. 2007. Focus marking in Indian English. *English Worldwide* 28(1). 89–118.

Lee, Sarah & Debra Ziegeler. 2006. Analysing a semantic corpus study across English dialects: Searching for paradigmatic parallels. In Andrew Wilson, Dawn Archer & Paul Rayson (eds.), *Corpus linguistics around the world*, 121–139. Amsterdam: Rodopi.

Lindqusit, Hans. 2009. *Corpus linguistics and the description of English*. Edinburgh: Edinburgh University Press.

McEnery, Tony, Richard Xiao & Yukio Tono. 2006. *Corpus-based language studies: An advanced resource book*. Abingdon: Routledge.

McEnery, Tony & Costas Gabrielatos. 2006. English corpus linguistics. In Bas Aarts & April McMahon (eds.), *The handbook of English linguistics*, 33–71. Malden, MA: Blackwell.

McEnery, Tony & Andrew Wilson. 2001. *Corpus linguistics*, 2nd edn. Edinburgh: Edinburgh University Press.

Mehl, Seth. 2017. Light verb semantics in the International Corpus of English: Onomasiological variation, identity evidence, and degrees of lightness. *English Language and Linguistics*. doi:https://doi.org/10.1017/S1360674317000302 (accessed 12 January, 2018).

Mehl, Seth. In press (accepted 2017). Corpus onomasiology in world Englishes and concrete verbs *make* and *give*. *World Englishes*.

Nelson, G., Bas Aarts & S. A. Wallis. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins.

Nordquist, D. 2004. Comparing elicited data and corpora. In M. Achard & S. Kemmer (eds.), *Language, culture and mind*, 211–224. Stanford: CSLI Publications.

Poutsma, H. 1926. *A grammar of Late Modern English*. Groningen: P. Noordhoff.

Ronan, Patricia & Gerold Schneider. 2015. Determining light verb constructions in contemporary British and Irish English. *International Journal of Corpus Linguistics* 20(3). 326–354.

Rosch, Eleanor. 1973. Natural categories. *Cognitive Psychology* 4(3). 328–350.

Rosch, Eleanor. 1975a. Cognitive reference points. *Cognitive Psychology* 7. 532–547.

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

Schmid, Hans-Jörg. 2007. Entrenchment, salience and basic levels. In Dirk Geeraerts & Hubert Cuyckens (eds.), *The Oxford Handbook of Cognitive Linguistics*, 117–138. Oxford: Oxford University Press.

Schneider, Edgar W. 1994. How to trace structural nativization: Particle verbs in World Englishes. *World Englishes* 23. 227–249.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Taylor, John. 2003. *Linguistic categorization*, 3rd edn. Oxford: Oxford University Press.

Taylor, John R. 2012. *The mental corpus: How language is represented in the mind*. Oxford: Oxford University Press.

Wallis, S. A. 2009. *Binomial confidence intervals and contingency tests: Mathematical fundamentals and the evaluation of alternative methods*. London: UCL Survey of English Usage. http://www.ucl.ac.uk/english-usage/staff/sean/resources/binomialpoisson.pdf (accessed 1 November 2016.).

Wallis, S. A. 2012. *That vexed problem of choice: Reflections on experimental design and statistics with corpora*. London: UCL Survey of English Usage. http://www.ucl.ac.uk/english-usage/staff/sean/resources/vexedchoice.pdf (accessed 1 November 2016.).

Werner, Janina & Joybrato Mukherjee. 2012. Highly polysemous verbs in New Englishes: A corpus-based pilot study of Sri Lankan and Indian English. In Sebastian Hoffman (ed.), *English corpus linguistics: Looking back, moving forward*, 249–266. Amsterdam: Rodopi.