Rachel E. Ramsey*

Individual differences in word senses

https://doi.org/10.1515/cog-2021-0020
Received February 14, 2021; accepted September 3, 2021; published online October 19, 2021

Abstract: Individual differences and polysemy have rich literatures in cognitive linguistics, but little is said about the prospect of individual differences in polysemy. This article reports an investigation that sought to establish whether people vary in the senses of a polysemous word that they find meaningful, and to develop a novel methodology to study polysemy. The methodology combined established tools: sentence-sorting tasks, a rarely used statistical model of inter-participant agreement, and network visualisation. Two hundred and five English-speaking participants completed one of twelve sentence-sorting tasks on two occasions, separated by a delay of two months. Participants varied in how similarly they sorted the sentences as compared to other participants, and mean agreement across all 24 tasks did not meet an established threshold of acceptable agreement. Between the two test phases, inter-participant agreement varied to a significant but trivial degree. Networks generated for each dataset varied in the degree to which they captured all participants' responses. This variation correlated with inter-participant agreement. The data collectively suggest that word senses may be subject to individual differences, as is the case in other linguistic phenomena. The methodology proved replicable and has a promise as a useful tool for studying polysemy.

Keywords: individual differences; network theory; polysemy; quantitative methods; word sense disambiguation

1 Introduction

Cognitive linguists have shown enduring curiosity about individual differences and polysemy but have said little about the prospect of individual differences *in* polysemy. Indeed, while noting "inevitable individual differences" in senses across adult native English speakers, Carston (2021: 11) acknowledges that there is a "bigger story to tell, … one that accommodates individual differences." (p. 16). Natural language processing (NLP) scholars, particularly those who develop word

***Corresponding author: Rachel E. Ramsey**, Northumbria University, Newcastle upon Tyne, UK; and Dubit Limited, Leeds, UK, E-mail: rachelramsey@gmail.com

sense disambiguation (WSD) algorithms, have long assumed that such differences exist. Moreover, evidence gathered using WSD methods indicates that this is the case. It is unclear why this has not been picked up in cognitive linguistics, since individual differences have been observed by scholars in the field across a wide range of linguistic phenomena. This article seeks to address this gap in the field's literature by asking a simple question: do individuals have different senses of a given polysemous word? Of course, the idea that they might is anything but simple. However, the theoretical scope of this article is intentionally focused and simple, seeking only to examine quantitative data to explore whether such differences might exist. Only then can the field begin to explore what factors shape these differences and consider the implications of individual differences for existing descriptions of polysemy, such as how senses are related (or not), and what model of categorisation best describes their representation. As part of this aim, and in response to the call for our conclusions to be borne out of data, quantitative or otherwise (Dąbrowska 2016), the article also asks whether some established methods can be combined to create a novel methodology, enabling new insights about polysemy.

1.1 Individual differences

The proposition that not all members of a linguistic community acquire the same linguistic system runs counter to generative theories and is undermined by empirical research. Kidd et al. (2018) review recent research on individual differences in language acquisition and processing, concluding that variation may pervade the entire linguistic system. LuCiD, the landmark study of 80 English-learning children from 6 months to 4.6 years, has generated a vast literature indicating the roles of various dimensions of difference in language acquisition (LuCiD 2021).

Individual differences persist into adulthood. ERP studies of language comprehension reviewed by Tanner et al. (2018) indicate both quantitative differences (i.e., in N400 and P600 amplitudes), and qualitative differences (i.e., different types of effects across individuals). In a review of investigations of differences in linguistic attainment between high academically achieving (HAA) and low academically achieving (LAA) populations, Dąbrowska (2012) notes that while HAA participants typically performed at ceiling level – meaning that their responses are typically homogenous and correct – LAA participants showed greater variation in their responses to a range of linguistic stimuli. Her recent work concludes that individual differences are observed even in everyday-occurring grammatical constructions. For example, 41% of native English-speaking

participants participating in a grammatical comprehension test performed below chance with the locative with quantifier construction (e.g., *every pencil is in a box*) (Dąbrowska 2018). Skoe et al. (2017) propose that variation in reading ability in adults may result from variation in auditory brainstem function. Personality type is implicated in explaining certain individual differences. Dunlop et al. (2020) describe systematic variation in first person pronoun use according to adult romantic attachment styles. Duffy et al. (2014) observe differences in interpretation of McGlone and Harding's (1998) ambiguous statement *The meeting originally scheduled for next Wednesday has been moved forward two days*. Participants displaying conscientious behaviours typically used the Moving Time perspective to interpret the statement, saying the meeting will now be held on Monday. Those showing procrastinating behaviours adopt the Moving Ego perspective, saying that the meeting will now be held on Friday. The same pattern was found in self-reported conscientiousness and procrastination (Duffy and Feist 2014). Stamenković et al. (2019) argue that differences in metaphor comprehension may be explained by differences in fluid and crystallised intelligence, i.e., in reasoning and information organisation abilities.

Several factors have been proposed to account for individual differences across the linguistic system, and biology, personality and intelligence appear to play key roles in shaping an individual's language. Of interest to cognitive linguists is the role of exposure in shaping both children's and adults' language. Diessel (2017: 2) summarises the argument, "As frequency strengthens the representation of linguistic elements in memory, it facilitates the activation and processing of words, categories, and constructions, which in turn can have long-lasting effects on the organization of linguistic knowledge in the language network."

This can be tested in studies where exposure to a particular linguistic phenomenon is expected to vary across participants. Street and Dąbrowska (2010) found that comprehension of passive constructions correlated with level of academic attainment and argued that this could be explained by the fact that HAA participants will have more exposure to the formal written texts in which the passive is generally found. In Dąbrowska's more recent work she concluded that vocabulary and collocational knowledge correlated with print exposure (Dąbrowska 2018).

If individuals have different linguistic systems, how is successful communication achieved?[1] In the context of the present study, it may seem problematic to posit an account of major individual differences in word senses. After all, if the meaning that a pair of individuals attribute to a single example of a polysemous

1 However, as Ferreira et al. (2002) note, communication is not always successful.

word differs, it is possible that the meaning intended by the speaker will not correspond to the listener's interpretation. However, successful communication takes place between individuals with markedly different linguistic systems (e.g., children and their caregivers), and when speech is disfluent (Ferreira et al. 2002). In a study comparing computational models' and native speakers' performance in their choice of six synonymous Russian *try*-verbs, Divjak et al. (2016: 27) proposed that multiple computational models of grammar can explain usage. Indeed, they recommend that the pursuit of a "single 'best' model" of human grammar should be side-lined in favour of developing multiple models, representing individual variation in grammar. In a somewhat different explanation of successful communication despite individual differences in grammar, Ferreira et al. (2002) propose that successful sentence interpretation in normal communicative situations is based on a "good enough" model of sentence processing, and that successful processing and interpretation is supported by contextual information.

1.2 Individual differences in word senses

Polysemy has a long history in cognitive linguistics. This longevity is perhaps motivated by the argument that polysemy is an example of linguistic categorisation (e.g., Taylor 2003). This argument aligns with the cognitive commitment (Lakoff 1990), which demands that theories about language must be compatible with what is known about cognition more broadly. To find that polysemy *is* an example of categorisation—a fundamental cognitive process—would bring credibility to cognitive linguists' argument that language is one of several interrelated cognitive functions. Consequently, much of the energy spent on exploring polysemy has been dedicated to establishing the nature of categorisation, particularly which model of categorisation best accommodates what we know about polysemy (Gries 2015). The field has, however, said little about the prospect of individual differences in polysemy, despite its longstanding interest in variation elsewhere in language, and despite the interesting theoretical and methodological implications that such differences would have. Indeed, the canonical literature by Brugman (1981), Brugman and Lakoff (1988), and Tyler and Evans (2001), while offering rich and 'principled' analyses of the senses of *over*, does not leave open the possibility that individuals may have different senses of a polysemous word. They further propose network representations of the senses which, in the absence of any caveat to the contrary, we might assume to be invariant. A network representation is a reasonable suggestion if we assume that senses are related to each other, but we should be mindful that their development has typically drawn heavily on

introspection (see Section 1.3.1 for further discussion).[2] Beyond cognitive linguistics, one could wonder whether the very existence of lists of senses in dictionaries presupposes that senses are constant across a language community.

While the prospect of individual differences in word senses has received little scrutiny in cognitive linguistics, research in natural language processing (NLP) has produced, perhaps inadvertently, a body of evidence supporting this idea. Word sense disambiguation (WSD) scholars have found that agreement with "gold standards", i.e., with expert judgements about which sense tag should be applied to each individual example of a polysemous word, varies across individuals (conventionally referred to as annotators), whether they are also "experts", or "naïve", non-expert annotators. This (dis)agreement is typically measured using one of a family of inter-annotator agreement measures that includes Krippendorff's alpha, α, Cohen's kappa, κ, Rand, and adjustments of Rand such as Morey & Agresti used in this research (Artstein and Poesio 2008; Morey and Agresti 1984; Rand 1971; see Section 1.3.4 for discussion of how agreement values are interpreted). When asked to assign examples of a word to a predetermined category based on the meaning of that word in the example, annotators agree with the gold standard to varying extents. Exceptionally high inter-annotator agreement was found in Snow et al.'s (2008) study, in which naïve annotators were asked to classify examples of the word *president*. They reached complete consensus with the gold standard; however, *president* was to be classified into one of only three possible categories: (1) executive officer of a firm, corporation or university, (2) head of a country (other than the U.S.), or (3) head of the U.S., President of the United States. An example at the opposite end of the scale comes from Passonneau et al. (2012a) WSD study of *normal*; trained annotators agreed with each other less than would be expected by chance. Between these two extremes, research demonstrates intermediate agreement. Using α, Bhardwaj et al. (2010: 4) found agreement amongst trained annotators ranged from "about 0.5 to 0.7 for nouns and adjectives, and about 0.37 to 0.46 for verbs", versus from 0.08 to 0.15 for adjectives amongst untrained annotators. Passonneau et al. (2010), also using α, found that trained annotators' decisions gave agreement scores ranging from 0.37 to 0.68. The fact that in these and similar studies the annotators each undertake identical tasks indicates differences in the extent to which individuals agree with each other about how the pre-set sense categories are used, and therefore they vary in how well they agree with the gold standard. Further, the extent of their

2 A reviewer rightly observed that a contemporary reader of these seminal works should consider the technological constraints that might have necessitated an introspective approach. Brugman (1981), for example undertook research in a pre-internet landscape that had neither widely accessible corpora nor other means of gathering the large volumes of data we might expect today.

agreement varies according to the word they are examining (Bhardwaj et al. 2010; Passonneau et al. 2009). Passonneau et al. (2009: 3) propose three factors that might affect agreement:

1. Greater specificity in the contexts of use leads to higher agreement,
2. More concrete senses give rise to higher agreement,
3. A sense inventory with closely related senses (e.g., relatively lower average inter-sense similarity scores) gives rise to lower agreement.

Later, the authors explore the effect of individual words on inter-annotator agreement and suggest that words that can be found in contexts that are more open to subjective interpretation may explain some cases of disagreement; this likely ties in with the second factor, cited above (Passonneau et al. 2012b). They suggest, for example, that the adjective *fair* is inherently more subjective than the adjective *long*, which describes a measurable physical property. Other research notes that the subjectivity of a word may be a product of different "perception[s] and experience[s] of individuals", giving *justice* as an example of a word that derives its "meaning from cultural norms that may differ from community to community" (Bhardwaj et al. 2010: 2).

These studies suggest that different people assign the same examples of a word to different sense categories, suggestive of individual differences in word senses. To a reader versed in NLP literature, these findings will not come as a surprise. As Passonneau et al. (2009: 3) put it, "It is widely recognized that achieving high [agreement] scores […] is difficult for word sense annotation," while Passonneau et al. (2010: 1) say "variation in word sense annotation across annotators should be expected as a consequence of usage variation." They note that in lexicographic and linguistic literature, "it is taken for granted that there will be differences in judgment across language users regarding word sense" (Passonneau et al. 2012b: 11). In the cognitive linguistic literature, typical studies of polysemy make no reference to the idea that different speakers would judge a single example of a polysemous word to exemplify different senses.

1.3 Methodological practise in studying polysemy

This article seeks to develop a robust methodology for investigating individual differences in polysemy. This section reviews some existing approaches to analysing polysemous words and their senses.

1.3.1 Introspection

Introspection, which Talmy (2007) defines as "conscious attention directed by a language user to particular aspects of language as manifest in her own cognition" has a rich methodological tradition within and beyond cognitive linguistics. Of its role in studying word meaning, Talmy argues that "introspection has the advantage over other methodologies in seemingly being the only one able to access [meaning] directly." (2007: xiii). Some of the canonical literature in polysemy (e.g., Brugman 1981; Brugman and Lakoff 1988; Tyler and Evans 2001) propose senses of *over*, and describe how those senses relate to each other, on the basis of introspection. Introspection certainly has a place in the linguists' toolkit and indeed the technological and methodological constraints at play during early cognitive linguistic research on polysemy may themselves explain why introspection was the methodology of choice. However, when description of a linguistic phenomenon rests solely on the outcome of introspection by a handful of experts at most, we should use caution in speculating about their cognitive reality. This caution is justified; research on expert introspection in the description of linguistic phenomena has found the resulting analyses wanting (Bradac et al. 1980; Dąbrowska 2010; Gibbs 2006; Gordon and Hendrick 1997; Labov 1972; Miller 1962; Ross 1979; Schütze 1996; Schwarz-Friesel 2012; Spencer 1973; Sprouse and Schütze 2017). Moreover, semanticists are not immune to semantic satiation (James 1962); indeed, it seems fair to wonder whether a close focus on a particular word makes them more prone to it.

1.3.2 Sorting tasks

Sentence-sorting tasks are an obvious choice for studying polysemy. Cognitive linguists take polysemy as an example of linguistic categorisation, so it seems only natural that the sorting tasks used in cognitive psychological studies of categorisation are used to understand categories in language. Work by Rice and colleagues used sentence-sorting tasks to investigate the polysemous prepositions *at*, *on* and *in*. In each case, every participant sorted examples of each word; 20 examples for each word in Sandra and Rice (1995), and 50 each in Rice (1996) and Rice et al. (1999). Baker (1999) recruited participants to sort up to 244 examples of *see* in an open-sort task. Beyond polysemy, Divjak and Gries (2008) used a smaller but progressive task in their study of near-synonymous Russian verbs; in that case, each participant sorted nine sentences, representing examples of different near-synonymous verbs, each time according to more specific instructions.

Sentence-sorting tasks reveal the categories people make when they are asked to sort stimuli according to a criterion, using pre-set categories (in closed-sort tasks), or using their own categories (in open-sort tasks). In this study, participants

sorted sentences based on the meaning of a target word: either *over*, *under*, *above*, or *below*. This study assumed that the categories people make are indicators of (some of) the senses they find meaningful.

1.3.3 Statistical analysis

Sorting task data is straightforward to interpret using agreement statistics and data visualisation techniques. Baker (1999) introduced Morey and Agresti's adjusted Rand to the study of polysemy, but it has been absent from linguistics since. It is similar to Kappa, κ (Cohen 1960); both are magnitude statistics of pairwise agreement, do not offer significance values, and have the same range of values, ranging from –1.0 (total disagreement) through 0 (chance agreement) to 1.0 (perfect agreement). Unlike κ, the adjusted Rand is used in open-sort tasks. It does not require subjects to use the same number of groups.

Ironically, what constitutes an acceptable level of agreement is the subject of disagreement. Research has not studied acceptable levels of agreement using Morey and Agresti's adjusted Rand but focuses on κ; due to their similarity, conclusions about acceptability of κ values are extended here to the adjusted Rand. Artstein and Poesio describe the interpretation of agreement scores as "little more than a black art" (2008: 576). They note that agreement scores in computational linguistics research tend to follow the interpretation conventions adopted in content analysis, in which values of 0.8 or higher constitute good agreement, and in which tentative conclusions may be drawn from values between 0.67 and 0.8. This scale continues to be used, and indeed generalised to other agreement models such as Sklar's Omega, a Gaussian copula-based framework (Hughes 2018). However, they observe that other authors propose more stringent interpretations; Neuendorf (2002: 3) recommends considering values of 0.9 or more as acceptable in all situations, 0.8 to 0.9 as acceptable in most situations, and values less than 0.8 to constitute great disagreement. Following Neuendorf, the threshold for acceptable agreement in this article is 0.8.

1.3.4 Network analysis

Networks have an established place in the study of semantic knowledge and word meaning (see, for example, Brugman 1981; Collins and Quillian 1969; Hollan 1975; Lindner 1981; Quillian 1969). Brugman's (1981) analysis of *over* has inspired a rich body of work exploring the notion that senses of polysemous words can be represented as networks (Brugman and Lakoff 1988; Li and Joanisse 2021; Rice 2003). Tyler and Evans (2001) proposed a 'principled' means of identifying a central sense, delineating senses, and establishing relationships between senses.

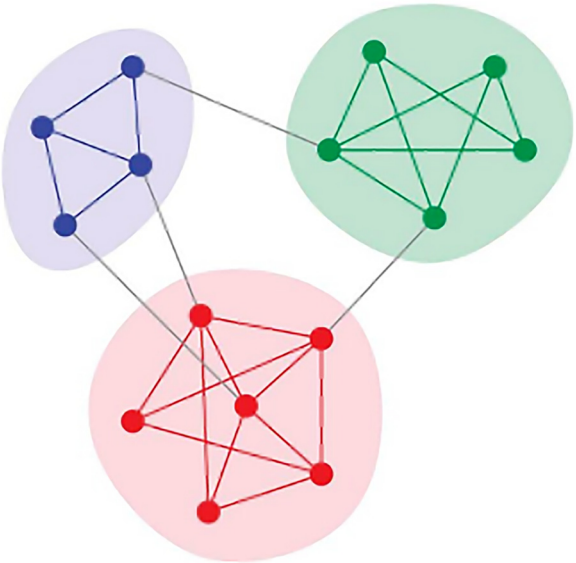

Figure 1: Nodes, edges and communities in a network (Newman 2012).

Describing a polysemous word as a network is reasonable if we assume that its senses are related. However, these proposals are driven by introspective analyses by the authors. They therefore lack objectivity and are absent of any tests of representativity amongst native speakers. Nonetheless, technology exists to explore whether networks do have a role in the study of polysemy.

Network theory is conventionally associated with the study of complex systems and has been used, for example, to study biological processes and food webs (Newman 2012). Networks, illustrated in Figure 1, consist of nodes connected by edges. For example, in a food web, nodes represent species in an ecosystem, while edges represent predator–prey relations. Network algorithms can also produce communities, "groups of nodes with dense intra-community edges and sparse inter-community connections" (Kauffman et al. 2014); in the network in Figure 1, nodes are organised into three communities. Communities provide further detail about the structure of the network, and can reveal unknown or unexpected modules (Blondel et al. 2008). In this study, nodes represent the stimulus sentences and edges represent a participant's decision to categorise them into the same group. Good et al. (2010: 1) note that communities within a network might provide a "principled way to reduce or coarse-grain a system by dividing global heterogeneity into relatively homogenous substructures." Following this, communities are understood to capture similar sentences, and might be understood as senses.

Modularity values are a means of detecting and characterising community structure (Newman 2006). In this study, a modularity value quantifies how well the network captures all participants' sorting decisions and produces distinct communities (a difficult task if participants disagree with each other). Values range from –1 to 1. A "good" set of communities has a modularity value closer to 1 and will feature groups that have more internal connections than would be expected at random. A "bad" set of communities has a modularity value closer to zero, reflecting no more connections between community members than expected by chance (Good et al. 2010). Networks with a modularity value of less than zero feature fewer connections between nodes than would be expected by chance. However, Good et al. (2010) and Levallois (2013) urge caution when interpreting modularity values and suggest taking them lightly. A means of testing the utility of modularity values is to compare them with established measures. In this study, agreement between participants over how stimuli should be categorised is measured statistically, using Morey and Agresti's adjusted Rand. This study thus provides the opportunity to understand whether there is a relationship between statistical agreement calculated for each task, and the modularity value of each associated network. A significant relationship would both inform how modularity values can be interpreted and indicate a role for network visualisations for studying polysemy.

1.3.5 A new method for studying polysemy?

This section has described the precedents for using sentence-sorting tasks in polysemy research, using statistical models of agreement to interpret sorting data, and invoking networks as a means of describing the organisation and representation of polysemous words and their senses. Until now, these tools have operated in isolation (excepting Baker 1999, who combines sorting tasks with the adjusted Rand). Cognitive linguists increasingly call for empirical approaches to studying linguistic phenomena (Dąbrowska 2016). The seminal polysemy networks produced by Brugman (1981), Brugman and Lakoff (1988), and Tyler and Evans (2001), methodologically constrained as they were, cannot meet this imperative. Network theory proper, though, proposes that networks emerge from data representing connections between entities, so as long as we have the data, we needn't throw the baby out with the bathwater. Networks might have a role to play in cognitive linguistic models of polysemy, and there is an opportunity to use data gathered using sorting tasks to establish whether network visualisations *are* useful tools.

1.4 Summary and objectives

Evidence from the cognitive linguistic literature indicates that at least some aspects of language vary across individuals. To date this prospect has not received consideration in the discipline's work on polysemy. Individual differences, conversely, are widely assumed, and indeed observed, in NLP research. However, NLP research focuses on the practical issues that such differences pose to developing disambiguation algorithms, rather than their theoretical implications. This article aims to build on NLP work by examining individual differences in polysemy from a theoretical perspective; specifically, from a cognitive linguistic perspective. Its theoretical aim is sharply focused on the question of whether individuals vary in the senses of a polysemous word that they find meaningful, i.e., whether there are individual differences in word senses. This article also seeks to make a practical contribution. It examines whether three tools typically used in isolation in the study of polysemy can be drawn together to form a robust methodology to answer this question.

2 Methodology

This section describes a novel methodology for studying word senses, combining sentence-sorting tasks with statistical and network analyses.

2.1 The task

Sentence-sorting tasks, described in Section 1.3.2, can be administered as 'open-sort tasks', in which participants organise sentences into groups of their own making. A 'closed-sort task' requires participants to categorise sentences into predetermined categories. Open-sort tasks are used here.

2.2 Participants

Two hundred and five native English speakers aged 18 or over completed the experiment in full. Participants were recruited using Reddit Sample Size; the sample reflects the website's demographics, and the majority (84%) of participants was born in North America. Level of educational attainment varied from incomplete secondary school-level education to doctoral qualification. No incentive was given.

2.3 Stimuli

Stimuli consisted of 12 sets of 36 examples of one of *over, under, above,* or *below*. Sentences were extracted from the Internet and the British National Corpus, edited to make the examples well-formed sentences. The target word was printed in upper case, the remaining text in sentence case. Where necessary, sentences were edited to reduce their length; this maximised the number of sentences that could be displayed simultaneously. Three sets of stimuli were used for each of the four target words: a set of 36 spatial uses of the target word; a set of 36 non-spatial uses; and a mixed set of an equal number of spatial and non-spatial uses taken from the spatial and non-spatial sets. These coarse distinctions were made by the author, informed by small-scale preliminary work (Ramsey 2016). Participants were randomly assigned to one of these 12 tasks.

2.4 Procedure

Participants completed the task online using Optimal Sort, a web-based card sorting platform, using a laptop or desktop computer. Before starting the task, information about the task and how responses would be stored was presented on the screen. Participants were required to read and confirm that they understood this and provide their informed consent to participate. Participants were then shown written instructions about how to complete the task. Briefly, they were instructed to sort a set of sentences into one or more groups, based on what the capitalised target word meant in each sentence. It was emphasised that the goal of the task was to sort all sentences into groups in which the meaning of the capitalised target word was the same in each member of the group. The instructions disappeared when the participant moved the first sentence but could be recalled at any time.

The task was then revealed on the screen. Sentences were presented in random order in a column on the left side of the screen, with a large sorting pane to the centre and right of the screen. Participants were advised to read all sentences, considering carefully what the target word meant in each. They were then required to move each sentence into the sorting pane to create a group, after which further sentences could be dragged and dropped into each group. Sentences could be moved in and out of groups until the participant was satisfied with their sorting decisions. Participants were required to label each group in a way that captured the meaning of the target word in all member sentences. Participants were required to sort all stimuli before they could submit their responses.

2.5 Statistical analysis

Data were analysed for inter-participant agreement in R (R Core Team 2013) using Morey and Agresti's adjusted Rand (Morey and Agresti 1984) in the clues package (Chang et al. 2010).

2.6 Network analysis

Networks were produced in Gephi (Bastian et al. 2009) using the Force Atlas algorithm; one network was produced for each task. Two types of data were input. Node data corresponded to the 36 sentences participants sorted. Edge data was the proportion of participants that sorted each possible pair of sentences into the same group.

2.7 Replication

The study aimed to add to cognitive linguistic methodology by presenting a new tool for studying word senses, comprising a web-based sorting task analysed using an under-used statistical model and a network visualisation algorithm. To test the rigour of this new methodology, the study was replicated after two months with the same participants. Participants were not informed that the two tasks they would complete were identical.

3 Results

The remainder of this article presents the findings of the study and explores their theoretical and methodological implications. Each of the tasks to which a participant could be assigned, and each testing phase, are named using the formula [*word*] [type of stimuli] [testing phase]. For example, participants who sorted examples of *over* that reflected a combination of spatial AND non-spatial senses first completed *over* mixed T1, and two months later completed *over* mixed T2.

Participants' data was quality checked against two criteria. Data that fell foul of either or both criteria was discarded.

1. The logic of sentence-sorting decisions is discernible.[3]

3 A study of individual differences in word senses inherently allows for participants' sorting decisions to differ from my own. However, in cases where it was impossible to understand the participants' sorting rationale, their data was discarded.

2. All free-text entries (in pre-task demographic questions, and group labels) are coherent.

Before I present the analyses, I provide in Table 1 a sample of the data on which these analyses were performed, to enable a visual understanding of what the following, more abstract, analyses are based on. The table shows the groups that participants BMi15 and BMi16 sorted the stimuli into in the *below* mixed T1 task.

Table 1: Groups made by participants BMi15 and BMi16 in below mixed T1.

	BMi15	**BMi16**
BELOW the crags a well-built tunnel could be seen.	Lower down, but not directly underneath	under – physical
BELOW the front windows the extension was divided into two sections.	Directly lower down on some physical vertical surface	under – physical
Congress is somewhere BELOW cockroaches and traffic jams in Americans' esteem.	Lower (metaphorical) ranking	under – metaphorical
Don't paint BELOW the windowsill.	Directly lower down on some physical vertical surface	under – physical
Fill in BELOW all the tasks that you do in a typical day.	Further down on written instructions	next
Give us your fun verdict by dialling the numbers BELOW.	Further down on written instructions	next
He is an unenthusiastic and BELOW average soldier.	Lower (metaphorical) ranking	under – metaphorical
He performed BELOW par last time.	Lower (metaphorical) ranking	under – metaphorical
He set a price BELOW the existing supplier's 1994 prices.	Lower (metaphorical) prices	under – metaphorical
I called out to the people on the beach BELOW, but they didn't hear me.	Lower down, but not directly underneath	beyond
I pinned my name badge BELOW the logo on my tshirt.	Directly lower down on some physical vertical surface	under – physical
In the situations listed BELOW identify what your information needs would be.	Further down on written instructions	next
Instead of being up high the box was down BELOW.	Unclear	under – physical
It will be argued BELOW that economic reconstruction was a success.	Further down on written instructions	next
Look at the sentence BELOW, what does it say?	Further down on written instructions	next
Paul had performed BELOW expectation.	Lower (metaphorical) ranking	under – metaphorical

Table 1: (continued)

	BMi15	BMi16
Serve with Sharp sauce (see BELOW).	Further down on written instructions	next
She had a mole just BELOW her right eye.	Directly lower down on some physical vertical surface	under – physical
Tabith stood BELOW, watching him.	Lower down, but not directly underneath	under – physical
The campus was shrinking BELOW me into a collection of children's play houses.	Lower down, but not directly underneath	under – physical
The crocodile sank BELOW the surface.	Lower down through some opaque medium	under – physical
The loss is a little BELOW £3,200.	Lower (metaphorical) prices	under – metaphorical
The people in the flat BELOW wouldn't stop shouting.	Lower down through some opaque medium	under – physical
The sales value was well BELOW target.	Lower (metaphorical) prices	under – metaphorical
The sleeves gradually get tighter and end BELOW the elbow.	Beyond, on a dimension which is implicitly downwards	beyond
The walk provides wonderful views of Mallerstang BELOW.	Lower down, but not directly underneath	beyond
There are two iron rings on the wall BELOW the painting.	Directly lower down on some physical vertical surface	under – physical
There is a £20 surcharge on orders BELOW £50.	Lower (metaphorical) prices	under – metaphorical
There's no level BELOW which the wages may not fall.	Lower (metaphorical) prices	under – metaphorical
They established an iron foundry in the valley BELOW the church in 1790.	Lower down, but not directly underneath	beyond
We brought in forty million pounds BELOW the target amount.	Lower (metaphorical) prices	under – metaphorical
We dredged BELOW the mud at the bottom of the river.	Lower down through some opaque medium	under – physical
We must set standards of achievement BELOW which they must not fall.	Lower (metaphorical) ranking	under – metaphorical
When we got BELOW the next layer the concentrations became stronger.	Lower down through some opaque medium	under – physical
You wouldn't be doing the job if you were BELOW that level.	Lower (metaphorical) ranking	under – metaphorical
Your mates are down BELOW, watching you.	Lower down, but not directly underneath	under – physical

Statistical analysis: inter-participant agreement.

Table 2: Average inter-participant agreement, and network modularity.

	T1			T2		
	Mean	**SD**	**Modularity**	**Mean**	**SD**	**Modularity**
Above non-spatial	0.61	0.16	0.51	0.63	0.13	0.54
Above mixed	0.59	0.21	0.43	0.65	0.17	0.48
Above spatial	0.62	0.13	0.43	0.67	0.13	0.46
Below non-spatial	0.46	0.17	0.28	0.68	0.15	0.48
Below mixed	0.49	0.21	0.39	0.45	0.19	0.36
Below spatial	0.52	0.13	0.42	0.52	0.13	0.45
Over non-spatial	0.74	0.16	0.64	0.77	0.13	0.67
Over mixed	0.79	0.11	0.66	0.74	0.12	0.64
Over spatial	0.50	0.20	0.38	0.45	0.22	0.32
Under non-spatial	0.46	0.16	0.33	0.47	0.15	0.37
Under mixed	0.63	0.15	0.46	0.65	0.15	0.43
Under spatial	0.25	0.16	0.14	0.25	0.16	0.16

Mean pairwise agreement, summarised in Table 2, varied considerably across tasks, from 0.25 (*under* spatial T1 and T2) to 0.79 (*over* mixed T1). For reference, agreement between the above participants, BMi15 and BMi16, is 0.49.

Neuendorf (2002) proposes that agreement below 0.8 reflects great disagreement. Using this threshold, the data indicate that, on the whole, participants do not reach consensus with each other about how the sentences should be sorted. Taking the categories participants create as the senses they find meaningful, this indicates that there may be individual differences in word senses.

Further, standard deviations reveal differences in *how* varied pairwise agreement is across the tasks. In other words, the degree to which pairs of participants agree about how the sentences should be sorted appears to vary according to the target word, and the broad sense type of this word (spatial, non-spatial, or mixed). Comparing *below* mixed (SD = 0.21) and *below* spatial (SD = 0.13) at T1, for example, the distribution of agreement values amongst participants who sorted a combination of spatial and non-spatial examples of *below* is more volatile than those who only sorted spatial examples. Across the target words, there is greatest variation in pairwise agreement in *over* spatial and *under* spatial, and *above* mixed and *below* mixed.

3.1 Network analysis

In this research, a network's modularity value quantifies how well the network captures all participants' sorting decisions and produces distinct communities. Networks that represent every participant's sorting decisions perfectly would have

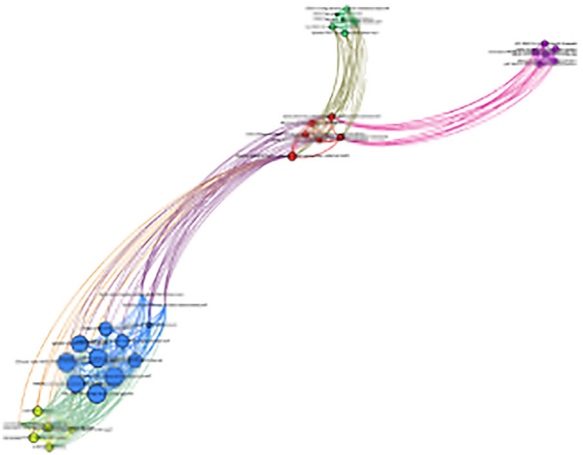

Figure 2: Network visualisation of Over non-spatial T2 (modularity: 0.67).

a modularity value of 1; those that correspond less well to all participants' decisions will have a lower modularity value. Visually, these manifest as networks with discrete communities, and as increasingly undifferentiated communities, respectively. Figures 2 and 3 indicate the structural contrast between networks with the highest and lowest modularity values observed in this study. Figure 2 shows a network with five communities of varying degrees of delineation. The two communities in Figure 3 are broad and poorly differentiated.[4] In this study, we might expect that networks with low modularity values will be generated using data that records low inter-participant agreement.

As shown in Table 2, networks produced using this data varied in their modularity, ranging from 0.14 for *Under* spatial T1, to 0.67 for *Over* non-spatial T2. Pearson's *r* was used to investigate whether there was a relationship between inter-participant agreement recorded for each task, and the modularity value of its associated network. This produced a strong and significant correlation ($r = 0.953$, $p = 0.000$).

3.2 Replication

The study sought to test the rigour of this new methodology and was replicated after two months. The approach taken here was to establish whether there are

4 Network visualisations are provided as Supplementary Materials. Due to their size, they are best viewed on a monitor.

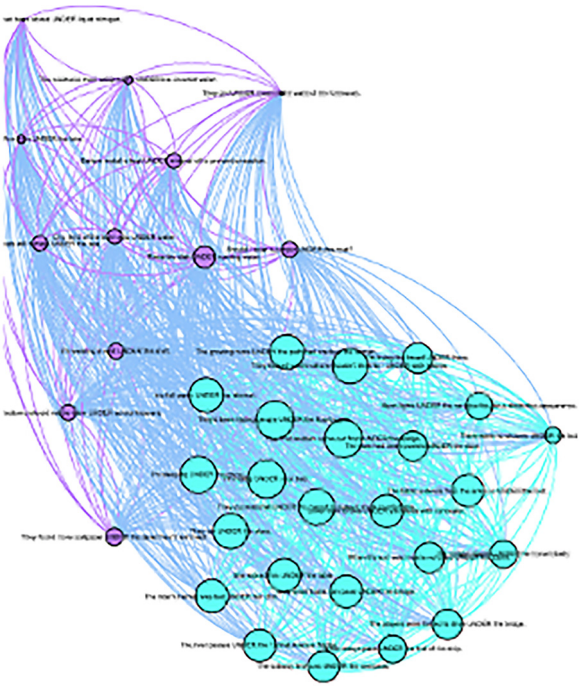

Figure 3: Network visualisation of Under spatial T1 (modularity: 0.14).

significant and meaningful differences in the levels of agreement between participants at T1 and T2, and whether the networks differed in their modularity at T1 and T2. Such differences, if observed, would undermine the rigour of the methodology. A paired samples t-test measured whether the degree to which each participant agreed with every other participant in their task differed significantly between testing phases. Cohen's *d* measured the effect size of any such difference, i.e., whether the difference was meaningful. Effect size was interpreted following Cohen (1988), who proposed that *d* values of 0.8, 0.5, and 0.2 represented large, medium and small effect sizes respectively. Mean agreement was 0.56 (SD = 0.21) at T1, and 0.60 (SD = 0.21) at T2. While the t-test indicated a significant difference in agreement between T1 and T2, the effect size indicates that the difference is trivial: $t(1759) = 4.10$, $p > 0.0001$, $d = 0.19$. This outcome indicates that the degree to which participants (dis)agree about how the sentences should be sorted varies between testing phases, but the difference is minor. Indeed, this outcome can be reasonably expected in research that studies decisions such as these by a group of human participants. Consistency of network modularity was likewise tested with a paired samples t-test. Mean modularity was 0.42 (SD = 0.14) at T1, and 0.45 (SD = 0.14) at

T2. The t-test indicated no significant differences between network modularity at T1 and T2: $t(11) = 1.28$, $p = 0.23$, $d = 0.21$.

4 Discussion

The following section discusses the implications of these findings for cognitive linguistic theory and methodology, before speculating on their practical implications for WSD research.

4.1 Individual differences in word senses

These findings augment the large body of work that has found individual differences in linguistic phenomena. Kidd et al. (2018) argue that all aspects of adult and child linguistic systems are characterised by individual differences. The data gathered here indicate individual differences in word senses, but further research is needed to understand what causal factors determine these differences. If linguistic categories form through exposure and usage, those who work in "language rich" (Kidd et al. 2018) environments may display different sorting decisions. For example, a participant employed as a manual labourer may make different sorting decisions than a participant who is employed as a secretary, due to differences in language exposure across these occupations. Moreover, these differences might manifest at an item level. Uses of *above*, for example, that are predominantly found in written text such as *The process, described ABOVE, is clear to all*, might be sorted differently by these participants due to difference in exposure to print, even if they agree about how other examples should be sorted.

4.2 Successful communication despite individual differences in word senses

Ferreira et al. (2002) offer an interesting explanation of how we succeed in communicating despite individual differences in grammar, proposing that sentence interpretation is based on a "good enough" model of sentence processing, and that successful processing and interpretation is supported by contextual information. We might extend this argument to account for the incongruity between variation in word senses and (generally) successful communication. In the study reported here, participants engaged in a disambiguation experiment rather than in a conversation that required explicit disambiguation. This artificial scenario thus

lacks contextual information that a conversation might provide. In a natural communicative situation, context might provide information and disambiguation or categorisation biases that produce different categorisation decisions than those observed here.

Such an account assumes that successful disambiguation depends on fine-grained senses. Computational linguists have claimed that, on the contrary, disambiguation beyond even the homograph-like level is rarely necessary for human and computer understanding, and that division of these very coarse senses into more fine-grained sub-senses is only done if successful communication depends on it (Ide and Wilks 2007: 66). This may be compatible with Ferreira, Bailey, and Ferraro's argument. Let us consider Ide and Wilks' very coarse-grained senses, which have the potential to consist of more fine-grained senses. It seems feasible that the senses that emerge in the sorting task reported here might, at least by some participants, be collapsed into a smaller number of coarser senses. If we were to pair up two such participants in a conversational scenario, it is possible that these coarser senses might overlap. Let us take as an example the categorisation decisions made by participants BMi15 and BMi16, shown in Table 1, looking specifically at the category to which they assigned the sentence *The sleeves gradually get tighter and end below the elbow*. Participant BMi15 assigned it to a single-member group labelled BEYOND, ON A DIMENSION WHICH IS IMPLICITLY DOWNWARD. Participant BMi16 assigned that sentence to a group labelled BEYOND, along with three other sentences. The categorisation decisions, when adopting a fine-grained approach, therefore differ. However, it might be the case that both participants would agree that this sentence is an example of a broader INFERIOR POSITION sense. If this sentence arose in conversation between BMi15 and BMi16, and if the success of the conversation depends on accessing a fine-grained sense, a mismatch in interpretation might occur. However, if accessing a broad sense allows for "good enough" interpretation, no mismatch will occur, since they both belong to the same broader, collapsed sense.

These ideas, that word senses vary across individuals, and that successful communication might depend on accessing a coarse-grained sense, cast doubt on the idea that a fine-grained exposition of word senses is necessary in a cognitively realistic description of language. The cognitive status of fine-grained distinctions has long been questioned in the field (e.g., Sandra and Rice 1995). It is beyond the scope of this article to establish the level of granularity needed in a realistic description, and while slight differences in meaning might not be necessary for typical communication, they might have a role in other forms of communication such as wordplay. Such an understanding would be of value to linguistic theory, and to resolving the practical constraints that hold back the long-awaited step-change in NLP progress.

4.3 Introspection as a methodology

Individuals appear to classify examples of polysemous words in different ways, despite receiving identical guidelines instructing them to categorise the stimuli only according to the meaning of the target word. This presents a challenge to introspection-based studies of polysemy at both a theoretical level and a practical level; for example, in building "gold standards", databases of annotated corpora used to train and evaluate (semi-) supervised WSD algorithms. Indeed, even though computational linguists have long acknowledged the problem of human annotation (Kilgarriff 1998) contemporary WSD models continue to be trained and evaluated on human-annotated gold standards (e.g., Moreno-Ortiz et al. 2020).[5]

These findings discourage an introspection-based analysis of polysemous words and indicate that scholars should acknowledge that the distinctions that they find meaningful might not be meaningful to other speakers. Consequently, I argue that an introspection-based analysis of the senses of polysemous words might be neither generalisable nor cognitively realistic. I also respond to Talmy's claim that "introspection has the advantage over other methodologies in seemingly being the only one able to access [meaning] directly" (2007: xiii). The fact that participants in this research produced groups of sentences that appear to follow logical and non-random categorisation principles suggests that Talmy might be right. However, his claim, which forms part of his larger argument over the utility of introspection in linguistic research, does not account for the possibility – which has been realised in this study – that when an individual encounters a pair of examples of a polysemous word, the meaning(s) they attribute to them may be different from the meanings attributed by another person. This finding entails that, while introspection may be an important tool for understanding word meaning, meanings identified by introspection are subject to variation across individuals. Consequently, the use of introspection alone as a means of studying word meaning may produce results that do not represent meanings held by other speakers. This finding is compatible with and adds to existing literature that problematises the utility of expert introspection in the description of linguistic – primarily syntactic – phenomena.

---

5 Burgeoning work on automatically annotated corpora may provide one solution to this problem. For example, Pasini and Navigli (2020) have developed Train-O-Matic, "a knowledge-based and language-independent approach that is able to provide millions of training instances annotated automatically with word meanings. The approach is fully automatic, i.e., no human intervention is required, and the only type of human knowledge used is a task-independent WordNet-like resource."

4.4 The role of network visualisations in understanding polysemy

Levallois (2013) has questioned the utility of modularity values. Comparison of network modularity against an established metric, such as an agreement statistic, enables us to respond to this uncertainty. When based on the sentence-sorting data reported here, modularity values measure how closely the network's communities correspond to all participants' groups. The fact that modularity and agreement correlate is therefore unsurprising; if participants make different sorting decisions, which is quantified by the agreement scores, a network will not be able to accurately represent every participant's groups but will instead produce something akin to an "average" of the groups all participants created. While this outcome is not particularly remarkable, it does indicate that modularity values, in this study at least, are meaningful measures. Indeed, if we take an agreement score of at least 0.8 as acceptable, following Neuendorf (2002), we might also argue that a network whose modularity value is 0.67 or higher might be useful for identifying, for example, the senses of a polysemous word.

When a network generated using sentence-sorting data has low modularity (e.g., *under* spatial T1), it seems counterintuitive that we might use that network to study characteristics of the polysemy of that word. However, networks of both high *and* low modularity may allow us to understand more about how, in the face of individual differences in word senses, communication nonetheless proceeds successfully. Since the membership of communities in the networks presented here are tentatively taken to represent word senses, networks may be useful for identifying coarse-grained senses that are used for "good enough" processing, regardless of their modularity. In high modularity networks, proximity between groups of well-defined communities might indicate that they could be collapsed into a coarse-grained sense. In low modularity networks, poorly differentiated communities might themselves reflect coarse-grained senses.

4.5 A new methodology for studying polysemy

The present study introduces a new methodology for studying the nature of polysemous words and their senses. It brings together sentence-sorting tasks, a little-known statistical model to measure inter-participant agreement in scenarios where participants create their own categories, and network visualisations. The data and analyses produced using this methodology are singly and collectively coherent. Inspection of individual participants' sorting decisions, undertaken as a

quality check, strongly suggests that they took a systematic and decisive approach to organising the stimuli. Finally, network visualisations have an established place in the polysemy canon, but those posited to date result from introspective analysis. Network visualisations of the data reported here varied in their coherence, as measured by their modularity value. Crucially, though, this variance corresponded to the degree of inter-participant agreement in the dataset, and the greater the average agreement, the higher the network's modularity. The study was replicated after two months to test the rigour of the methodology. If the methodology is reliable, we should expect to gather consistent results at both T1 and T2. This consistency was tested with paired samples t-tests, performed on pairwise inter-participant agreement and network modularity values. Although there was a significant difference in inter-participant agreement between T1 and T2, the difference was trivial. No significant difference in modularity values was found between T1 and T2.

The method appears to be replicable but requires replication by other scholars with their own data to further test its rigour. However, the findings presented here indicate that it might be a useful methodology for quantitative research on polysemy.

4.6 Automated word sense disambiguation

Automated WSD is a necessary component of a successful artificial intelligence system. Human introspection about word senses is harnessed to train supervised WSD algorithms, and to assess the disambiguation results. Traditionally, experts were the source of this introspection, but over the last fifteen years the discipline has increasingly drawn on crowd-sourced introspection, with or without input from experts. The data shown here problematise the utility not only of introspective analyses by experts, but also the role of naïve annotators' introspection in the development of a word sense inventory. If individuals have different senses of a given word, how useful are their responses in trying to develop a gold standard?

Based on the findings presented here, I propose that fine-grained distinctions may not provide a useful tagging scheme, and that a fine-grained approach to word senses may never be fruitful, no matter how much training the algorithm receives, nor how many human participants (expert or otherwise) contribute to the training set and sense inventory. I would not, however, argue that good automated WSD algorithms based on an inventory compiled by human informants are impossible. Instead, I propose that coarse-grained senses are more useful. I reach this conclusion based on the conflicting realities of both individual differences in word senses, as observed here, and nonetheless effective communication between

individuals. I, therefore, offer an evidence-based argument compatible with Ide and Wilks' (2007) recommendation that computational WSD tasks do not use "the standard fine-grained division of senses" but focus instead on "broad discriminations" (p. 47). This idea has recently been formalised by Lacerra et al. (2020). The authors bring together two resources used widely in WSD research: Roget's Thesaurus, with its 1,075 categories, and WordNet, known for its fine-grained senses. The authors produced an inventory of 'labels' comprising groups of Roget's categories, mapped to WordNet synsets. This inventory, which they named the Coarse Sense Inventory (CSI), resulted in better inter-annotator agreement and better WSD system performance than another coarse sense inventory (BabelDomains) and a fine-grained inventory (WordNet). The resulting inter-annotator agreement levels indicate that human disambiguation is achieved at a coarse level, in line with Ide and Wilks' (2007) argument. However, Lacerra et al.'s (2020) model still relies on human introspection. A solution to this problem might come in the form of a model, such as a network, that collapses fine-grained senses emerging from automatically annotated training corpora, such as Pasini and Navigli's (2020) Train-O-Matic, into increasingly coarse ones. Where a fine-grained sense tag is consistently inaccurate, such a model provides an alternative, coarser-level sense; alternatively, it might suggest a different fine-grained sense that is subsumed under the same coarse sense.

5 Conclusions

Cognitive linguists have shown an enduring interest in individual differences and polysemy, but not in the prospect of individual differences *in* polysemy. This gap exists even though data gathered using WSD methods suggests that speakers appear to have different senses of polysemous words. This article is intended to address the gap in cognitive linguistic literature using an investigation firmly situated in the field's theoretical and methodological traditions. It asks whether individuals have different senses of a given polysemous word, and whether we can combine three established tools typically used in isolation to build a robust and replicable methodology for understanding polysemy. The methodology developed, drawing together open sentence-sorting tasks, a rarely used statistical model, and network visualisation, appears to be a useful way of exploring phenomena in polysemy. In this case, it indicated that there may be individual differences in word senses. The methodology requires replication, and further work is needed to understand the factors that shape these differences. Nonetheless, the data gathered in this study adds further weight to Kidd et al.'s (2018) suggestion that individual differences pervade the entire linguistic system.

As an example of linguistic categorisation, research on polysemy promises to make a substantial contribution towards our understanding of the relationship between language and cognition. This research had a narrow theoretical aim, to understand whether word senses, like other linguistic phenomena, are subject to individual differences. I argue here that they are. If replication using other data sustains this argument, we can then explore the implications of these individual differences for other unresolved questions about polysemy, such as how senses are related, and what theoretical model of categorisation best describes their representation.

Acknowledgements: I am grateful to Ewa Dąbrowska, Amanda Patten, Dagmar Divjak, Sarah Duffy, John Newman, Laura Janda and an anonymous reviewer for their constructive comments on the research and earlier versions of the manuscript, and to the participants for their involvement.
Research funding: Some of this work was funded by Northumbria University, UK under a University Studentship (2013–2016), http://dx.doi.org/10.13039/100010052.
Data availability statement: The data for the analyses, except for 3 datasets for *over* at T2 which were lost due to a technical error, are available at https://osf.io/6p9xy/. Please note that the data will be permanently removed on 1 January 2023 in accordance with conditions of ethical clearance for this study as imposed by Northumbria University.

References

Artstein, Ron & Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34(4). 555–596.

Baker, Collin. 1999. *Seeing clearly: Frame semantic, psycholinguistic, and cross-linguistic approaches to the semantics of the English verb see*. Berkeley: University of California Dissertation.

Bastian, Mathieu, Sebastien Heymann & Mathieu Jacomy. 2009. Gephi: An open source software for exploring and manipulating networks. In *Third international aaai conference on weblogs and social media*, 361–362. Palo Alto, California: Association for the Advancement of Artificial Intelligence.

Bhardwaj, Vikas, Rebecca J. Passonneau, Ansaf Salleb-Aouissi & Nancy Ide. 2010. Anveshan: A framework for analysis of multiple annotators' labeling behavior. In *Proceedings of the fourth linguistic annotation workshop (LAW IV)*, 47–55. Stroudsburg, Pennsylvania: Association for Computational Linguistics.

Blondel, Vincent D., Guillaume Jean-Loup, Lambiotte Renaud & Lefebvre Etienne. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 10. P10008.

Bradac, James J., Larry W. Martin, Norman D. Elliott & H. Tardy Charles. 1980. On the neglected side of linguistic science: Multivariate studies of sentence judgment. *Linguistics* 18(11–12). 967–995.

Brugman, Claudia. 1981. *The story of over*. Berkeley: University of California MA thesis.

Brugman, Claudia & Lakoff George. 1988. Cognitive topology and lexical networks. In Steven L. Small, Garrison W. Cottrell & Michael K. Tanenhaus (eds.), *Lexical ambiguity resolution: Perspectives from psycholinguistics, neuropsychology, and artificial intelligence*, 477–508. San Mateo, CA: Morgan Kaufmann.

Carston, Robyn. 2021. Polysemy: Pragmatics and sense conventions. *Mind & Language* 36(1). 108–133.

Chang, Fang, Weiliang Qiu, Ruben H. Zamar, Ross Lazarus & Xiaogang Wang. 2010. clues: An R package for nonparametric clustering based on local shrinking. *Journal of Statistical Software* 33. 1–16.

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

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences*. New York, NY: Routledge Academic.

Collins, Allan M. & M. Ross Quillian. 1969. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior* 8(2). 240–247.

Dąbrowska, Ewa. 2016. Cognitive linguistics' seven deadly sins. *Cognitive Linguistics* 27(4). 479–491.

Dąbrowska, Ewa. 2010. Naive v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review* 27(1). 1–23.

Dąbrowska, Ewa. 2012. Different speakers, different grammars: Individual differences in native language attainment. *Linguistic Approaches to Bilingualism* 2(2). 219–253.

Dąbrowska, Ewa. 2018. Experience, aptitude and individual differences in native language ultimate attainment. *Cognition* 178. 222–235.

Diessel, Holger. 2017. Usage-based linguistics. In Harley Heidi, Shigeru Miyagawa & Mark Aronoff (eds.), *Oxford research encyclopedia of linguistics*, 1–26. New York: Oxford University Press.

Divjak, Dagmar, Ewa Dąbrowska & Antti Arppe. 2016. Machine Meets Man: Evaluating the psychological reality of corpus-based probabilistic models. *Cognitive Linguistics* 27(1). 1–33.

Duffy, Sarah. E. & Michele I. Feist. 2014. Individual differences in the interpretation of ambiguous statements about time. *Cognitive Linguistics* 25(1). 29–54.

Divjak, Dagmar & Stephan Th. Gries. 2008. Clusters in the mind? Converging evidence from near-synonymy in Russian. *The Mental Lexicon* 3(2). 188–213.

Duffy, Sarah. E., Michele I. Feist & Steven McCarthy. 2014. Moving through time: The role of personality in three real life contexts. *Cognitive Science* 38(8). 1662–1674.

Dunlop, William L., Alexander Karan, Dulce Wilkinson & Nicole Harake. 2020. Love in the first degree: Individual differences in first-person pronoun use and adult romantic attachment styles. *Social Psychological and Personality Science* 11(2). 254–265.

Ferreira, Fernanda, Karl G. D. Bailey & Vittoria Ferraro. 2002. Good-enough representations in language comprehension. *Current Directions in Psychological Science* 11(1). 11–15.

Gibbs, Raymond W. J. 2006. Introspection and cognitive linguistics: Should we trust our own intuitions? In Francisco José Ruiz de Mendoza Ibáñez (ed.), *Annual Review of cognitive linguistics*, vol. 4, 135–151. Amsterdam: John Benjamins.

Good, Benjamin H., Yves-Alexandre de Montjoye & Aaron Clauset. 2010. The performance of modularity maximization in practical contexts. *Physical Review E* 81. 046106.

Gordon, Peter C. & Hendrick Randall. 1997. Intuitive knowledge of linguistic co-reference. *Cognition* 62(3). 325–370.
Gries, Stephan Th. 2015. Polysemy. In Ewa Dąbrowska & Dagmar Divjak (eds.), *Handbook of cognitive linguistics*, 472–490. Berlin: De Gruyter Mouton.
Hollan, James D. 1975. Features and semantic memory: Set-theoretic or network model? *Psychological Review* 82(2). 154–155.
Hughes, John. 2018. SKLAR's omega: a gaussian copula-based framework for assessing agreement. *arXiv preprint* (1803.02734), In press.
Ide, Nancy & Yorick Wilks. 2007. Making sense about sense. In Eneko Agirre & Philip Edmonds (eds.), *Word sense disambiguation algorithms and applications*, 47–73. Dordrecht: Springer.
James, Leon J. 1962. *Effects of repeated stimulation on cognitive aspects of behavior: Some experiments on the phenomenon of semantic satiation*. Montreal, Canada: McGill University Dissertation.
Kauffman, Julie, Aristotelis Kittas, Laura Bennett & Sophia Tsoka. 2014. DyCoNet: A Gephi plugin for community detection in dynamic complex networks. *PLOS One* 9(7). https://doi.org/10.1371/journal.pone.0101357.
Kidd, Evan, Seamus Donnelly, Morten H. Christiansen. 2018. Individual differences in language acquisition and processing. *Trends in Cognitive Sciences* 22(2). 154–169.
Kilgarriff, Adam. 1998. Gold standard datasets for evaluating word sense disambiguation programs. *Computer Speech & Language* 12(3). 453–472.
Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
Lacerra, Caterina, Michele Bevilacqua, Tommaso Pasini & Roberto Navigli. 2020. CSI: A coarse sense inventory for 85% word sense disambiguation. In *Proceedings of the AAAI conference on artificial intelligence*, 8123–8130. Palo Alto, California: Association for the Advancement of Artificial Intelligence.
Lakoff, George. 1990. The invariance hypothesis: Is abstract reason based on image-schemas? *Cognitive Linguistics* 1(1). 39–74.
Levallois, Clément. 2013. Modularity score. *Gephi Forums*. http://gephi.forumatic.com/viewtopic.php?f=32&t=3090 (accessed 31 May 2016).
Li, Jiangtian & Marc F. Joanisse. 2021. Word senses as clusters of meaning modulations: A computational model of polysemy. *Cognitive Science* 45(4). e12955.
Lindner, Susan J. 1981. *A lexico-semantic analysis of English verb particle constructions with out and up*. San Diego: University of California Dissertation.
LuCiD. 2021. LuCiD: The ESRC international centre for language and communicative development. http://www.lucid.ac.uk/resources/for-researchers/outputs-database/ (accessed 12 February 2021).
McGlone, Matthew S. & Jennifer L. Harding. 1998. Back (or forward?) to the future: The role of perspective in temporal language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24(5). 1211–1223.
Miller, George A. 1962. *Psychology: The science of mental life*. New York: Harper & Row.
Moreno-Ortiz, Antonio, Javier Fernández-Cruz & Chantal Pérez Chantal Hernández. 2020. Design and evaluation of SentiEcon: A fine-grained economic/financial sentiment lexicon from a corpus of business news. In *LREC 2020 - 12th International conference on language resources and evaluation*, 5065–5072. Paris, France: European Language Resource Association.

Morey, Leslie C. & Agresti Alan. 1984. The measurement of classification agreement: An adjustment to the Rand statistic for chance agreement. *Educational and Psychological Measurement* 44. 33–37.

Neuendorf, Kimberly A. 2002. *The content analysis guidebook*. Thousand Oaks, California: Sage Publications Inc.

Newman, Mark E. J. 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America* 103(23). 8577–8582.

Newman, Mark E. J. 2012. Communities, modules and large-scale structure in networks. *Nature Physics* 8(1). 25–31.

Pasini, Tommaso & Roberto Navigli. 2020. Train-O-Matic: Supervised word sense disambiguation with no (manual) effort. *Artificial Intelligence* 279. https://doi.org/10.1016/j.artint.2019.103215.

Passonneau, Rebecca J., Ansaf Salleb-Aouissi, Vikas Bhardwaj & Nancy Ide. 2010. Word sense annotation of polysemous words by multiple annotators. In *Proceedings of the seventh international conference on language resources and evaluation (LREC)*, 3244–3249. Paris, France: European Language Resources Association.

Passonneau, Rebecca J., Ansaf Salleb-Aouissi & Nancy Ide. 2009. Making sense of word sense variation. In *Proceedings of the workshop on semantic evaluations: Recent achievements and future directions*, 2–9. Stroudsburg, Pennsylvania: Association for Computational Linguistics.

Passonneau, Rebecca J., Collin Baker, Christiane Fellbaum & Nancy Ide. 2012a. The MASC word sense sentence corpus. In *Proceedings of the eighth international conference on language resources and evaluation (LREC)*, 3025–3030. Paris, France: European Language Resources Association.

Passonneau, Rebecca J., Vikas Bhardwaj, Ansaf Salleb-Aouissi & Nancy Ide. 2012b. Multiplicity and word sense: Evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation* 46. 219–252.

Quillian, M. Ross. 1969. The teachable language comprehender: A simulation program and theory of language. *Communications of the ACM* 12(8). 459–476.

Ramsey, Rachel E. 2016. *An exemplar-theoretic account of word senses*. Newcastle upon Tyne, UK: Northumbria University.

Rand, William. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66. 846–850.

Rice, Sally. 1996. Prepositional prototypes. In Pütz Martin & René Dirven (eds.), *The construal of space in language and thought*, 135–165. Berlin/New York: Mouton De Gruyter.

Rice, Sally. 2003. Growth of a lexical network: Nine English prepositions in acquisition. In Hubert Cuyckens, René Dirven & John R. Taylor (eds.), *Cognitive approaches to linguistic semantics*, 243–280. Berlin: Mouton De Gruyter.

Rice, Sally, Dominiek Sandra & M. Mia Vanrespaille. 1999. Prepositional semantics and the fragile link between space and time. In Masako K. Hiraga, Chris Sinha & Sherman Wilcox (eds.), *Cultural, psychological and typological issues in cognitive linguistics*, 107–127. Amsterdam: John Benjamins.

Ross, Joh Robert. 1979. Where's English? In Fillmore J. Charles, Kempler Daniel & William S.-Y. Wang (eds.), *Individual differences in language ability and language behaviour*, 127–163. New York: Academic Press.

Sandra, Dominiek & Sally Rice. 1995. Network analyses of prepositional meaning: Mirroring whose mind - the linguist's or the language user's? *Cognitive Linguistics* 6(1). 89–130.

Schütze, Cardson T. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.
Schwarz-Friesel, Monika. 2012. On the status of external evidence in the theories of cognitive linguistics: Compatibility problems or signs of stagnation in the field? Or: Why do some linguists behave like Fodor's input systems? *Language Sciences* 34(6). 656–664.
Skoe, Erika, Lisa Brody & Rachel M. Theodore. 2017. Reading ability reflects individual differences in auditory brainstem function, even into adulthood. *Brain and Language* 164. 25–31.
Snow, Rion, Brendan O'Connor, Dan Jurafsky, Andrew Y. Ng. 2008. Cheap and fast - but is it good? Evaluation non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, 254–263. Stroudsburg, Pennsylvania: Association for Computational Linguistics.
Spencer, Nancy Jane. 1973. Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability. *Journal of Psycholinguistic Research* 2(2). 83–98.
Sprouse, Jon & Carson T. Schütze. 2017. Grammar and the use of data. In Bas Aarts, Jill Bowie & Gergana Popova (eds.), *Oxford handbook of English grammar*. Oxford: Oxford University Press.
Stamenković, Dušan, Nicholas Ichien & KeithJ. Holyoak. 2019. Metaphor comprehension: An individual-differences approach. *Journal of Memory and Language* 105. 108–118.
Street, James A. & Dąbrowska Ewa. 2010. More individual differences in language attainment: How much do adult native speakers of English know about passives and quantifiers? *Lingua* 120(8). 2080–2094.
Talmy, Leonard. 2007. Foreword. In Monica Gonzalez-Marquez, Irene Mittelberg, Michael J. Spivey & Seana Coulson (eds.), *Methods in cognitive linguistics*, xi–xxi. Amsterdam: John Benjamins.
Tanner, Darren, Maria Goldshtein & Benjamin Weissman. 2018. Individual differences in the real-time neural dynamics of language comprehension. *Psychology of Learning and Motivation* 68. 299–335.
Taylor, John R. 2003. *Linguistic categorization*, 3rd edn. Oxford: Oxford University Press.
R Team. 2013. *R: A language and environment for statistical computing*. Vienna, Austria: Foundation for Statistical Computing. Available at: http://www.r-project.org/.
Tyler, Andrea & Vyvyan Evans. 2001. Reconsidering prepositional polysemy networks: The case of over. *Language* 77(4). 724–765.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/cog-2021-0020).