Word Sense Clustering and Clusterability

Diana McCarthy*
University of Cambridge

Marianna Apidianaki**
LIMSI, CNRS, Université Paris-Saclay

Katrin Erk†
University of Texas at Austin

Word sense disambiguation and the related field of automated word sense induction traditionally assume that the occurrences of a lemma can be partitioned into senses. But this seems to be a much easier task for some lemmas than others. Our work builds on recent work that proposes describing word meaning in a graded fashion rather than through a strict partition into senses; in this article we argue that not all lemmas may need the more complex graded analysis, depending on their partitionability. Although there is plenty of evidence from previous studies and from the linguistics literature that there is a spectrum of partitionability of word meanings, this is the first attempt to measure the phenomenon and to couple the machine learning literature on clusterability with word usage data used in computational linguistics.

We propose to operationalize partitionability as clusterability, a measure of how easy the occurrences of a lemma are to cluster. We test two ways of measuring clusterability: (1) existing measures from the machine learning literature that aim to measure the goodness of optimal k-means clusterings, and (2) the idea that if a lemma is more clusterable, two clusterings based on two different "views" of the same data points will be more congruent. The two views that we use are two different sets of manually constructed lexical substitutes for the target lemma, on the one hand monolingual paraphrases, and on the other hand translations. We apply automatic clustering to the manual annotations. We use manual annotations because we want the representations of the instances that we cluster to be as informative and "clean" as possible. We show that when we control for polysemy, our measures of clusterability tend to correlate with partitionability, in particular some of the type-(1) clusterability measures, and that these measures outperform a baseline that relies on the amount of overlap in a soft clustering.

* Department of Theoretical and Applied Linguistics, University of Cambridge, UK. E-mail: diana@dianamccarthy.co.uk.
** LIMSI, CNRS, Université Paris-Saclay, France. E-mail: marianna.apidianaki@limsi.fr.
† Department of Linguistics, University of Texas at Austin, USA. E-mail: katrin.erk@mail.utexas.edu.

Submission received: 13 June 2014; revised version received: 3 August 2015; accepted for publication: 25 January 2016.

doi:10.1162/COLI_a_00247

1. Introduction

In computational linguistics, the field of word sense disambiguation (WSD)—where a computer selects the appropriate sense from an inventory for a word in a given context—has received considerable attention.[1] Initially, most work focused on manually constructed inventories such as WordNet (Fellbaum 1998) but there has subsequently been a great deal of work on the related field of word sense induction (WSI) (Pedersen 2006; Manandhar et al. 2010; Jurgens and Klapaftis 2013) prior to disambiguation. This article concerns the phenomenon of word meaning and current practice in the fields of WSD and WSI.

Computational approaches to determining word meaning in context have traditionally relied on a fixed sense inventory produced by humans or by a WSI system that groups token instances into hard clusters. Either sense inventory can then be applied to tag sentences on the premise that there will be one best-fitting sense for each token instance. However, word meanings do not always take the form of discrete senses but vary on a continuum between clear-cut ambiguity and vagueness (Tuggy 1993). For example, the noun *crane* is a clear-cut case of ambiguity between lifting device and bird, whereas the exact meaning of the noun *thing* can only be retrieved via the context of use rather than via a representation in the mental lexicon of speakers. Cases of polysemy such as the verb *paint*, which can mean painting a picture, decorating a room, or painting a mural on a house, lie somewhere between these two poles. Tuggy highlights the fact that boundaries between these different categories are blurred. Although specific context clearly plays a role (Copestake and Briscoe 1995; Passonneau et al. 2010) some lemmas are inherently much harder to partition than others (Kilgarriff 1998; Cruse 2000). There are recent attempts to address some of these issues by using alternative characterizations of word meaning that do not involve creating a partition of usages into senses (McCarthy and Navigli 2009; Erk, McCarthy, and Gaylord 2013), and by asking WSI systems to produce soft or graded clusterings (Jurgens and Klapaftis 2013) where tokens can belong to a mixture of the clusters. However, these approaches do not overtly consider the location of a lemma on the continuum, but doing so should help in determining an appropriate representation. Whereas the broad senses of the noun *crane* could easily be represented by a hard clustering, this would not make any sense for the noun *thing*; meanwhile, the verb *paint* might benefit from a more graded representation.

In this article, we propose the notion of **partitionability** of a lemma, that is, the ease with which usages can be grouped into senses. We exploit data from annotation studies to explore the partitionability of different lemmas and see where on the ambiguity–vagueness cline a lemma is. This should be useful in helping to determine the appropriate computational representation for a word's meanings—for example, whether a hard clustering will suffice, whether a soft clustering would be more appropriate, or whether a clustering representation does not make sense. To our knowledge, there has been no study on detecting partitionability of word senses.

We operationalize partitionability as **clusterability**, a measure of how much structure there is in the data and therefore how easy it is to cluster (Ackerman and Ben-David 2009a), and test to what extent clusterability can predict partitionability. For deriving a gold estimate of partitionability, we turn to the Usage Similarity (hereafter **Usim**) data set (Erk, McCarthy, and Gaylord 2009), for which annotators have rated the similarity of

1 See McCarthy (2009) for a high level overview, Navigli (2009) for a detailed summary, and Agirre and Edmonds (2006) for further background and discussion.

pairs of instances of a word using a graded scale (an example is given in Section 2.2). We use inter-annotator agreement (IAA) on this data set as an indication of partitionability. Passonneau et al. (2010) demonstrated that IAA is correlated with sense confusability. Because this data set consists of similarity judgments on a scale, rather than annotation with traditional word senses, it gives rise to a second indication of partitionability: We can use the degree to which annotators have used intermediate points on a scale, which indicate that two instances are neither identical in meaning nor completely different, but somewhat related.

We want to know to what extent measures of clusterability of instances can predict the partitionability of a lemma. As our focus in this article is to test the predictive power of clusterability measures in the best possible case, we want the representations of the instances that we cluster to be as informative and "clean" as possible. For this reason, we represent instances through manually annotated translations (Mihalcea, Sinha, and McCarthy 2010) and paraphrases (McCarthy and Navigli 2007). Both translations (Resnik and Yarowsky 2000; Carpuat and Wu 2007; Apidianaki 2008) and monolingual paraphrases (Yuret 2007; Biemann and Nygaard 2010; Apidianaki, Verzeni, and McCarthy 2014) have previously been used as a way of inducing word senses, so they should be well suited for the task. Since the suggestion by Resnik and Yarowsky (1997) to limit WSD to senses lexicalized in other languages, numerous works have exploited translations for semantic analysis. Dyvik (1998) discovers word senses and their relationships through translations in a parallel corpus and Ide, Erjavec, and Tufiş (2002) group the occurrences of words into senses by using translation vectors built from a multilingual corpus. More recent works focus on discovering the relationships between the translations and grouping them into clusters either automatically (Bannard and Callison-Burch 2005; Apidianaki 2009; Bansal, DeNero, and Lin 2012) or manually (Lefever and Hoste 2010). McCarthy (2011) shows that overlap of translations compared to overlap of paraphrases on sentence pairs for a given lemma are correlated with inter-annotator agreement of graded lemma usage similarity judgments (Erk, McCarthy, and Gaylord 2009) but does not attempt to cluster the translation or paraphrase data or examine the findings in terms of clusterability. In this initial study of the clusteribility phenomenon, we represent instances through translation and paraphrase annotations; in the future, we will move to automatically generated instance representations.

There is a small amount of work on clusterability in the area of machine learning theory (Epter, Krishnamoorthy, and Zaki 1999; Zhang 2001; Ostrovsky et al. 2006; Ackerman and Ben-David 2009a), and all existing measures are based on k-means clustering. Two of them (variance ratio and worst pair ratio) test how tight the clusters are and how far different clusters are from each other (Epter, Krishnamoorthy, and Zaki 1999; Zhang 2001), and one (separability) tests how much the value of the objective function changes as the number k of clusters changes (Ostrovsky et al. 2006). We test all three of these **intra-clustering** (hereafter intra-clust) measures of clusterability. In addition, we test the intuition that for a well-clusterable lemma, the clusterings based on two different "views" of the same data points—in our case, a clustering based on monolingual paraphrases and a clustering based on translations—should be similar. For this **inter-clustering** (inter-clust) notion of clusterability, we use a simple graphical method that does not have the requirement of needing a specified number of clusters. We use this same graphical clustering to provide the k for our intra-clust measures because the existing definitions of clusterability from machine learning theory need the number of clusters to be fixed in advance. There are a vast number of clustering algorithms with which we could experiment. The clustering algorithm itself is not being evaluated here. Instead, the hypothesis is that if a data set is more clusterable, then it

should be computationally easier to cluster (Ackerman and Ben-David 2009b) because the structure in the data is more obvious, so any reasonable algorithm should be able to partition the data to reflect that structure. We contrast the performance of the three intra-clust measures and the inter-clust measure with a simplistic baseline that relies on the amount of overlapping items in a soft clustering of the instance data, since such a baseline would be immediately available if one applied soft clustering to all lemmas.

We show that when controlling for polysemy, our indicators of higher clusterability tend to correlate with our two gold standard partitionability estimates. In particular, clusterability tends to correlate positively with higher inter-annotator agreement and negatively with a greater proportion of mid-range judgments on a graded scale of instance similarity. Although all our measures show some positive results, it is the intra-clust measures (particularly two of these) that are most promising.

2. Characterizing Word Meaning

2.1 The Difficulty of Characterizing Word Meaning

There has been an enormous amount of work in the fields of WSD and WSI relying on a fixed inventory of senses and on the assumption of a single best sense for a given instance (for example, see the large body of work described in Navigli [2009]) though doubts have been expressed about this methodology when looking at the linguistic data (Kilgarriff 1998; Hanks 2000; Kilgarriff 2006). One major issue arises from the fact that there is a spectrum of word meaning phenomena (Tuggy 1993) from clear-cut cases of ambiguity where meanings are distinct and separable, to cases where meanings are intertwined (highly interrelated) (Cruse 2000; Kilgarriff 1998), to cases of vagueness at the other extreme where meanings are underspecified. For example, at the ambiguous end of the spectrum are words like *bank* (noun) with the distinct senses of financial institution and side of a river. In such cases, it is relatively straightforward to differentiate corpus examples and come up with clear definitions for a dictionary or other lexical resource.[2] These clearly ambiguous words are commonplace in articles promoting WSD because the ambiguity is evident and the need to resolve it is compelling. On the other end of the spectrum are cases where meaning is unspecified (vague); for example, Tuggy gives the example that *aunt* can be father's sister or mother's sister. There may be no contextual evidence to determine the intended reading and this does not trouble hearers and should not trouble computers (the exact meaning can be left unspecified). Cases of polysemy are somewhere in between. Examples from Tuggy include the noun *set* (a chess set, a set in tennis, a set of dishes, and a set in logic) and the verb *break* (a stick, a law, a horse, water, ranks, a code, and a record), each having many connections between the related senses. Although it is assumed in many cases that one meaning has spawned the other by a metaphorical process (Lakoff 1987)—for example, the mouth of a river from the mouth of a person—the process is not always transparent and neither is the point at which the spawned meaning takes an independent existence.

From the linguistics literature, it seems that the boundaries on this continuum are not clear-cut and tests aimed at distinguishing the different categories are not definitive (Cruse 2000). Meanwhile, in computational linguistics, researchers point to

2 Different etymology can help in determining such homonymous cases where several meanings have coincidentally ended up having the same word form, but there are many cases where etymologically related meanings are just as distinct to speakers (Ide and Wilks 2006).

there being differences in distinguishing meanings with some words being much harder than others (Landes, Leacock, and Randee 1998), resulting in differences in inter-tagger agreement (Passonneau et al. 2010, 2012), issues in manually partitioning the semantic space (Chen and Palmer 2009), and difficulties in making alignments between lexical resources (Palmer, Dang, and Rosenzweig 2000; Eom, Dickinson, and Katz 2012). For example, OntoNotes is a project aimed at producing a sense inventory by iteratively grouping corpus instances into senses and then ensuring that these senses can be reliably distinguished by annotators to give an impressive 90% inter-annotator agreement (Hovy et al. 2006). Although the process is straightforward in many cases, for some lemmas this is not possible even after multiple re-partitionings (Chen and Palmer 2009).

Recent work on graded annotations (Erk, McCarthy, and Gaylord 2009, 2013) and graded word sense induction (Jurgens and Klapaftis 2013) has aimed to allow word sense annotations where it is assumed that more than one sense can apply and where the senses do not have to be equally applicable. In the graded annotation study, the annotators are assigned various tasks including two independent sense labeling tasks where they are given corpus instances of a target lemma and sense definitions (WordNet) and are asked to (1) find the most appropriate sense for the context and (2) assign a score out of 5 as to the applicability of every sense for that lemma. In graded word sense induction (Jurgens and Klapaftis 2013), computer systems and annotators preparing the gold standard have to assign tokens in context to clusters (WordNet senses) but each token is assigned to as many senses as deemed appropriate and with a graded level of applicability on a Likert scale (1–5). This scenario allows for overlapping sense assignments and sense clusters, which is a more natural fit for lemmas with related senses, but inter-annotator agreement is highly variable depending on the lemma, varying between 0.903 and 0.0 on Krippendorff's α (Krippendorff 1980). This concurs with the variation seen in other annotation efforts, such as the MASC word sense corpus (Passonneau et al. 2012). Erk, McCarthy, and Gaylord (2009) demonstrated that annotators produced more categorical decisions (5 - identical vs. 1 - completely different) for some words and more mid-range decisions (4 - very similar, 3 - similar, 2 - mostly different) for others. This is not solely due to granularity. In a later article (Erk, McCarthy, and Gaylord 2013), the authors demonstrated that when coarse-grained inventories are used, there are some words where, unsurprisingly, usages in the same coarse senses tend to have higher similarity than those in different coarse senses, but for some lemmas, the reverse happens. Although graded annotations (Erk, McCarthy, and Gaylord 2009, 2013) and soft clusterings (Jurgens and Klapaftis 2013) allow for representing subtler relationships between senses, not all words necessitate such a complicated framework. This article is aimed at finding metrics that can measure how difficult a word's meanings are to partition.

2.2 Alternative Word Meaning Characterizations

Several groups have proposed alternative characterizations of word meaning that do not rely on a partition of instances into senses. We use three of these approaches in the current article: two to provide instance annotations that we use as the basis for clustering and one to provide a gold standard indication of partitionability. Crucially, these three data sets are all produced by adding annotations to samples taken from the same set of sentences used for the English lexical substitution task (McCarthy and

Table 1
Sentences for *post.n* from LEXSUB.

s#	LEXSUB sentence
701	However, both posts include a one-year hand over period and consequently the elections need to be held one year in advance of the end of their terms.
702	Application Details CLOSING DATE : FRIDAY 2 September 2005 (Applications must be post marked on or before this day - no late applications can be considered.)
703	So I put fence posts all the way around the clearing.
704	And I put a second rail around the posts.
705	Received 2 the other day from the AURA Mansfield to Buller ultra in the post at no charge.
706	26/8/2004 Base Jumping Goodness Filed in : Sport by Reevo | Link to this post | Comments (0) There's nothing quite like spending ten minutes watching base jumpers doing their thing all around the world, check them out, you won't regret it.
707	PRO Centre Manager The Board for this post had taken place and the successful applicant would be in post in November.
708	It's becoming really frustrating and they keep on moving the goal post with regard to what they require as security.
709	A consultants post, with a special interest in Otology at St Georges Hospital was advertised in February.
710	The next morning we arrived at the border post at 7:30.

Navigli 2007), hereafter **LEXSUB**. Ten sentences for the target lemma *post.n*[3] are shown in Table 1, with the corresponding sentence ids (s#) in the LEXSUB data set and the target token underlined.

In LEXSUB, human annotators saw a target lemma in a given sentence context and were asked to provide one or more substitutes for the lemma in that context. There were 10 instances for each lemma, and the lemmas were manually selected by the task organizers. The cross-lingual lexical substitution task (Mihalcea, Sinha, and McCarthy 2010) (**CLLS**) is similar, except that whereas in LEXSUB both the original sentence and the substitutes were in English, CLLS used Spanish substitutes. For both tasks, multiple annotators provided substitutes for each target instance. Table 2 shows the English substitutes from LEXSUB alongside the Spanish substitutes from CLLS for the sentences for *post.n* displayed in Table 1.

In the Usim annotation (Erk, McCarthy, and Gaylord 2009, 2013), annotators saw a pair of sentences at a time that both contained an instance of the same target word. Annotators then provided a graded judgment on a scale of 1–5 of how similar the usage of the target lemma was in the two sentences. Multiple annotators rated each sentence pair. Table 3 shows the average judgments for the *post.n* example between each pair of sentence ids in Table 1.[4]

3 We use *n*, *v*, *a*, *r* suffixes to denote nouns, verbs, adjectives, and adverbs, respectively.
4 There are no judgments for a sentence paired with itself and we do not repeat values where a judgment has appeared already in the table (for example, 702-701, given that we already have 701-702 displayed).

Table 2
Paraphrases and translations for sentences with the lemma *post.n* from the LEXSUB and CLLS data. The same sentences were used to elicit both substitute sets.

s#	LEXSUB	CLLS
701	position 3; job 2; role 1;	puesto 2; cargo 1; posicion 1; anuncio 1;
702	mail 2; postal service 1; date 1; post office 1;	enviado 1; mostrando 1; publicado 1; saliendo 1; anuncio 1; correo 1; marcado por correo 1;
703	pole 3; support 2; stake 1; upright 1;	poste 3; cerco 1; colocando 1; desplegando 1;
704	support 2; pole 2; stake 2; upright 1;	poste 3; cerco 1; tabla 1;
705	mail 4; mail carrier 1;	correo 2; posicion 1; puesto 1; publication 1; anuncio 1;
706	message 2; electronic mail 1; mail 1; announcement 1; electronic bulletin 1;	entrada 4;
707	position 3; the job 2; employment 1; job 1;	puesto3; cargo 2; posicion 1; publicacion 1; oficina 1;
708	support 2; marker 1; target 1; pole 1; boundary 1; upright 1;	poste 3;
709	position 3; job 2; appointment 1; situation 1; role 1;	puesto 3; posicion 2; cargo 1; anuncio 1;
710	crossing 3; station 2; lookout 1; fence 1;	caseta 2; puesto fronterizo 1; poste 1; correo 1; frontera 1; cerco 1; puesto 1; caseta fronteriza 1;

The three data sets overlap in the sentences that they cover: Both Usim and CLLS are drawn from a subset of the data from LEXSUB.[5] The overlap between all three data sets is 45 lemmas each in the context of ten sentences.[6] In this article we only use data from this common subset as it provides us with a gold-standard (Usim) and two different representations of the instances (LEXSUB and CLLS substitutes). The 45 lemmas in this subset include 14 nouns, 14 adjectives, 15 verbs, and 2 adverbs.[7]

In our experiments herein, we use the Usim data as a gold-standard of how difficult to partition usages of a lemma is. We use both LEXSUB and CLLS independently as the basis for intra-clust clusterability experiments. We compare clusterings based on LEXSUB and CLLS for the inter-clust clusterability experiments.

3. Measuring Clusterability

We present two main approaches to estimating clusterability of word usages using the translation and paraphrase data from CLLS and LEXSUB. Firstly, we estimate clusterability using intra-clust measures from machine learning. Secondly, our inter-clust

5 Some sentences in LEXSUB did not have two or more responses and for that reason were omitted from the data.

6 Usim data were collected in two rounds. For the four lemmas where there is both round 1 and round 2 Usim and CLLS data, we use round 2 data only because there are more annotators (8 for round 2 in contrast to 3 for round 1) (Erk, McCarthy, and Gaylord 2013).

7 These are the lemmas used in our experiments: account.n, call.v, charge.v, check.v, clear.v, coach.n, dismiss.v, draw.v, dry.a, execution.n, field.n, figure.n, fire.v, flat.a, fresh.a, function.n, hard.r, heavy.a, hold.v, investigator.n, lead.n, light.a, match.n, new.a, order.v, paper.n, poor.a, post.n, put.v, range.n, raw.a, right.r, ring.n, rude.a, shade.n, shed.v, skip.v, soft.a, solid.a, special.a, stiff.a, strong.a, tap.v, throw.v, work.v.

Table 3
Average Usim judgments for *post.n.*

s#	701	702	703	704	705	706	707	708	709	710
701	-	1.0	1.0	1.0	1.3	1.3	4.7	1.3	4.0	1.0
702	″	-	1.0	1.0	3.0	1.7	1.0	1.0	1.0	1.0
703	″	″	-	5.0	1.0	1.0	1.0	3.0	1.0	2.3
704	″	″	″	-	1.0	1.0	1.0	3.0	1.0	3.3
705	″	″	″	″	-	2.0	1.0	1.0	1.0	1.0

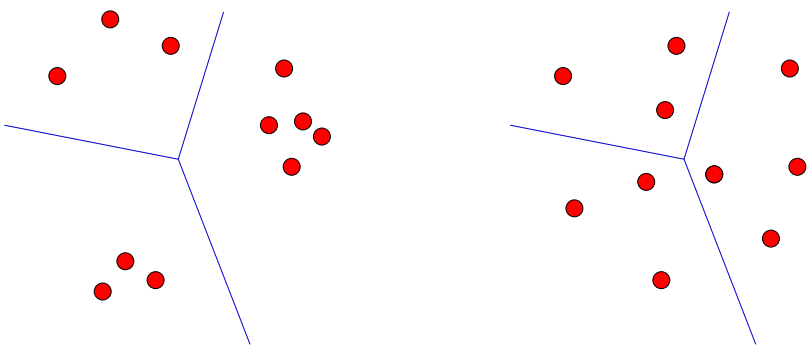
method uses clustering evaluation metrics to compare agreement between two clusterings obtained from CLLS and LEXSUB based on the intuition that less clusterable lemmas will have lower congruence between solutions from the two data sets (which provide different views of the same underlying data).

3.1 Intra-Clustering Clusterability Measures

The notion of the general clusterability of a data set (as opposed to the goodness of any particular clustering) is explored within the field of machine learning by Ackerman and Ben-David (2009a). Consider for example the plots in Figure 1, where the data points on the left should be more clusterable than those on the right because the partitions are easier to make. All the notions of clusterability that Ackerman and Ben-David consider are based on *k*-means and involve optimum clusterings for a fixed *k*.

We consider three measures of clusterability that all assume a *k*-means clustering. Let X be a set of data points, then a *k*-means *k*-clustering of X is a partitioning of X into k sets. We write $C = \{X_1, \ldots, X_k\}$ for a *k*-clustering of X, with $\bigcup_{i=1}^{k} X_i = X$. The *k*-means loss function for a *k*-clustering C is the sum of squared distances of all data points from the centroid of their cluster,

$$\mathscr{L}(C) = \sum_{i=1}^{k} \sum_{x \in X_i} ||x - \text{centroid}(X_i)||^2 \quad (1)$$



(a) More clusterable data (b) Less clusterable data

Figure 1
A more clusterable data set compared with a less clusterable one.

where the centroid or center mass of a set Y of points is

$$\text{centroid}(Y) = \frac{1}{|Y|}\sum_{y \in Y} y \tag{2}$$

A "k-means optimal k-clustering" of the set X is a k-clustering of X that has the minimal k-means loss of all k-clusterings of X. There may be multiple such clusterings.

The first measure of clusterability that we consider is **variance ratio** (VR), introduced by Zhang (2001). Its underlying intuition is that in a good clustering, points should be close to the centroid of their cluster, and clusters should be far apart. For a set Y of points,

$$\sigma^2(Y) = \frac{1}{|Y|}\sum_{y \in Y} ||y - \text{centroid}(Y)||^2 \tag{3}$$

is the variance of Y. For a k-clustering C of X, we write $p_i = \frac{|X_i|}{|X|}$, and define within-cluster variance $W(C)$ and between-cluster variance $B(C)$ of C as follows:

$$\begin{array}{rcl} W(C) & = & \sum_{i=1}^{k} p_i \sigma^2(X_i) \\ B(C) & = & \sum_{i=1}^{k} p_i ||\text{centroid}(X_i) - \text{centroid}(X)||^2 \end{array} \tag{4}$$

Then the variance ratio of the data set X for the number k of clusters is

$$\text{VR}(X,k) = \max_{C \in \mathscr{C}_k} \frac{B(C)}{W(C)} \tag{5}$$

where $\mathscr{C}_k$ is the set of k-means optimal k-clusterings of X. A higher variance ratio indicates better clusterability because variance ratio rises as the distance between clusters increases ($B(C)$) and the distance within clusters decreases ($W(C)$).

Worst pair ratio (WPR) uses a similar intuition as variance ratio, in that it, too, considers a ratio of a within-cluster measure and a between-cluster measure. But it focuses on "worst pairs" (Epter, Krishnamoorthy, and Zaki 1999), the closest pair of points that are in different clusters, and the most distant points that are in the same cluster. For two data points $x, y \in X$ and a k-clustering C of X, we write $x \sim_C y$ if x and y are in the same cluster of C, and $x \not\sim_C y$ otherwise. Then the split of C is the minimum distance of two data points in different clusters, and the width of C is the maximum distance of two data points in the same cluster:

$$\begin{array}{rcl} \text{split}(C) & = & \min_{x,y \in X, x \not\sim_C y} ||x - y|| \\ \text{width}(C) & = & \max_{x,y \in X, x \sim_C y} ||x - y|| \end{array} \tag{6}$$

We use the variant of worst pair ratio given by Ackerman and Ben-David (2009b), as their definition is analogous to variance ratio:

$$\text{WPR}(X,k) = \max_{C \in \mathscr{C}_k} \frac{\text{split}(C)}{\text{width}(C)} \tag{7}$$

where $\mathscr{C}_k$ is the set of k-means optimal k-clusterings of X. Worst pair ratio is similar to variance ratio but can be expected to be more affected by noise in the data, as it only looks at two pairs of data points while variance ratio averages over all data points.

The third clusterability measure that we use is **separability** (SEP), due to Ostrovsky et al. (2006). Its intuition is different from that of variance ratio and worst pair ratio: It measures the improvement in clustering (in terms of the k-means loss function) when we move from $(k-1)$ clusters to k clusters. We write $\text{Opt}_k(X) = \min_{C\ k\text{-clustering of } X} \mathscr{L}(C)$ for the k-means loss of a k-means optimal k-clustering of X. Then a data set X is (k, ε) separable if $\text{Opt}_k(X) \leq \varepsilon\ \text{Opt}_{k-1}(X)$. Separability-based clusterability is defined by

$$\text{SEP}(X, k) = \begin{array}{l}\text{the smallest } \varepsilon \text{ such that} \\ X \text{ is } (k, \varepsilon)\text{-separable}\end{array} \tag{8}$$

Whereas for variance ratio and worst pair ratio higher values indicate better clusterability, the opposite is true for separability: Lower values of separability signal a larger drop in k-means loss when moving from $(k-1)$ to k clusters.[8]

The clusterability measures that we describe here all rely on k-means optimal clusterings, as they were all designed to prove properties of clusterings in the area of clustering theory. To use them to test clusterability of concrete data sets in practice, we use an external measure to determine k (described in Section 4.3), and we approximate k-means optimality by performing many clusterings of the same data set with different random starting points, and using the clustering with minimal k-means loss $\mathscr{L}$.

3.2 Inter-Clustering Clusterability Measures

If the instances of a lemma are highly clusterable, then an instance clustering derived from monolingual paraphrase substitutes and a second clustering of the same instances derived from translation substitutes should be relatively similar. We compare two clustering solutions using the SemEval 2010 WSI task (Manandhar et al. 2010) measures: **V-measure** (V) (Rosenberg and Hirschberg 2007) and **paired F score** (pF) (Artiles, Amigó, and Gonzalo 2009).

V is the harmonic mean of homogeneity and completeness. Homogeneity refers to the degree that each cluster consists of data points primarily belonging to a single gold-standard class, and completeness refers to the degree that each gold-standard class consists of data points primarily assigned to a single cluster. The V measure is noted to depend on both entropy and number of clusters: Systems that provide more clusters do better. For this reason, Manandhar et al. (2010) also used the paired F score (pF), which is the harmonic mean of precision and recall. Precision is the number of common instance pairs between clustering solution and gold-standard classes divided by the number of pairs in the clustering solution, and recall is the same numerator but divided by the total

8 Ackerman and Ben-David (2009b) proposed an additional clusterability measure, center perturbation. However, this measure is not scale invariant, in that its clusterability scores depend on the overall distance between data points in X. As we found this dependency to be very strong, we are not using center perturbation in our experiments in this article.

number of pairs in the gold-standard. *pF* penalizes a difference in number of clusters to the gold-standard in either direction.[9]

4. Experimental Design

In our experiments reported here, we test both intra-clust and inter-clust clusterability measures. All clusterability results are computed on the basis of LEXSUB and CLLS data. The clusterings that we use for the intra-clust measures are *k*-means clusterings. We use *k*-means because this is how these measures have been defined in the machine learning literature; as *k*-means is a widely used clustering, this is not an onerous restriction. The similarity between sentences used by *k*-means is defined in Section 4.2. The *k*-means method needs the number *k* of clusters as input; we determine this number for each lemma by a simple graph-partitioning method that groups all instances that have a minimum number of substitutes in common (Section 4.3). The graph-partitioning method is also used for the inter-clust approach, since it provides the simplest partitioning of the data and determines the number of partitions (clusters) automatically.

In addition to the intra-clust and inter-clust clusterability measures, we test a baseline measure based on degree of overlap in an overlapping clustering (Section 4.4).

We compare the clusterability ratings to two gold standard partitionability estimates, both of which are derived from Usim (Section 4.1). We perform two experiments to measure how well clusterability tracks partitionability (Section 4.6).

4.1 The Gold Standard: Estimating Partitionability from Usim

We turn Usim data into partitionability information in two ways. First, we model partitionability as inter-tagger agreement on Usim (**Uiaa**): Uiaa is the inter-tagger agreement for a given lemma taken as the average pairwise Spearman's correlation between the ranked judgments of the annotators. Second, we model partitionability through the proportion of mid-range judgments over all instances for a lemma and all annotators (**Umid**). We follow McCarthy (2011) in calculating Umid as follows. Mid-range judgments are between 2 and 4, that is not 1 (completely different usages) and not 5 (the same usage). Let $a \in A$ be an annotator from the set A of all annotators, and $j_a \in P_l$ be the judgment of annotator a for a sentence pair for a lemma from all possible such pairings for that lemma (P_l). Then the Umid score for that lemma is calculated as

$$\text{Umid} = \frac{\sum_{a \in A} \sum_{j_a \in P_l} 1 \; if \; j_a \in \{2,3,4\}}{|A| \cdot |P_l|} \tag{9}$$

Umid is a more direct indication of partitionability than Uiaa in that one might have high values of inter-tagger agreement where annotators all agree on mid-range scores. Uiaa is useful as it demonstrates clearly that these measures can indicate "tricky" lemmas that might prove problematic for human annotators and computational linguistic systems.

9 Because both measures (*V* and *pF*) use the harmonic mean, it does not matter whether we use CLLS as the gold standard against LEXSUB or vice versa: The harmonic mean of homogeneity and completeness, or precision and recall, is the same regardless of which clustering solution is considered as "gold."

4.2 Similarity of Sentences Through LEXSUB and CLLS for *k*-Means Clustering

The LEXSUB data for a sentence, for example, an instance of *post.n*, is turned into a vector as follows. Each possible LEXSUB substitute for *post.n* over all its ten instances becomes a dimension. For a given sentence, for example sentence 701 in Table 2, the value for dimension t is the number of times t was named as a substitute for sentence 701. So the vector for sentence 701 has an entry of 3 in the dimension *position*, an entry of 2 in the dimension *job*, and a value of 1 in the dimension *role*, and zero in all other dimensions, and analogously for the other instances. The CLLS data is turned into one vector per instance in the same way. This results in vectors of the same dimensionality for all instances of the same lemma, though the instances of different lemmas can be in different spaces (which does not matter, as they will never be compared). The distance (d_{vec}) between two instances s, s' of the same lemma ℓ is calculated as the Euclidean distance between their vectors. If there are n substitutes overall for ℓ across all its instances, then the distance of s and s' is

$$d_{vec}(s, s') = \sqrt{\sum_{i=1}^{n} (s_i - s'_i)^2} \tag{10}$$

4.3 Graphical Partitioning

This subsection describes the method that we use for determining the number of clusters (k) for a given lemma needed by the intra-clust approach described in Section 3.1, and for providing data partitions for the inter-clust measure of clusterability described in Section 3.2. We adopt a simple graph-based approach to partitioning word usages according to their distance, following Di Marco and Navigli (2013). Traditionally, graph-based WSI algorithms reveal a word's senses by partitioning a co-occurrence graph built from its contexts into vertex sets that group semantically related words (Véronis 2004). In these experiments we build graphs for the LEXSUB and CLLS target lemmas and partition them based on the distance of the instances, reflected in the substitute annotations. Although the graphical approach is straightforward and representative of the sort of WSI methods used in our field, the exact graph partitioning method is not being evaluated here. Other graph partitioning or clustering algorithms could equally be used.

For a given lemma l, we build two undirected graphs using the LEXSUB and CLLS substitutes for l. An instance of l is identified by a sentence id (s#) and is represented by a vertex in the graph. Each instance is associated with a set of substitutes (from either LEXSUB or CLLS) as shown in Table 2 for the noun *post*. Two vertices are linked by an edge if their distance is found to be low enough.

The graph partitioning method that we describe here uses a different and simpler estimate of distance than the k-means clustering. The distance of two vertices is estimated based on the overlap of their substitute sets. As the number of substitutes in each set varies, we use the size of the whole sets along with the size of the intersection for calculating the distance. Let s be an instance (sentence) from a data set (LEXSUB or CLLS) and T be the set of substitute types[10] provided for that

10 We have not used the frequency of each substitute, which is the number of annotators that provided it in LEXSUB or CLLS, though it would be possible to experiment with this in future work.

Table 4
Hard and overlapping partitions (COMPS and CLIQUES) obtained for *post.n* from the LEXSUB data.

Partitions	Sentence ids	Elements
COMPS	706, 705, 702	mail carrier, date, post office, electronic mail, mail, electronic bulletin, message, postal service, announcement
	704, 703, 708	support, target, marker, boundary, stake, pole, upright
	710	lookout, station, fence, crossing
	701, 709, 707	appointment, position, employment, situation, job, the job, role
CLIQUES	705, 702	mail carrier, date, post office, postal service, mail
	705, 706	mail carrier, electronic bulletin, message, electronic mail, announcement, mail
	704, 703, 708	support, target, marker, boundary, stake, pole, upright
	701, 709, 707	appointment, position, employment, situation, job, the job, role
	710	lookout, station, fence, crossing

instance in LEXSUB or CLLS. The distance (d_{node}) between two instances (nodes) s and s' with substitute sets T and T' corresponds to the number of moves necessary to convert T into T'. We use the metric proposed by Goldberg, Hayvanovych, and Magdon-Ismail (2010), which considers the elements that are shared by, and are unique to, each of the sets.

$$d_{node}(T, T') = |T| + |T'| - 2|T \cap T'| \tag{11}$$

We consider two instances as similar enough to be linked by an edge if their intersection is not empty (i.e., they have at least one common substitute) and their distance is below a threshold. After observation of the distance results for different lemmas, the threshold was defined to be equal to 7.[11] A pair of instances with a distance below the threshold is linked by an edge in the graph. For example, instances 705 and 706 of *post.n* are linked in the graph built from the LEXSUB data (cf. Table 2) because their intersection is not empty (they share *mail*) and they have a distance of 5. The graph built for a lemma is partitioned into connected components (hereafter **COMP**). As the COMPS do not share any instances, they correspond to a hard (non-overlapping) clustering solution over the set of instances. Two instances belong to the same component if there is a path between their vertices. The top part of Table 4 displays the COMPS obtained for *post.n* from the LEXSUB data. The 10 instances of the lemma in Table 2 are grouped into four COMPS. Instances 705 and 706 that were linked in the graph are found in the same connected component. On the contrary, 710 shares no substitutes with any other instance as shown in Table 2, and, as a consequence, does not satisfy either the intersection or the distance criterion. Instance 710 is thus isolated as it is linked to no other instances, and forms a separate component.

11 In future work, we intend to explore ways for defining the distance threshold dynamically, on a per lemma basis.

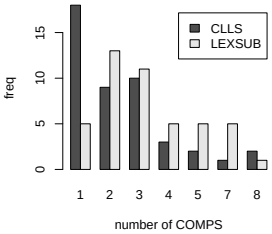


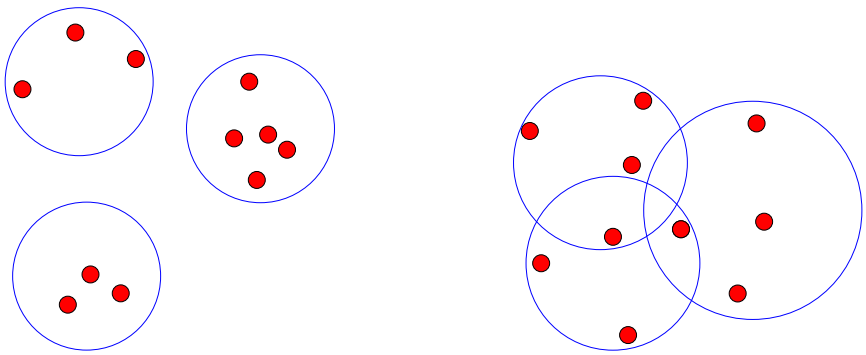
Figure 2
Frequency distribution over number of COMPS: How many lemmas had a given number of COMPS in the two data sets.

Figure 2 shows the frequency distribution of lemmas over number of COMPS.

4.4 A Baseline Measure Based on Cluster Overlap

Our proposed clusterability measures (both intra- and inter-clust) are applicable to hard clusterings. WSI in computational linguistics has traditionally focused on a hard partition of usages into senses but there have been recent attempts to allow for graded annotation (Erk, McCarthy, and Gaylord 2009, 2013) and soft clustering (Jurgens and Klapaftis 2013). We wanted to see how well the extent of overlap between clusters might be used as a measure of clusterability because this information is present for any soft clustering. If this simple criterion worked well, it would avoid the need for an independent measure of clusterability. If the amount of overlap is an indicator of clusterability then soft clustering can be applied and lemmas with clear-cut sense distinctions will be identified as having little or no overlap between clusters, as depicted in Figure 3.

For this baseline, we measure overlap from a second set of node groupings of the graphs described in Section 4.3, where an instance can fall into more than one of the groups. We refer to this soft grouping solution as CLIQUES. A clique consists of a



(a) More clusterable data (b) Less clusterable data

Figure 3
A more clusterable data set compared with a less clusterable one, allowing for cluster overlap.

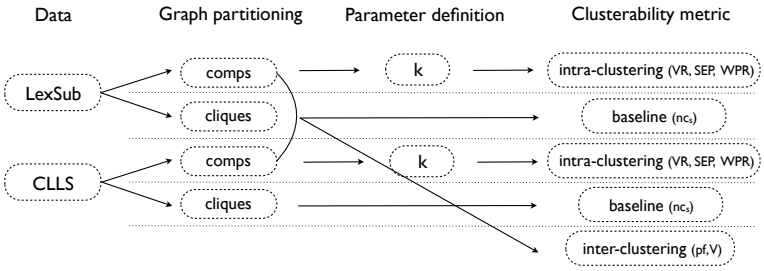


Figure 4
Illustration of the processing pipeline from input data to clusterability estimation.

maximal set of nodes that are pairwise adjacent.[12] They are typically finer grained than the COMPS because there may be vertices in a component that have a path between them without being adjacent.[13]

The lower part of Table 4 contains the CLIQUES obtained for *post.n* in LEXSUB. The two solutions, COMPS and CLIQUES, presented for the lemma in this table are very similar except that there is a further distinction in the CLIQUES as the first cluster in the COMPS is subdivided between two different senses of *mail* (broadly speaking, the `physical` and `electronic` senses). Note that these two CLIQUES overlap and share instance 705.

We wish to see if using the extent of overlap in the CLIQUES reflects the partitionability numbers derived from the Usim data to the same extent as the clusterability metrics already presented. If it does, then the overlapping clustering approach itself could be used to determine how easily the senses partition and clusterability would be reflected by the extent of instance overlap in the clustering solution. Let C_s be the set of partitions (CLIQUES) to which a sentence s from the sentences for a given lemma (S_l) is automatically assigned. Then $nc_s(l)$ measures the average number of CLIQUES to which the sentences for a given lemma are assigned.

$$nc_s(l) = \frac{\sum_{s \in S_l} |C_s|}{|S_l|} \tag{12}$$

We assume that lemmas that are less easy to partition will have higher values of nc_s compared with lemmas with a similar number of clusters over all sentences but with lower values of nc_s.

4.5 Experimental Design Overview

In Figure 4 we give an overview of the whole processing pipeline, from the input data to the clusterability estimation. The graphs built for each lemma from the LEXSUB and CLLS

12 Cliques are computed directly from a graph, not from the COMPS.
13 Note that two different COMPS and CLIQUES can share substitutes (translations or paraphrases). Substitutes serve to determine the distance of the instances. If the distance is high, two instances are not linked in the graph despite their shared substitutes.

Table 5
Overview of gold partitionability estimates and of clusterability measures to be evaluated.

Gold partitionability estimates	Umid: proportion of mid-range (2–4) instance similarity ratings for a lemma
	Uiaa: inter-annotator agreement on the Usim data set (average pairwise Spearman)
Intra-clust clusterability measures	VR, WPR, SEP based on k-means clustering
	k estimated as COMPS
	clustering computed based on either LEXSUB or CLLS substitutes
Inter-clust clusterability measures	comparing COMPS partitioning of CLLS with COMPS partitioning of LEXSUB
	comparison either through V or pF
Baseline	average number nc_s of CLIQUES clusters, computed either from LEXSUB or CLLS data

data are partitioned twice creating COMPS and CLIQUES. The COMPS serve to define the k per lemma needed by the intra-clust clusterability metrics (VR, SEP, WPR). The inter-clust metrics (V and pF) compare the two sets of COMPS created for a lemma from the LEXSUB and CLLS data. The overlaps present in the CLIQUES are exploited by the baseline metric (nc_s).

4.6 Evaluation

Table 5 provides a summary of the two gold standard partitionability estimates and the two types of clusterability measures, along with the baseline clusterability measure that we test. The partitionability estimates and the clusterability measures vary in their directions: In some cases, high values denote high partitionability; in other cases high values indicate low partitionability. Because WPR and VR are predicted to have high values for more clusterable lemmas and SEP has low values, we expect WPR and VR to positively correlate with Uiaa and negatively with Umid and the direction of correlation to be reversed for SEP. Our clustering evaluation metrics (V and pF) should provide correlations with the gold standards in the same direction as WPR and VR since a high congruence between the two solutions for a lemma from different annotations of the same sentences should be indicative of higher clusterability and consequently higher values of Uiaa and lower values of Umid. As regards the baseline approach based on cluster overlap, because we assume that lemmas that are less easy to partition will have higher values of nc_s, high values of nc_s should be positively correlated with Umid and negatively correlated with Uiaa (like SEP). Table 6 gives an overview of the expected directions.

We perform two sets of experiments, which differ in the way in which we control for polysemy. Partitionability estimates as well as clusterability predictions can be expected to be influenced by polysemy. Polysemy has an influence on inter-annotator agreement in that agreement is lower with higher attested polysemy (Passonneau et al. 2010). The number of clusters also influences all our measures of clusterability. Manandhar et al. (2010) note that V and pF are influenced by polysemy. Also, all intra-clust clusterability measures are influenced by k. Variance ratio and worst pair ratio both improve monotonically with k because the distance of points from the center mass of their cluster

Table 6
Directions of partitionability estimates and clusterability measures: ↗ means that high values denote high partitionability, and ↘ means that a high value denotes low partitionability.

Gold partitionability estimates		Clusterability measures	
Umid:	↘	VR:	↗
Uiaa:	↗	WPR:	↗
		SEP:	↘
		V:	↗
		pF:	↗
		nc_s:	↘

decreases as the number of clusters rises (this affects the within-cluster variance $W(C)$ and width(C)). Separability is always lowest for $k = n$ (number of data points), and almost always second-lowest for $k = n - 1$.

The first set of experiments measures correlation using Spearman's ρ between a ranking of partitionability estimates and a ranking of clusterability predictions. We do not perform correlation across all lemmas but control for polysemy by grouping lemmas into polysemy bands, and performing correlations only on lemmas with a polysemy within the bounds of the same band. Let k be the number of clusters for lemma l, which is the number of COMPS for all clusterability metrics other than nc_s, and the number of CLIQUES for nc_s. For the cluster congruence metrics (V and pF), we take the average number of clusters for a lemma in both LEXSUB and CLLS.[14] Then we define three polysemy bands:

- low: $2 \leq k < 4.3$
- mid: $4.3 \leq k < 6.6$
- high: $6.6 \leq k < 9$

Note that none of the intra-clust clusterability measures are applicable for $k = 1$, so in cases where the number of COMPS is one, the lemma is excluded from analysis. In these cases the clustering algorithm itself decides that the instances are not easy to partition.

The second set of experiments performs linear regression to link partitionability to clusterability, using the degree of polysemy k as an additional independent variable. As we expect polysemy to interfere with all clusterability measures, we are interested not so much in polysemy as a separate variable but in the interaction polysemy × clusterability. This lets us test experimentally whether our prediction that polysemy influences clusterability is borne out in the data. As the second set of experiments does not break the lemmas into polysemy bands, we have a single, larger set of data points undergoing analysis, which gives us a stronger basis for assessing significance.

14 Differences in granularity are quite possibly an indication of non-clusterability, but not necessarily. We have also tried using the difference in the number of clusters between CLLS and LEXSUB as an indicator of clusterability but the proposed measures allow a more complete estimation of disparity and so far seem more reliable.

5. Experiments

In this section we provide our main results evaluating the various clusterability measures against our gold-standard estimates. Section 5.1 discusses the evaluation via correlation with Spearman's ρ. In Section 5.2 we present the regression experiments. In Section 5.3 we provide examples and lemma rankings by two of our best performing metrics.

5.1 Correlation of Clusterability Measures Using Spearman's ρ

We calculated Spearman's correlation coefficient (ρ) for both gold standards (Uiaa and Umid) against all clusterability measures: intra-clust (VR, WPR, and SEP), inter-clust (V and pF), and the baseline nc_s. For all these measures except the inter-clust, we calculate ρ using LEXSUB and CLLS separately as our clusterability measure input. The inter-clust measures rely on two views of the data so we use LEXSUB and CLLS together as input. We calculate the correlation for lemmas in the polysemy bands (low, mid, and high, as described above in Section 4.6) subject to the constraint that there are at least five lemmas within the polysemy range for that band. We provide the details of all trials in Appendix A and report the main findings here.

Table 7 shows the average Spearman's ρ over all trials for each clusterability measure. Although there are a few non-significant results from individual trials that are in the unanticipated direction (as discussed in the following paragraph), all average ρ are in the anticipated direction, specified in Table 6; SEP and nc_s are positively correlated with Umid and negatively with Uiaa whereas for all other measures the direction of correlation is reversed. Some of the metrics show a promising level of correlation but the performance of the metrics varies. The baseline nc_s is particularly weak, highlighting that the amount of shared sentences in overlapping clusters is not a strong indication of clusterability. This is important because if this simple baseline had been a good indicator of clusterability, then a sensible approach to the phenomenon of partionability of word meaning would be to simply soft cluster a word's instances and the extent of overlap would be a direct indication that the meanings are highly intertwined. WPR is also quite

Table 7
The macro-averaged correlation of each clusterability metric with the Usim gold-standard rankings Uiaa and Umid: All correlations are in the expected direction. Also, the proportion (prop.) of trials from Tables A.1–A.5 in Appendix A with moderate or stronger correlation in the correct direction with a statistically significant result.

measure type	measure	average ρ Umid	average ρ Uiaa	prop. ρ > 0.4* or ** Umid	prop. ρ > 0.4* or ** Uiaa
intra-clust	VR	−0.483	0.365	2/3	2/3
	SEP	0.569	−0.390	2/3	1/3
	WPR	−0.322	0.210	1/3	0/3
inter-clust	pF	−0.318	0.540	0/2	1/2
	V	−0.123	0.493	0/2	0/2
baseline	nc_s	0.053	−0.164	0/6	1/6

weak, which is not unexpected: It only considers the worst pair rather than all data points, as noted in Section 3.1. Both inter-clust measures (*pF* and *V*) have a stronger correlation with Uiaa than with Umid, whereas for the machine learning measures the reverse is true and the correlation is stronger for Umid. As mentioned in Section 4.1, Umid is a more direct gold-standard indicator of partitionability but Uiaa is useful as a gold standard as it indicates how problematic annotation will be for humans. The machine learning metric SEP and our proposal for *pF* as an indication of clusterability provide the strongest average correlations, though the results for *pF* are less consistent over trials.[15]

Because we are controlling for polysemy, there is less data (lemmas) for each correlation measurement so many individual trials do not give significant results, but all significant correlations are in the anticipated direction. The final two columns of Table 7 show the proportion of cases that are significant at the 0.05 level or above and have $\rho > 0.4$[16] in the anticipated direction out of all individual trials meeting the constraint of five or more lemmas in the respective polysemy band for LEXSUB or CLLS input data. We are limited by the available gold-standard data and need to control for polysemy. So there are several results with a promising ρ which, however, are not significant, such that they are scored negatively in this more stringent summary. Nevertheless, from this summary of the results we can see that the machine learning metrics, particularly VR (which has a higher proportion of successful trials) and SEP (which has the highest average correlations) are most consistent in indicating partitionability using either gold-standard estimate (Umid or Uiaa) with VR achieving 66.7% success (2 out of 3 trials for each gold-standard ranking). WPR is less promising for the reasons stated above. Although there are some successful trials for the inter-clust approaches, the results are not consistent and only one trial showed a (highly) significant correlation. The baseline approach which measures cluster overlap has only one significant result in all 6 trials, but more worrisome for this measure is the fact that in 4 out of the 12 trials (2 for each Umid and Uiaa) the correlation was in the non-anticipated direction. In contrast there was only one result for WPR (on CLLS) in the non-anticipated direction and one result for *V* on the fence ($\rho = 0$) and all other individual results for the inter and intra-clust measures were in the anticipated direction.

There were typically more lemmas in the intra-clust trials with LEXSUB compared to CLLS, as shown in Appendix A due to the fact that many lemmas in CLLS have only one component (see Figure 2) and are therefore excluded from the intra-clust clusterability estimation.[17]

5.2 Linking Partitionability to Clusterability and Polysemy Through Regression

Our first round of experiments revealed some clear differences between approaches and implied good performance, particularly for the intra-clust measures VR and SEP. In the first round of experiments, however, we separated lemmas into polysemy bands and this resulted in the set of lemmas involved in each individual correlation experiment being somewhat small. This makes it hard to obtain significant results. Even for the

15 This can be seen in Table A.3 in Appendix A.

16 This is generally considered the lower bound of moderate correlation for Spearman's and is the level of inter-annotator agreement achieved in other semantics tasks (for example see Mitchell and Lapata [2008]).

17 As noted before, none of the intra-clust measures are applicable for the case of $k = 1$.

overall most successful measures, not all trials came out as significant. In this second round of experiments, we therefore change the set-up in a way that allows us to test on all lemmas in a single experiment, to see which clusterability measures will exhibit an overall significant ability to predict partitionability.

We use linear regression, an analysis closely related to correlation.[18] The dependent variable to be predicted is a partitionability estimate, either Umid or Uiaa. We use two types of independent variables (predictors). The first is the clusterability measure—here we call this variable **clust**. The second is the degree of polysemy, which we call **poly**. This way we can model an influence of polysemy on clusterability as an interaction of variables, and have all lemmas undergo analysis at the same time. This lets us obtain more reliable results: Previously, a non-significant result could indicate either a weak predictor or a data set that was too small after controlling for polysemy, but now the data set undergoing analysis is much bigger.[19] Furthermore, this experiment demonstrates how clusterability and polysemy can be used together as predictors.

The variable clust reflects the clusterability predictions of each measure. We use the actual values, not their rank among the clusterability values for all lemmas. This way we can test the ability of our clusterability measures to predict partitionability for individual lemmas, while the rank is always relative to other lemmas that are being analyzed at the same time. The values of the variable clust are obviously different for each clusterability measure, but the values of poly also vary across clusterability measures: For all intra-clust measures poly is the number of COMPS. For the inter-clust measures, it is the average number of COMPS between the numbers computed from LEXSUB and from CLLS. For the nc_s baseline it is the number of CLIQUES. In all cases, poly is the actual number of COMPS or CLIQUES, not the polysemy band.

We test three different models in our linear regression experiment. The first model has poly as its sole predictor. It tests to what extent partitionability issues can be explained solely by a larger number of COMPS or CLIQUES. Our hypothesis is that this simple model will not suffice. The second model has clust as its sole predictor, ignoring possible influences from polysemy. The third model uses the interaction poly $\times$ clust as a predictor (along with poly and clust as separate variables). Our hypothesis is that this third model should fare particularly well, given the influence of polysemy on clusterability measures that we derived theoretically above.[20]

We evaluate the linear regression models in two ways. The first is the F test. Given a model M predicting Y from predictors $X_1, \ldots, X_m$ as $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m$, it tests the null hypothesis that $\beta_0 = \beta_1 = \ldots = \beta_m = 0$. That is, it tests whether M is statistically indistinguishable from a model with no predictors.[21] Second, we use the Akaike Information Criterion (AIC) to compare models. AIC tests how well a model

18 The regression coefficient is a standardization of Pearson's r, a correlation coefficient, related via a ratio of standard deviations.

19 Also, the first round of experiments had to drop some lemmas from the analysis when they were in a polysemy band with too few members; the second round of experiments does not have this issue.

20 We also tested a model with predictors poly+clust, without interaction. We do not report on results for this model here as it did not yield any interesting results. It was basically always between clust and poly $\times$ clust.

21 We will say an F test "reached significance" to mean that the null hypothesis was rejected for some model.

Table 8
Regression results for the **Umid** partitionability estimate. Significance of F statistic, and AIC for the following models: polysemy only (poly), clusterability only (clust), and interaction (poly × clust). **Bolded**: model that is best by AIC and has significant F, separately for each substitute set. We use * for statistical significance with $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$.

		poly		clust		poly × clust	
data	cl. measure	F	AIC	F	AIC	F	AIC
CLLS	VR	-	-24.9	-	-25.7	**	**-35.1**
CLLS	SEP	-	-24.9	**	-34.2	*	-30.3
CLLS	WPR	-	-24.9	-	-27.5	-	-25.4
CLLS	nc_s	-	-30.6	-	-25.8	-	-26.8
LEXSUB	VR	-	-24.8	*	-31.4	***	**-34.5**
LEXSUB	SEP	-	-24.8	**	-32.6	***	-32.2
LEXSUB	WPR	-	-24.8	***	-34.0	*	-30.2
LEXSUB	nc_s	**	-26.5	**	-27.9	*	-28.2
both	pF	-	-24.8	-	-25.3	-	-21.5
both	V	-	-24.8	*	**-28.0**	-	-25.7

will likely generalize (rather than overfit) by penalizing models with more predictors. AIC uses the log likelihood of the model under the data, corrected for model complexity computed as its number of predictors. Given again a model M predicting Y (in our case, either Umid or Uiaa) from m predictors, the AIC is

$$AIC = -2 \log p(Y|M) + 2m$$

The lower the AIC value, the better the generalization of the model. The model preferred by AIC is the one that minimizes the Kullback-Leibler divergence between the model and the data. AIC allows us to compare all models that model the same data, that is, all models predicting Umid can be compared to each other, and likewise all models predicting Uiaa.

The number of data points in each model depends on the partitioning (as lemmas with $k = 1$ cannot enter into intra-clust clusterability analysis), which differs between CLLS and LEXSUB. AIC depends on the sample size (through $p(Y|M)$), so in order to be able to compare all models that model the same partitionability estimate, we compute AIC only on the subset of lemmas that enters in all analyses.[22] In contrast, we compute the F test on all lemmas where the clusterability measure is valid,[23] in order to use the largest possible set of lemmas to test the viability of a model.[24]

Table 8 shows the results for models predicting Umid, and Table 9 shows the results for the prediction of Uiaa. The bolded figures are the best AIC values for each substitute set (CLLS, LEXSUB, both) where the corresponding F-tests reach significance.[25]

22 This subset comprises 27 lemmas: charge.v, clear.v, draw.v, dry.a, fire.v, flat.a, hard.r, heavy.a, hold.v, lead.n, light.a, match.n, paper.n, post.n, range.n, raw.a, right.r, ring.n, rude.a, shade.n, shed.v, skip.v, soft.a, solid.a, stiff.a, tap.v, throw.v.

23 For the intra-clust measures, this is only lemmas where $k > 1$.

24 We also computed AIC separately for substitute sets LEXSUB, CLLS, and both (for inter-clust). The relative ordering of models within each substitute set remained mostly the same.

25 Log-likelihood values can be positive, as they are in our case, leading to negative AIC values. See, for example, `http://blog.stata.com/2011/02/16/positive-log-likelihood-values-happen/`.

Table 9
Regression results for the **Uiaa** partitionability estimate. Significance of F statistic, and AIC for the following models: polysemy only (poly), clusterability only (clust), and interaction (poly × clust). Bolded: model that is best by AIC and has significant F, separately for each substitute set. We use * for statistical significance with $p < 0.05$ and ** for $p < 0.01$.

		poly		clust		poly × clust	
data	cl. measure	F	AIC	F	AIC	F	AIC
CLLS	VR	-	-20.2	-	-21.2	-	-20.7
CLLS	SEP	-	-20.2	-	-23.0	-	-20.9
CLLS	WPR	-	-20.2	-	-24.1	-	-21.8
CLLS	nc_s	-	-20.8	-	-20.3	-	-19.3
LEXSUB	VR	-	-20.4	-	-21.7	*	-26.9
LEXSUB	SEP	-	-20.4	**	-27.7	*	-25.4
LEXSUB	WPR	-	-20.4	*	**-29.7**	-	-27.0
LEXSUB	nc_s	-	-22.7	**	-21.4	**	-24.8
both	pF	-	-20.0	-	-22.1	-	-18.8
both	V	-	-20.0	-	-24.8	-	-21.9

Confirming the results from our first round of experiments, we obtain the best results for SEP and VR: The best AIC results in predicting Umid are reached by VR, while SEP shows a particularly reliable performance. In predicting Umid, all SEP models that use clust reach significance, and in predicting Uiaa, all SEP models that use clust reach significance if they are based on LEXSUB substitutes. WPR reaches the best AIC values on predicting Uiaa, but on the F test, which takes into account more lemmas, its results are less often significant.

As in the first round of experiments, the performance of the two inter-clust measures is not as strong as that of the intra-clust measures. Here the inter-clust measures are in fact often comparable to the nc_s baseline. However, as CLLS seems to be harder to use as a basis than LEXSUB (we comment on this subsequently), the inter-clust measures may be hampered by problems with the CLLS data.

The baseline nc_s measure does not have as dismal a performance here as it did in the first round of experiments, but its performance is still worse throughout than that of the intra-clust measures. Interestingly, the poly variable that we use for nc_s, which is the absolute number of CLIQUES for a lemma, is informative to some extent for Umid but not for Uiaa, and the clust variable is informative to some extent for Uiaa but not for Umid.

The regression experiments overall confirm the influence of polysemy on the clusterability measures. Although clusterability as a predictor on its own (the clust models) often reaches significance in predicting partitionability, taking polysemy into account (in the poly × clust models) often strengthens the model in predicting Umid and achieves the overall best results (the two bolded models); however for Uiaa the results are more ambivalent, where of the four clusterability measures that produce significant models, two improve when the interaction with polysemy is taken into account, and the two others do not. We also note that COMPS alone (the poly variable for the intra-clust models) never manages to predict partitionability in any way, for either Umid or Uiaa. In contrast, the number of CLIQUES (the poly variable of the nc_s model) emerges as a predictor of Umid, though not of Uiaa.

In comparing Umid versus Uiaa, we see that Umid seems to be generally easier to predict, as it has more models with a significant F test.

Comparing the CLLS and LEXSUB substitutions, we see that the use of LEXSUB leads to much better predictions than CLLS. Most strikingly, in predicting Uiaa no model achieves significance using CLLS. We have commented on this issue before: The reason for this effect is that many lemmas in CLLS have only one component and are therefore excluded from the intra-clust clusterability estimation.

Clusterability in practice. As this round of experiments used the raw clusterability figures to predict partitionability, rather than their rank, it points the way to using clusterability in practice: Given a lemma, collect instance data (for example paraphrases, translations, or vectors). Estimate the number of clusters, for example using a graphical clustering approach. Then use a clusterability measure (SEP or VR recommended) to determine its degree of clusterability, and use a regression classifier to predict a partitionability estimate. It may help to take the interaction of clust and poly into account. If the estimate is high, then a hard clustering is more likely to be appropriate, and sense tagging for training or testing should not be difficult. Where the estimate is low it is more likely that a more complex graded representation is needed, and in extreme cases clustering should be avoided altogether. Determining where the boundaries are would depend on the purpose of the lexical representation and is not addressed in this article. Our contribution is an approach to determine the relative location of lemmas on a continuum of partitionability.

5.3 Lemma Clusterability Rankings and Some Examples

Our clusterability metrics, in particular VR and SEP, are useful for determining the partitionability of lemmas. In this section we show the rankings for these two metrics with our lemmas and provide a couple of more detailed examples with the LEXSUB and CLLS data.

In Table 10 we show the lemmas that have $k > 1$ when partitioned into COMPS using the LEXSUB substitutes, their respective gold standard Umid and Uiaa values, and the SEP and VR values calculated for them on the basis of LEXSUB substitutes. The "L by Uiaa" and "L by Umid" columns display the lemmas reranked according to the two gold-standard estimates, and the "L by VR" and "L by SEP" columns do likewise for the VR and SEP clusterability measures. We have reversed the order of the ranking by Umid and SEP because these measures are high when clusterability is low and vice versa. Lemmas with high partitionability should therefore be near the bottom of the table in columns 7–10 and lemmas with low partitionability should be near the top. There are differences and all rankings are influenced by polysemy, but we can see from this table that on the whole the metrics rank lemmas similarly to the gold-standard rankings with highly clusterable lemmas (such as *fire.v*) at the bottom of the table and less clusterable lemmas (such as *work.v*) nearer the top.

We now take a closer look at two example lemmas, *fire.v* and *solid.a*. Table 11 provides the COMPS from both the LEXSUB and the CLLS data. Both lemmas have a polysemy of 2 according to the COMPS clustering. *fire.v* is an example of a highly clusterable lemma whereas *solid.a* is a less-clusterable lemma. Table 12 shows the values for the clusterability measures. The intra-clust metrics are calculated for both LEXSUB and CLLS independently whereas the inter-clust metrics (*pF* and *V*) compare the two independent clustering solutions with each other. *fire.v* is more clusterable as can be seen by the clusters over the LEXSUB and CLLS data (Table 11), which denote a clear sense distinction, and by the Uiaa and Umid from the Usim gold standard. The measures WPR, VR, *V*, and *pF* are all higher for the more clusterable *fire.v* compared with *solid.a*,

Table 10
Ranking of lemmas (L) by the gold-standards, and by VR and SEP for LEXSUB data.

lemma	k	Umid	Uiaa	VR	SEP	L by Umid	L by Uiaa	L by VR	L by SEP
account.n	2	0.39	0.66	0.67	0.60	raw.a	function.n	throw.v	hold.v
check.v	2	0.52	0.35	0.68	0.59	strong.a	field.n	put.v	strong.a
dismiss.v	2	0.61	0.52	1.26	0.44	special.a	work.v	work.v	flat.a
fire.v	2	0.17	0.93	7.18	0.12	throw.v	raw.a	soft.a	field.n
heavy.a	2	0.60	0.57	0.50	0.67	work.v	strong.a	heavy.a	throw.v
put.v	2	0.62	0.34	0.42	0.70	hard.r	throw.v	hold.v	special.a
right.r	2	0.48	0.65	0.85	0.54	solid.a	put.v	account.n	execution.n
shed.v	2	0.49	0.53	0.71	0.58	put.v	hard.r	check.v	hard.r
skip.v	2	0.38	0.70	0.81	0.55	dismiss.v	check.v	shed.v	function.n
soft.a	2	0.44	0.51	0.48	0.68	heavy.a	special.a	solid.a	ring.n
solid.a	2	0.63	0.49	0.71	0.58	function.n	stiff.a	skip.v	put.v
throw.v	2	0.70	0.32	0.36	0.74	rude.a	shade.n	right.r	clear.v
work.v	2	0.64	0.27	0.46	0.68	draw.v	poor.a	ring.n	match.n
call.v	3	0.18	0.65	3.44	0.54	check.v	hold.v	stiff.a	draw.v
execution.n	3	0.46	0.78	1.33	0.73	stiff.a	lead.n	dismiss.v	lead.n
figure.n	3	0.39	0.50	2.80	0.51	shed.v	solid.a	match.n	work.v
hard.r	3	0.64	0.34	1.28	0.72	lead.n	light.a	hard.r	raw.a
hold.v	3	0.48	0.47	0.64	0.80	right.r	figure.n	execution.n	soft.a
match.n	3	0.33	0.59	1.27	0.69	hold.v	draw.v	paper.n	rude.a
paper.n	3	0.44	0.63	1.84	0.58	field.n	soft.a	rude.a	range.n
poor.a	3	0.34	0.43	2.18	0.65	execution.n	dismiss.v	poor.a	heavy.a
ring.n	3	0.33	0.53	1.03	0.71	tap.v	shed.v	tap.v	light.a
stiff.a	3	0.50	0.40	1.24	0.63	clear.v	ring.n	function.n	dry.a
tap.v	3	0.45	0.70	2.35	0.56	paper.n	heavy.a	flat.a	poor.a
flat.a	4	0.44	0.85	2.48	0.75	soft.a	match.n	range.n	stiff.a
function.n	4	0.53	0.14	2.46	0.71	flat.a	dry.a	strong.a	account.n
post.n	4	0.22	0.69	3.19	0.60	figure.n	rude.a	figure.n	post.n
rude.a	4	0.53	0.61	2.10	0.67	account.n	paper.n	post.n	check.v
shade.n	4	0.30	0.42	4.22	0.48	skip.v	clear.v	special.a	shed.v
charge.v	5	0.24	0.68	7.26	0.57	dry.a	right.r	call.v	solid.a
dry.a	5	0.38	0.59	4.74	0.65	light.a	call.v	raw.a	paper.n
field.n	5	0.47	0.25	3.86	0.74	range.n	account.n	field.n	charge.v
range.n	5	0.34	0.74	2.60	0.67	poor.a	charge.v	shade.n	tap.v
special.a	5	0.70	0.37	3.36	0.73	match.n	post.n	dry.a	skip.v
clear.v	6	0.45	0.63	4.76	0.70	ring.n	skip.v	clear.v	right.r
lead.n	6	0.49	0.47	5.15	0.68	shade.n	tap.v	light.a	call.v
light.a	6	0.36	0.49	4.87	0.66	charge.v	range.n	lead.n	figure.n
raw.a	6	0.73	0.29	3.56	0.68	post.n	execution.n	fire.v	shade.n
strong.a	6	0.73	0.31	2.76	0.76	call.v	flat.a	charge.v	dismiss.v
draw.v	7	0.53	0.50	8.50	0.69	fire.v	fire.v	draw.v	fire.v

Table 11
COMPS obtained from LEXSUB and CLLS for *fire.v* and *solid.a.*

	LEXSUB		CLLS	
	s#	substitutes	s#	substitutes
fire.v	1857, 1852, 1859, 1855, 1851, 1860	discharge, shoot at, launch, shoot	1857, 1852, 1859, 1855, 1851, 1860	balear, lanzar, aparecer, prender fuego, disparar, golpear, apuntar, detonar, abrir fuego
	1858, 1856, 1853, 1854	sack, dismiss, lay off	1858, 1856, 1853, 1854	correr, dejar ir, delar sin trabajo, despedir, desemplear, liquidar, dejar sin trabajo, echar, dejar sin empleo
solid.a	1081, 1083, 1087	solid, sound, set, strong, firm, rigid, dry, concrete, hard	1084	estable, solido, integro, formal, seguro, firme, real, consistente, fuerte, fundado
	1090, 1082, 1088, 1085, 1084, 1089, 1086	fixed, secure, substantial, valid, reliable, good, sturdy, respectable, convincing, sound, substantive, dependable, strong, genuine, cemented, firm, accurate, stable	1090, 1081, 1087, 1086, 1082, 1083, 1085, 1088, 1089	fidedigno, con cuerpo, conciso, estable, macizo, solido, con fundamentos, tempano, fundamentado, confiable, real, seguro, firme, consistente, fuerte, estricto, congelado, en estado solido, resistente, duro, bien fundado, fundado

whereas SEP is lower as anticipated. The two lemmas were selected as examples with the same number of COMPS to allow for a comparison of the values. The overlap measure nc_s is higher for *solid.a* as anticipated.[26]

Note that for the highly clusterable lemma *fire.v* there are no substitutes in common in the two groupings with either the LEXSUB or CLLS data because there is no substitute overlap in the sentences, which results in the COMPS and CLIQUES solutions being equivalent, whereas for *solid.a* there are several substitutes shared by the groupings for LEXSUB (e.g., *strong*) and CLLS (e.g., *solido*).

6. Conclusions and Future Work

In this article, we have introduced the theoretical notion of clusterability from machine learning discussed by Ackerman and Ben-David (2009a) and argued that it is relevant to WSI since lemmas vary as to the degree of partitionability, as highlighted in the linguistics literature (Tuggy 1993) and supported by evidence from annotation studies (Chen and Palmer 2009; Erk, McCarthy, and Gaylord 2009, 2013). We have demonstrated here how clustering of translation or paraphrase data can be used with clusterability measures to estimate how easily a word's usages can be partitioned into discrete senses. In addition to the intra-clust measures from the machine learning literature, we have also operationalized clusterability as consistency in clustering across information sources

26 The CLIQUES clustering gives a different number of clusters to the two lemmas, so these two lemmas would be in different polysemy bands for the correlation experiments on nc_s since we control for polysemy.

Table 12
Values of clusterability metrics for the examples *fire.v* and *solid.a.*

COMPS				
	LEXSUB		CLLS	
intra-clust	*fire.v*	*solid.a*	*fire.v*	*solid.a*
SEP	0.122	0.584	0.179	0.685
VR	7.178	0.713	4.579	0.459
WPR	1.732	0.845	1.795	0.707

LEXSUB and CLLS		
inter-clust	*fire.v*	*solid.a*
pF	1	0.081
V	1	0.590

CLIQUES				
	LEXSUB		CLLS	
baseline	*fire.v* (2 #cl)	*solid.a* (4 #cl)	*fire.v* (2 #cl)	*solid.a* (7 #cl)
nc_s	1.0	1.5	1	2.1

Gold-Standard from Usim		
Gold-Standard	*fire.v*	*solid.a*
Uiaa	0.930	0.490
Umid	0.169	0.630

using clustering solutions from translation and paraphrase data together. We refer to this second set of measures as inter-clust measures.

We conducted two sets of experiments. In the first we controlled for polysemy by performing correlations between clusterability estimates and our gold standard on our lemmas in three polysemy bands, which allows us to look at correlation independent of polysemy. In the second set of experiments we used linear regression on the data from all lemmas together, which allows us to see how polysemy and clusterability can work together as predictors. We find that the machine learning metrics SEP and VR produce the most promising results. The inter-clust metrics (*V* and *pF*) are interesting in that they consider the congruence of different views of the same underlying usages, but although there are some promising results, the measures are not as consistent and in particular in the second set of experiments do not outperform the baseline. This may be due to their reliance on CLLS, which generally produces weaker results compared to LEXSUB. Our baseline, which measures the amount of overlap in overlapping clustering solutions, shows consistently weaker performance than the intra-clust measures.

A variant of the inter-clust measures we would like to explore is a comparison of results from different clustering algorithms. Because more clusterable data is computationally easier to cluster (Ackerman and Ben-David 2009a), we assume more clusterable data should produce closer results across different algorithms operating on the same input data. We plan to test this empirically in future.

Clusterability metrics should be useful in planning annotation projects (and estimating their costs) as well as for determining the appropriate lexical representation for a lemma. A more clusterable lemma is anticipated to be better-suited to the traditional hard-clustering winner-takes-all WSD methodology compared with a less clusterable lemma where a more complex soft-clustering approach should be considered and more time and expertise is anticipated for any annotation and verification tasks. For some tasks, it may be worthwhile to focus disambiguation efforts only on lemmas with a reasonable level of partitionability.

We believe that notions of clusterability from machine learning are particularly relevant to WSI and the field of word meaning representation in general. These notions might prove useful in other areas of computational linguistics and lexical semantics in particular. One such area to explore would be clustering predicate-argument data (Sun and Korhonen 2009; Schulte im Walde 2006).

All the metrics and gold standards measure clusterability on a continuum. We have yet to address the issue of where the cut-off points on that continuum for alternate representations might be. There is also the issue that for a given word, there may be some meanings which are distinct and others that are intertwined. It may in future be possible to find contiguous regions of the data that are clusterable, even if there are other regions where the meanings are less distinguishable.

The paraphrase and translation data we have used to examine clusterability metrics have been produced manually. In future work, the measures could be applied to automatically generated paraphrases and translations or to vector-space or word (or phrase) embedding representations of the instances. Use of automatically produced data would allow us to measure clusterability over a larger vocabulary and corpus of instances but we would need to find an appropriate gold standard. One option might be evidence of inter-tagger agreement from corpus annotation studies (Passonneau et al. 2012) or data on ease of word sense alignment (Eom, Dickinson, and Katz 2012).

Appendix A: Individual Spearman's Correlation Trials

Tables A1–A5 provide the details of the individual Spearman's correlation trials of clusterability measures against the gold standards reported in Section 5.1. All correlations in the anticipated direction are marked in blue, and those in the counter-intuitive direction are marked in red and noted by *opp* in the final column. In the same column, we use * for statistical significance with $p < 0.05$ and ** for $p < 0.01$. We use only those polysemy bands where there are at least five lemmas within the polysemy range for that band. The number of lemmas (#) in each band is shown within parentheses.

Table A.1
Correlation of the intra-clust *k*-means metrics on CLLS against the Usim gold-standard rankings Uiaa and Umid.

Band (#)	Clusterability measure	Usim measure	ρ	sig/*opp*
low (22)	VR	Umid	–0.4349	*
low (22)	VR	Uiaa	0.4539	*
low (22)	SEP	Umid	0.6077	**
low (22)	SEP	Uiaa	–0.2041	
low (22)	WPR	Umid	–0.1187	
low (22)	WPR	Uiaa	–0.0266	*opp*

Table A.2
Correlation of the intra-clust *k*-means metrics on LEXSUB with the Usim gold-standard estimates Uiaa and Umid.

Band (#)	measure1	measure2	ρ	sig/*opp*
low (29)	VR	Umid	–0.6058	**
mid (10)	VR	Umid	–0.4073	
low (29)	VR	Uiaa	0.4049	*
mid (10)	VR	Uiaa	0.2364	
low (29)	SEP	Umid	0.359	
mid (10)	SEP	Umid	0.7416	*
low (29)	SEP	Uiaa	–0.3038	
mid (10)	SEP	Uiaa	–0.6606	*
low (29)	WPR	Umid	–0.4161	*
mid (10)	WPR	Umid	–0.4316	
low (29)	WPR	Uiaa	0.2739	
mid (10)	WPR	Uiaa	0.3818	

Table A.3
Correlation of the inter-clust metrics on LEXSUB-CLLS with the Usim gold-standards: Uiaa and Umid.

Band (#l)	measure1	measure2	ρ	sig/*opp*
low (29)	*pF*	Umid	–0.1365	
mid (5)	*pF*	Umid	–0.5	
low (29)	*pF*	Uiaa	0.1796	
mid (5)	*pF*	Uiaa	0.9	*
low (29)	*V*	Umid	–0.2456	
mid (5)	*V*	Umid	0	
low (29)	*V*	Uiaa	0.3849	*
mid (5)	*V*	Uiaa	0.6	

Table A.4
Correlation of the baseline nc_s operating on CLLS with the Usim gold-standard: Uiaa and Umid.

Band (#)	measure1	measure2	ρ	sig/*opp*
low (14)	nc_s	Umid	0.4381	
mid (17)	nc_s	Umid	0.0308	
high (9)	nc_s	Umid	–0.4622	*opp*
low (14)	nc_s	Uiaa	–0.3455	
mid (17)	nc_s	Uiaa	–0.4948	*
high (9)	nc_s	Uiaa	0.3713	*opp*

Table A.5
Correlation of the baseline nc_s operating on LEXSUB with the Usim gold-standard Uiaa and Umid.

Band (#)	measure1	measure2	ρ	sig/*opp*
low (14)	nc_s	Umid	0.2668	
mid (19)	nc_s	Umid	0.2204	
high (10)	nc_s	Umid	−0.179	*opp*
low (14)	nc_s	Uiaa	−0.3327	
mid (19)	nc_s	Uiaa	−0.2447	
high (10)	nc_s	Uiaa	0.0617	*opp*

Acknowledgments

This work was partially supported by National Science Foundation grant IIS-0845925 to K. E. We thank the anonymous reviewers for many helpful comments and suggestions.

References

Ackerman, Margareta and Shai Ben-David. 2009a. Clusterability: A theoretical study. *Journal of Machine Learning Research - Proceedings Track*, 5:1–8.

Ackerman, Margareta and Shai Ben-David. 2009b. Clusterability: A theoretical study. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1–8, Clearwater Beach, FL.

Agirre, Eneko and Philip Edmonds, editors. 2006. *Word Sense Disambiguation, Algorithms and Applications*. Springer.

Apidianaki, Marianna. 2008. Translation-oriented word sense induction based on parallel corpora. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 3269–3275, Marrakech.

Apidianaki, Marianna. 2009. Data-driven semantic analysis for multilingual WSD and lexical selection in translation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL'09)*, pages 77–85, Athens.

Apidianaki, Marianna, Emilia Verzeni, and Diana McCarthy. 2014. Semantic clustering of pivot paraphrases. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4270–4275, Reykjavik.

Artiles, Javier, Enrique Amigó, and Julio Gonzalo. 2009. The role of named entities in Web People Search. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*, pages 534–542, Singapore.

Bannard, Colin and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, MI.

Bansal, Mohit, John DeNero, and Dekang Lin. 2012. Unsupervised translation sense clustering. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL: HLT)*, pages 773–782, Montréal.

Biemann, Chris and Valerie Nygaard. 2010. Crowdsourcing WordNet. In *Proceedings of the 5th Global WordNet Conference*, Mumbai.

Carpuat, Marine and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 61–72, Prague.

Chen, Jinying and Martha Palmer. 2009. Improving English verb sense disambiguation performance with linguistically motivated features and clear sense distinction boundaries. *Journal of Language Resources and Evaluation, Special Issue on SemEval-2007*, 43:181–208.

Copestake, Ann and Ted Briscoe. 1995. Semi-productive polysemy and sense extension. *Journal of Semantics*, 12:15–67.

Cruse, D. A. 2000. Aspects of the microstructure of word meanings. In Yael Ravin and Claudia Leacock, editors, *Polysemy: Theoretical and Computational Approaches*. Oxford University Press, pages 30–51.

Di Marco, Antonio and Roberto Navigli. 2013. Clustering and diversifying Web search results with graph-based word sense induction. *Computational Linguistics*, 39(3):709–754.

Dyvik, Helge. 1998. Translations as semantic mirrors: From parallel corpus to Wordnet. In *Proceedings of the Workshop Multilinguality in the Lexicon II at the 13th Biennial European Conference on Artificial Intelligence (ECAI'98)*, pages 24–44, Brighton.

Eom, Soojeong, Markus Dickinson, and Graham Katz. 2012. Using semi-experts to derive judgments on word sense alignment: A pilot study. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 605–611, Istanbul.

Epter, Scott, Mukkai Krishnamoorthy, and Mohammed Zaki. 1999. Clusterability detection and initial seed selection in large data sets. Technical Report 99-6, Rensselaer Polytechnic Institute, Computer Science Department, Troy, NY.

Erk, Katrin, Diana McCarthy, and Nick Gaylord. 2009. Investigations on word senses and word usages. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 10–18, Suntec.

Erk, Katrin, Diana McCarthy, and Nick Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics*, 39(3):511–554.

Fellbaum, Christiane, editor. 1998. *WordNet, An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Goldberg, Mark K., Mykola Hayvanovych, and Malik Magdon-Ismail. 2010. Measuring similarity between sets of overlapping clusters. In *SocialCom/PASSAT*, pages 303–308, Minneapolis, MN.

Hanks, Patrick. 2000. Do word meanings exist? *Computers and the Humanities. Senseval Special Issue*, 34(1–2):205–215.

Hovy, Eduard, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL: HLT)*, pages 57–60, New York, NY.

Ide, Nancy, Tomaž Erjavec, and Dan Tufiş. 2002. Sense discrimination with parallel corpora. In *Proceedings of the ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 54–60, Philadelphia, PA.

Ide, Nancy and Yorick Wilks. 2006. Making sense about sense. In Eneko Agirre and Phil Edmonds, editors, *Word Sense Disambiguation, Algorithms and Applications*. Springer, pages 47–73.

Jurgens, David and Ioannis Klapaftis. 2013. Semeval-2013 task 13: Word sense induction for graded and non-graded senses. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 290–299, Atlanta, GA.

Kilgarriff, Adam. 1998. 'I don't believe in word senses'. *Computers and the Humanities*, 31(2):91–113.

Kilgarriff, Adam. 2006. Word Senses. In Eneko Agirre and Phil Edmonds, editors, *Word Sense Disambiguation, Algorithms and Applications*. Springer, pages 29–46.

Krippendorff, Klaus. 1980. *Content Analysis: An Introduction to Methodology*. Sage Publications, Inc., Beverly Hills, CA.

Lakoff, George. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago Press.

Landes, Shari, Claudia Leacock, and I. Tengi Randee. 1998. Building Semantic Concordances. In Christiane Fellbaum, editor, *WordNet: an Electronic Lexical Database*. MIT Press, Cambridge, MA, pages 199–237.

Lefever, Els and Veronique Hoste. 2010. SemEval-2007 Task 3: Cross-lingual word sense disambiguation. In *Proceedings of the Fifth International Workshop on Semantic Evaluations (SemEval-2010)*, pages 15–20, Uppsala.

Manandhar, Suresh, Ioannis Klapaftis, Dmitriy Dligach, and Sameer Pradhan. 2010. SemEval-2010 Task 14: Word sense induction and disambiguation. In *Proceedings of the Fifth International Workshop on Semantic Evaluation (SemEval-2010)*, pages 63–68, Uppsala.

McCarthy, Diana. 2009. Word sense disambiguation: An overview. *Language and Linguistics Compass*, 3(2):537–558.

McCarthy, Diana. 2011. Measuring similarity of word meaning in context with lexical substitutes and translations. In

Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing, CICLing 2011, Pt. I (Lecture Notes in Computer Science, LNTCS 6608)*. Springer, pages 238–252.

McCarthy, Diana and Roberto Navigli. 2007. SemEval-2007 Task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague.

McCarthy, Diana and Roberto Navigli. 2009. The English lexical substitution task. *Language Resources and Evaluation Special Issue on Computational Semantic Analysis of Language: SemEval-2007 and Beyond*, 43(2):139–159.

Mihalcea, Rada, Ravi Sinha, and Diana McCarthy. 2010. SemEval-2010 Task 2: Cross-lingual lexical substitution. In *Proceedings of the Fifth International Workshop on Semantic Evaluation (SemEval-2010)*, pages 9–14, Uppsala.

Mitchell, Jeff and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL'08: HLT)*, pages 236–244, Columbus, OH.

Navigli, Roberto. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.

Ostrovsky, Rafail, Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. 2006. The effectiveness of Lloyd-type methods for the k-means problem. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 165–176, Berkeley, CA.

Palmer, Martha, Hoa Trang Dang, and Joseph Rosenzweig. 2000. Semantic tagging for the Penn treebank. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, pages 699–704, Athens.

Passonneau, Rebecca, Ansaf Salleb-Aouissi, Vikas Bhardwaj, and 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'10)*, pages 3244–3249, Valleta.

Passonneau, Rebecca J., Collin F. Baker, Christiane Fellbaum, and Nancy Ide. 2012. The MASC word sense corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3025–3030, Istanbul.

Pedersen, Ted. 2006. Unsupervised corpus-based methods for WSD. In Eneko Agirre and Phil Edmonds, editors, *Word Sense Disambiguation, Algorithms and Applications*. Springer, pages 131–166.

Resnik, Philip and David Yarowsky. 1997. A perspective on word sense disambiguation methods and their evaluation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*, pages 79–86, Washington, DC.

Resnik, Philip and David Yarowsky. 2000. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(3):113–133.

Rosenberg, Andrew and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague.

Schulte im Walde, Sabine. 2006. Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics*, 32(2):159–194.

Sun, Lin and Anna Korhonen. 2009. Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*, pages 638–647, Singapore.

Tuggy, David H. 1993. Ambiguity, Polysemy and Vagueness. *Cognitive Linguistics*, 4(2):273–290.

Véronis, Jean. 2004. HyperLex: Lexical Cartography for Information Retrieval. *Computer Speech & Language*, 18(3):223–252.

Yuret, Deniz. 2007. KU: Word sense disambiguation by substitution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 207–214, Prague.

Zhang, Bin. 2001. Dependence of clustering algorithm performance on clustered-ness of data. Technical report HP-2001-91, Hewlett-Packard Labs.