A larger-scale evaluation resource of terms and their shift direction for diachronic lexical semantics

Astrid van Aggelen[♠] **Antske Fokkens**[♢] **Laura Hollink**[♠] **Jacco van Ossenbruggen** [♠♢]
♠ Centrum Wiskunde Informatica, Amsterdam, Netherlands
♢ Vrije Universiteit Amsterdam, Netherlands
a.e.van.aggelen@cwi.nl, antske.fokkens@vu.nl

Abstract

Determining how words have changed their meaning is an important topic in Natural Language Processing. However, evaluations of methods to characterise such change have been limited to small, handcrafted resources. We introduce an English evaluation set which is larger, more varied, and more realistic than seen to date, with terms derived from a historical thesaurus. Moreover, the dataset is unique in that it represents change as a shift from the term of interest to a WordNet synset. Using the synset lemmas, we can use this set to evaluate (standard) methods that detect change between word pairs, as well as (adapted) methods that detect the change between a term and a sense overall. We show that performance on the new data set is much lower than earlier reported findings, setting a new standard.

1 Introduction

Determining how words have changed their meaning is an important topic in Natural Language Processing (Tang, 2018; Kutuzov et al., 2018; Tahmasebi et al., 2018). Using large diachronic corpora, computational linguistics has provided methods that can detect or qualitatively explain semantic change automatically. In particular, several approaches have been introduced that use distributional semantic models representing different time periods in diachronic corpora (Gulordava and Baroni, 2011; Mitra et al., 2014; Kulkarni et al., 2015, e.g.).

Researchers have illustrated through compelling examples that these methods can detect semantic shift, like *cell* obtaining the meaning of 'phone' and *gay* shifting from 'cheerful' to 'homosexual' (Mitra et al., 2014, e.g.) and have reported high accuracy on small evaluation sets of selected examples. Hamilton et al. (2016) even report 100% accuracy in detecting known change on 28 word pairs. As a result, these approaches have been enthusiastically adopted (Wohlgenannt et al., 2019; Orlikowski et al., 2018; Kutuzov et al., 2016; Martinez-Ortiz et al., 2016, e.g.). However, it has been called into question how reliable these methods really are (Hellrich and Hahn, 2016a; Dubossarsky et al., 2017).

These developments show that there is both a wide interest in using distributional semantic models to assess semantic change and an urgent need for better insight into the possibilities and limitations of these methods. It is therefore unsurprising that three recent survey papers on the topic all list the lack of proper evaluation and, in particular, the absence of large-scale evaluation sets, as a key challenge for this line of research (Tang, 2018; Kutuzov et al., 2018; Tahmasebi et al., 2018).

In this paper, we automatically derive HiT, the largest English evaluation set to date, from a historical thesaurus. HiT consists of terms linked to WordNet (Fellbaum, 2012) entries that represent senses they gained or lost. We introduce *sense shift assessment* as a task, enabled by this dataset, that identifies whether a sense of a term of interest was coming in our out of use, based on its changed relationship with all lemmas of the sense. This is a variation of a task introduced by Hamilton et al. (2016) that assesses the relationship of the terms of interest with individual other terms. The sense shift assessment instead uncovers the conceptual change that explains multiple observed trends between word pairs. Cross-checking and summarising individual observations also means drawing more informed conclusions. Furthermore, the use of WordNet sense representations allows for the dataset entries to be automatically derived rather than manually (expert) collected, hence limiting the effect of bias. We use HiT to answer two main research questions. First, how well can current

methods detect sense shift on a larger and more varied evaluation set? Second, how, by taking a full synset as a representation of meaning, does the task of detecting sense shift compare to studying word pairs in isolation? The main contributions of this paper are as follows. First, the new evaluation set, consisting of 756 target words and 3624 word pairs. Second, we show that current methods perform quite poorly on this more challenging set, thus confirming that this set introduces a new benchmark. We also identify lexical factors that contribute to these differences.

2 Related Work

This section provides an overview of previous work on detecting lexical semantic change through distributional semantic models.

Distributional models of meaning are motivated by the hypothesis that words with similar meanings will occur in similar contexts (Harris, 1954; Firth, 1957). Tahmasebi et al. (2011) and Mitra et al. (2014) induce clusters of terms that temporally co-occur in particular syntactic patterns, and (qualitatively or quantitatively) trace their development. Their approach forms a bridge from previous document-based approaches (Blei and Lafferty, 2006; Wang and McCallum, 2006, e.g.) to the window-based models that are currently widely used.

Gulordava and Baroni (2011) and Jatowt and Duh (2014) were among the first to use trends in similarity between distributional representations. The former detect change within single terms by tracing their self-similarity. The latter, like we do, interpret the change of a term by contrasting it with other terms. In recent work, the most common type of distributional models used to assess semantic shift are known as prediction models (Kim et al., 2014; Kulkarni et al., 2015; Hamilton et al., 2016, e.g). In this paper, we use embeddings that gave the best results in Hamilton et al. (2016) and are created through the skip-gram method included in word2vec (Mikolov et al., 2013).

Until recently, semantic shifts were determined by comparing the distributional nearest neighbours of a term in one time period to its neighbours in another (see e.g. Tahmasebi et al. (2011)). However, such an inspection is difficult to carry out at scale, is not suited for disappearing senses - distant neighbours are hard to assess - and is prone to bias, especially when the aim is to confirm hypothesized trends. Hamilton et al. (2016) use a predetermined list of terms to which the target term got more and less related over time. This variation alleviates the problem of bias and introduces 'more distant neighbours' into the analysis, but with just 28 term pairs it is still very small-scale.

Basile and McGillivray (2018) are, to our knowledge, the first to exploit a large historical dictionary as an evaluation source. The aim of their work is to detect changed terms and their change point, based on dips in the self-similarities, with the Oxford English Dictionary as the gold standard. To verify whether the observed change point corresponds to a new dictionary sense, the time-specific nearest neighbours of the term are contrasted with the dictionary definition. This work could have provided the evaluation set for the task addressed in this paper. However, as far as we know, the authors have not enriched the data with said nearest neighbours nor made them available. Hence, the current work is still the first to provide a large-scale evaluation set based on a dictionary.

3 A large-scale sense shift assessment set

This section describes a new evaluation set that links terms of interest (target terms) to rich synset representations of their old and new senses. This means that in an experimental setting (such as that in Section 5), the target term can be contrasted to a predetermined, varied set of terms. We also adapt two existing evaluation sets, HistWords (Hamilton et al., 2016) and the Word Sense Change Testset[1] (Tahmasebi and Risse, 2017), into datasets of the same format.

3.1 Deriving a sense shift assessment set

The new dataset, which we call **HiT**, is derived from **The Historical Thesaurus of the University of Glasgow** (Kay et al., 2019). This thesaurus lists (nearly all) English terms organised in a conceptual hierarchy of senses. It also documents the time period in which a term was attested and assumed to be active in the given sense. For instance, one entry says that the verb *bray* was used in the sense of 'Grind/pound' (in turn a subconcept of 'Create/make/bring about') for the period 1382 till 1850. The thesaurus does not indicate how any listed sense of a word relates to previous, concurrent or future ones. Hence, it is unclear whether

[1] http://doi.org/10.5281/zenodo.495572

the term underwent a process of semantic narrowing or broadening, or whether it lost or gained a sense altogether. The change that is considered here is a broad notion of rising or declining senses.

For HiT we use all terms in senses that came up between 1900 and 1959 and all terms in senses that disappeared - which were less numerous - between 1850 and 1959. To enrich the thesaurus terms, we identify their WordNet synsets and check if each of these expresses the intended meaning. This check is based on the overlap between the thesaurus definition and the synset terms. If any term from the synset (excluding the target term itself) overlaps with any of the terms from the thesaurus definition (in the example above, the terms are {*create*, *make*, *bring about*, *grind*, *pound*}), we assume that the synset in question provides the intended sense. The verb *bray* appears in WordNet under two polylexical synsets: {*bray*, *hee-haw*}, and {*bray*, *grind*, *mash*, *crunch*, *comminute*}. Due to the overlapping term *grind*, the thesaurus entry is matched with the latter, meaning 'reduce to small pieces or particles by pounding or abrading', but not with the former ('laugh loudly and harshly').

Newly emerging senses from the thesaurus provide gold standard instances that are supposed to attract the vector of the target term. Disappearing senses, on the other hand - such as 'grind' for *bray* after 1850 - are gold standard instances from which the target term should move away. Table 1 shows how the *bray* example translates to a HiT entry of a vanishing sense with the WordNet synonyms used as reference terms.

Only entries with at least one identified WordNet synset are included. This results in a dataset of 756 target terms exhibiting 979 sense shifts.

target term	POS	sense (WN synset) ⌊ reference term	shift	onset
bray	v.	`grind.v.05` ⌊ grind, mash, crunch, comminute	-1	1850

Table 1: Example excerpt from an entry of HiT. Shift label -1 means a move away from the given meaning.

Validation To establish the accuracy of the WordNet matching method, two raters independently annotated a subset of 191 entries. The agreement between the raters (i.e. the proportion of agreement above chance level) is assessed using Cohen κ for two raters (Cohen, 1960). Also, we assessed how well the raters' judgement corresponded to the output of the automated WordNet matching, i.e., the supposed gold standard.

Rater 1 verified whether the algorithm had selected the correct synset or not. To counteract an effect of bias from the gold standard, rater 2 did not work with the gold standard, rather indicating for any given WordNet synset whether it represented the given definition. These findings were then translated to a judgement of the algorithm output in line with rater 1. The annotators agreed on the evaluation by 97.9 per cent; by Cohen's chance-normalised norm ($\kappa = 0.789$, $z = 11$, $p < 0.001$), this is generally thought of as 'substantial' agreement. Except for 10 (rater 1) and 9 (rater 2) instances out of 191, the raters' judgements corresponded to that of the algorithm, an error rate of approximately 5 per cent. We consider this high enough to take the outcome of the synset linking method as the gold standard.

3.2 Transforming existing datasets

The thesaurus-derived dataset, HiT, qualitatively differs from existing evaluation sets in its automated construction and in its representation of senses by a synset rather than selected terms. In order to compare this set to previously used datasets, we adapt two standard evaluation sets semi-automatically to link the target words they contain to synsets representing the given old or new senses.

HistWords (Hamilton et al., 2016) (HW) contains 28 word pairs that saw their similarity increase or decrease over time, based on 9 target terms. HistWords states the onset of the change - no end date - and the gold standard shift direction. For instance, since 1800, *awful* has moved towards *mess* and *disgusting* and away from *impressive*.

The Word Sense Change Testset (WSCT) (Tahmasebi and Risse, 2017) lists terms that acquired a new sense and unchanged control words. It gives the type of change the term underwent (no change, new, broader, or narrower sense), a short explanation of the change and the onset date of the change. For instance, *memory* acquires a new related sense 'digital memory' in 1960 whilst keeping its existing sense 'human memory'.

From HW to HW+ and from WSCT to WSCT+. Every entry from WSCT and HW is treated as a separate change event, with an onset date and a description; some target terms have more than one change event. For each such event, the affected sense(s) are selected out of all can-

didate senses, i.e. all WordNet synsets that correspond to the target lemma (in the correct part of speech). This synset selection process happened manually, by comparing the lexical information in WordNet against the change description in the source data. More details about the annotation process are given below. The outcomes determine the change type listed for the combination of target term and synset in the enriched datasets **WSCT**+ and **HW**+. The target term is thought to move towards any synset (and towards all terms of the synset) that captures an increasingly common sense, and away from any synset that expresses an increasingly uncommon sense. For any synset that does not capture any described change, its relation to the target term is described as unchanged or unknown. See Table 2 for an example.

Annotation process and validation. The selection was carried out by the first author; this was then evaluated by two co-authors for HW+ and one co-author for WSCT+. In the case of three raters, we used Fleiss' extension of Cohen's method (Fleiss, 1971). The raters judged the shift direction of the target term with respect to all candidate concepts (synsets): towards (+1), away from (-1) one another, or no change (0). The evaluation set of word-sense combinations was larger than the final dataset (Table 3), as it included synsets with just the target term. The raters were given the following data: the change description and the time of change, given in HW or WSCT; the (given or inferred) part of speech; the candidate WordNet concept that connects the two terms; and corresponding WordNet data such as the definition and the set of terms in the synset. For WSCT+ (N=129 target-sense pairs), the two raters agreed by 88.4 per cent, which, chance-normalised (Cohen's $\kappa = 0.63$, $z = 7.26$, $p < 0.01$) is thought to be 'substantial'. The raters then agreed on the final set of gold standard labels. On HW+ (N=70 target-sense pairs), the three raters agreed almost perfectly (Fleiss' $\kappa = 0.83$, $z = 16$, $p < 0.01$), and the ratings by the first author were taken as the gold standard.

Resulting datasets The evaluation sets all provide two types of pairings: target terms paired with reference terms and target terms with synsets. The gold standard for the individual word pairs - target word and synonym - corresponds to the gold standard for the whole synset. After the inter-rater evaluation, synsets with just one term (the target term) were omitted, as the experiment requires reference terms. Table 3 provides an overview of the resulting evaluation sets next to HistWords as a baseline. HiT does not show any overlap with the other datasets except for a single target term lemma (verb *call*) in common with HW+, however with the described change in a different meaning.

target term	POS	sense (WN synset) ⌊ reference term	t	shift
memory	n.	`memory.n.03` ⌊ retention, retentiveness, retentivity	1960	0
memory	n.	`memory.n.04` ⌊ computer memory, storage, computer storage, store, memory board	1960	1

Table 2: Excerpts from two entries of WSCT+. Shift label 0 means we have no evidence the word changed with respect to the given sense. Label 1 means a shift towards the indicated meaning (and its associated terms).

dataset dataset type	HW existing	HW+ adapted	WSCT+ adapted	HiT new
target words (TWs)	9	9	23	756
TW+term	28	117	213	3624
converging	18	41	56	1173
diverging	10	24	0	2451
unchanged/-known	0	52	157	0
TW+sense	n.a.	42	93	979
converging	n.a.	10	23	282
diverging	n.a.	10	0	697
unchanged/-known	n.a.	22	70	0

Table 3: Contents of the evaluation sets, which come in two variants: target terms paired with other terms and paired with WordNet senses (synsets). This allows the datasets to be used for the two types of evaluation used here.

4 Experiment

Two tasks are addressed in the experiment. Word shift assessment (**WordShiftAssess**) (Hamilton et al., 2016) is summarised as follows: given a target term, a reference term, and a time period, did the two terms become closer in meaning (gold standard label 1) or did their meanings move apart (label -1)? Sense shift assessment (**SenseShiftAssess**) goes as follows: given a target term, a WordNet synset, and a time period, did the target term in the given period move towards or away from the given sense? To be comparable to previous findings, we evaluate the datasets on both tasks. This section outlines the methods and the experimental setup.

4.1 Change assessment for word-word pairs

Our method of determining shift direction was proposed by Hamilton et al. (2016). It depends on the availability of distributional representations for the target term and reference terms, corresponding

to the synset lemmas, at regular intervals between the start and the end of the period of interest. The successive cosine similarities of the embeddings of the target and reference term are (Spearman) correlated with the time index (e.g. 1800, 1810, .., 1990). If the correlation is positive, the target term is taken to have moved towards the reference term; if it is negative, away from it. Given the binary classification setting, the statistical significance of the correlation factor has no clear interpretation. However, we include it to comply with earlier reported findings and for readers to judge the potential of the method for a three-way classification with a null category.

4.2 Change assessment for word-sense pairs

To address SenseShiftAssess, we suggest two broad approaches. The first starts from the method outlined in Section 4.1. That is, for any target-sense pair, we start from the given target word paired with all lemmas of the synset, and the trend of the cosine similarities (Spearman ρ and p) for each of these word pairs. Then, we either take the most-observed sign of ρ as the outcome (**majority vote**), or we promote one word pair to exemplify the sense shift as a whole. Assuming that an observed strong trend is likely to be correct, **argmax(corr)** takes the sign of the highest absolute ρ value of all word pairs in the synset as the synset assessment. **argmin(p(corr))** does the same for the observation the correlation coefficient of which has the lowest p value.

The second approach we suggest, **average vec.**, operates on a lower level, as it aggregates the distributional representations of the synset lemmas into an average vector, for every time slice separately. The target term and the averaged representation are then treated like a word pair (Section 4.1).

4.3 Experimental setup

We apply word shift assessment on HW, HW+, WSCT+, and HiT. The reference terms of HW+, WSCT+ and HiT come from WordNet; those in HW are readily taken from the source. Figure 1 illustrates how the term *awful* from HW+ compares with its individual WordNet synonyms over time.

Sense shift assessment is applied to WSCT+, HW+ and HiT, i.e. all sets that could be enriched with sense information. To continue with the example in Figure 1, sense shift assessment translates the word-based observations into a single assessment of the changed relation of *awful* with respect to the whole synset.

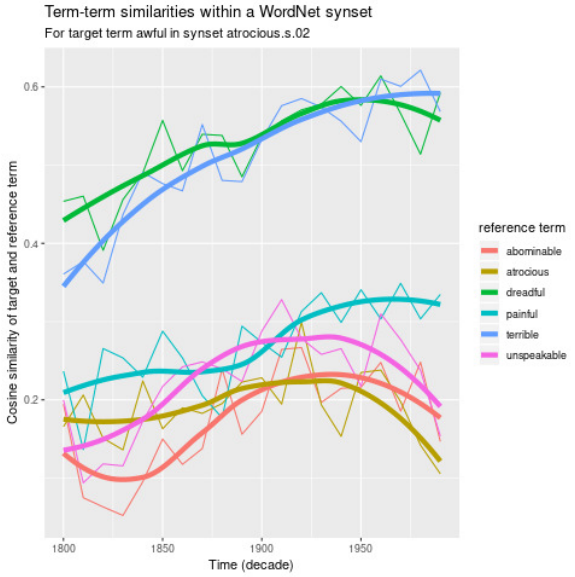

Figure 1: WordShiftAssess with WordNet-based reference terms: target term *awful* is individually contrasted with all terms from synset `atrocious.s.02`: *abominable*, *atrocious dreadful*, *painful*, etc. The fitted lines illustrate the observed trend in cosine similarities, such as the growing similarity between *awful* and *terrible*.

Distributional vectors. We use word embeddings provided by Hamilton et al. (2016) of size 300, trained through skip-gram with negative sampling (SGNS) for every decade separately on three corpora: the Corpus of Historical American English (Davies, 2015) (COHA), the complete English Google N-Gram corpus (Michel et al., 2011), and the English Fiction corpus, a subset of the Google N-Gram corpus. The embeddings are not (part of speech) disambiguated, and can stand for several lemmas at once. We employ the embeddings for every decade from the attested onset of change up to and including the last available embedding, trained on the 1990s subcorpus.

Handling of missing and infrequent data. Some terms appear infrequently in some slices of the corpus. The code that accompanies Hamilton et al. (2016) deals with these cases by padding the cosine time series with a zero for the dimension (i.e. time slice) in which either or both of the terms was insufficiently frequent (under 500 times, except for COHA). However, this biases the outcome, since zero is the smallest cosine similarity value. Given that low word frequencies are more common in the corpora of the first few decades, this setting makes it more likely to find

cosine time series								ρ
t_1	...	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	t_{15}	
0	...	**0**	0.25	0.29	0.29	0.20	0.18	**0.77**
NA	...	**NA**	0.25	0.29	0.29	0.20	0.18	**-0.7**

Table 4: Similarity values based on infrequent data must not be padded with zero as this biases the correlation value towards a positive value. In the word pair *delimit-define*, padding the values for decades t_1 (in fact, 1850) through to t_{11} (1940) with zeros would lead to a conclusion opposite to the ground truth stating that these terms move away from each other; hence these observations are treated as missing data (NA) instead.

a rising trend in cosine similarities. As this is an unwanted effect, we treated cosine values based on low-frequency numbers as missing values. Table 4 illustrates the difference between the caveat explained here and the approach taken. A further count filter ensures that all results (correlations) are based on at least five cosine values.

5 Results

Table 5 shows the proportion of word pair observations (**WordShiftAssess**) displaying the expected trend in cosine similarities for every dataset and training corpus. The significance reported is the proportion of *correct* findings (i.e. with an upper limit of 100%) with a Spearman ρ significant on the 0.05 level. Whether the correlation coefficient is significant depends on its magnitude as well as the number of cosine values considered. The latter in turn depends on the change onset - the longer the time series, the more observations - minus observations that were based on too little data and were left out (see Section 4.3). N expresses how many of the word pair entries from the datasets (Table 3) which displayed a real shift (unchanged words were not used) resulted in a cosine time series of at least five observations (see Section 4.3). This depends in part on the corpus, some of which have much greater coverage than other ones, particularly the complete English corpus, eng-all. For instance, the results for HiT for eng-all are based on 1461 word pairs as opposed to a mere 746 for COHA and 772 for English fiction, out of a dataset total of 3624 shifted terms. Moreover, eng-all resulted in more statistically significant correct outcomes than COHA and eng-fic. We therefore focus on the results based on eng-all in particular.

HiT appears more challenging than WSCT+ and HW+. On eng-all, just under 60 per cent of all entries were correctly assessed, as opposed to around 70 per cent for WSCT+ and 80 per cent

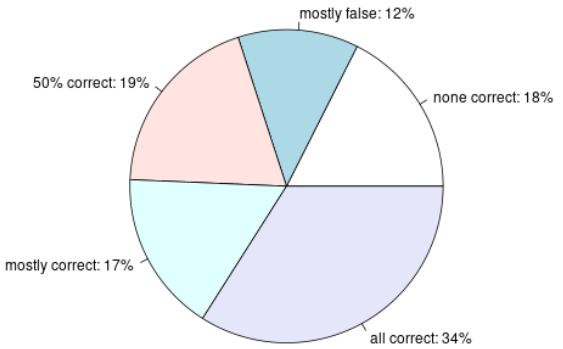

Figure 2: Proportions of correct word pairs (i.e. displaying the expected similarity trend) within synsets for HiT (on eng-all). In just over half of the cases, the synset contained more correct than incorrect observations (bottom half of the pie chart).

for HW+. The significance levels show a similar pattern, hence even the word pairs that showed the predicted trend did so less clearly for HiT than for the other datasets. The outcomes on COHA and eng-fic confirm the pattern for eng-all: HiT figures consistently lag behind WSCT+ and even further behind HW+. While eng-all and eng-fic give similar levels of accuracy, on COHA, the outcomes for HiT are below chance level.

HiT differs from WSCT+ and HW+ in that the target terms were not selected for the task at hand. Unsurprisingly, the automatically selected dictionary terms offer a more challenging evaluation set than the purposely selected terms in HW+ and WSCT+. The difference observed between HW and HW+ reveals a similar trend concerning hand-picked reference terms compared to (semi)-automatically selected ones: the performance on HW+ is about 20 per cent lower than for HW. In sum, the selection of term pairs has great impact.

Table 6 shows the proportion of target-sense entries that were correctly assessed (**SenseShiftAssess**), based on several possible aggregations of word pair level findings. Looking, firstly, at the different methods, argmin(p(corr)) performed best. Hence, the word pair within a synset that shows the most statistically robust change is the best indicator of the conceptual change of the target term. For HiT this resulted in 61.3% correct on eng-all and up to 64.0 % on eng-fic. The added performance over argmax(corr), whilst marginal (e.g. for HiT, 61.1 % on eng-all and 61.9% on eng-fic), suggests that balancing the correlation factor with the number of observations leads to better judgements than looking at the

WordShiftAssess	**eng-all**				**coha**				**eng-fic**			
	HiT	wsct+	HW+	HW	HiT	wsct+	HW+	HW	HiT	wsct+	HW+	HW
correct (%)	**58.0**	**69.2**	**79.5**	**100.0**	**45.6**	54.5	63.6	85.7	**59.5**	62.5	66.7	88.2
sig (%)	38.4	66.7	60.0	88.9	29.1	33.3	19.0	66.7	24.0	40.0	30.0	46.7
N	1461	13	44	27	746	11	33	21	772	8	30	17

Table 5: Results of determining the shift direction of a target word with respect to a reference word (WordShiftAssess).

SenseShiftAssess												
corpus	**eng-all**				**coha**				**eng-fic**			
dataset	HiT	wsct+	HW+	HW	HiT	wsct+	HW+	HW	HiT	wsct+	HW+	HW
average vec.	**57.8**	**100.0**	**88.9**	-	44.1	50.0	57.1	-	**58.9**	100.0	66.7	-
argmax(corr)	**61.1**	100.0	84.6	-	46.4	80.0	53.8	-	61.9	100.0	66.7	-
majority vote	**50.7**	100.0	76.9	-	**36.0**	80.0	53.8	-	**49.7**	75.0	66.7	-
argmin(p(corr))	**61.3**	100.0	84.6	-	46.8	80.0	53.8	-	**64.0**	75.0	66.7	-
N average vec.	374	2	9	-	204	2	7	-	214	1	6	
N other methods	450	5	13	-	278	5	13	-	286	4	12	-

Table 6: Results of determining the shift direction of a target word with respect to a reference word (SenseShiftAssess).

magnitude of the correlation alone. Averaging the vectors of all reference terms (average vec.) was less reliable an aggregation overall than promoting one word pair to represent the synset. However, it still did better than the majority vote, which required more than half of the word pairs in a synset to display the expected shift pattern. For HiT, this did not surpass chance level on any of the corpora. Hence, within a synset, false observations can be as numerous as or outweigh true ones, and heuristics are needed to find the signal in the noise.

SenseShiftAssess was expected to suffer less from noisy results that occur on the word level (WordShiftAssess). However, the improvement observed over the word pair results was marginal. For instance, for HiT on eng-all, the synset-level approach was correct in 61 per cent of cases at best, as opposed to 58 per cent on the word pair level. HW+ and WSCT+ did benefit more from the synset aggregation, but the small sample size makes it hard to draw conclusions from this. Figure 2 shows how much we can rely on the terms within a synset to display the anticipated change in relation to the target term (for HiT and eng-all). The vast majority of synsets - all except 18 per cent - contain at least one word pair that displays the true shift. This means that SenseShiftAssess on HiT is feasible, at least in theory, and the maximum accuracy attainable on eng-all is 82 per cent. A third of all synsets (34%) have all word pairs displaying the predicted shift; hence the lower limit is 34 per cent. There were more synsets with mostly correct than mostly false examples, and more synsets with just correct (33 per cent) than just false (17 per cent) ones. Based on the slightly higher odds of picking a correct than an incorrect example, our selection methods are perhaps not informed by the most determining factors. To know how we can find the signal in noisy word pair patterns, we must first understand what causes noise. This question is addressed next.

6 Follow-up analysis

We examine several lexical and corpus factors that may play a role in the outcomes. For every term in a word pair, we look at its polysemy, its frequency in the training corpus, and its typicality as a representation of the underlying concept. Table 7 breaks down the WordShiftEval results by the combined lexical properties of target and reference term. For every entry type we distinguish (e.g. a low-polysemous target term paired with a high-polysemous reference term), we examine the proportion of the result set it accounts for, and the observed tendency for such word pairs to display the expected shift pattern. Due to its fewer outcomes, WSCT+ was left out of this analysis.

When a term is ambiguous, i.e. when it tends to occur in various semantic and syntactic contexts, its distributional representation might be less suitable to reflect any single one of these. Polysemy is difficult to define (Ravin and Leacock, 2000). Here, we define the **polysemy** of a term by its total number of synsets divided by the number of different parts of speech it can occur in. We tested different thresholds for considering a term polysemous, from two to six synsets per part of speech, which all revealed similar results. The results we report are based on a minimum of three, four and

			entry type (N.B.: 'low' can mean more or less challenging, depending on the property)							
			low-low		low-high		high-low		high-high	
property	threshold	corpus	accuracy	proportion	accuracy	proportion	accuracy	proportion	accuracy	proportion
polysemy	3*	HiT	**63.8**	17.8	**60.9**	18.2	**62.8**	18.4	**52.7**	**45.6**
	3*	HW	100	37	100	29.6	100	18.5	100	**14.8**
	3*	HW+	70.6	38.6	71.4	31.8	100	20.5	100	**9.1**
polysemy	4*	HiT	**66.4**	**32.2**	**58.1**	21.1	**59.6**	17.5	**47.9**	**29.3**
	4*	HW	100	51.9	100	25.9	100	11.1	100	11.1
	4*	HW+	82.1	63.6	60	22.7	100	9.1	100	4.5
polysemy	5*	HiT	**64.8**	**45.3**	**58.7**	19.7	**53**	14.9	**45.9**	**20.1**
	5*	HW	100	70.4	100	14.8	100	14.8	-	0
	5*	HW+	81.8	75	40	11.4	100	11.4	100	2.3
frequency	100k	HiT	65.7	**33.9**	**61.4**	17.2	58.6	17.1	47.6	**31.8**
	100k	HW	100	**18.5**	100	11.1	100	22.2	100	**48.1**
	100k	HW+	80	34.1	66.7	13.6	81.2	36.4	85.7	15.9
frequency	10k	HiT	**65.1**	**7.5**	**74.5**	14.8	63.5	12.9	**52.4**	64.8
	10k	HW	-	**0**	-	0	100	11.1	100	88.9
	10k	HW+	-	**0**	-	0	76.9	29.5	80.6	70.5
frequency	5k	HiT	52.5	**2.7**	**74.3**	9.9	68.8	9.7	54.8	77.8
	5k	HW	-	**0**	-	0	100	7.4	100	92.6
	5k	HW	-	**0**	-	0	66.7	13.6	81.6	86.4
centrality	1*	HiT	**56.4**	75.8	**62.4**	10.2	**66.9**	9.7	**56.5**	4.2
	1*	HW	-	-	-	-	-	-	-	-
	1*	HW+	76	56.8	75	27.3	100	15.9	-	0
	2*	HiT	**54.5**	**54.3**	60.2	18.4	63.5	16.5	**63.9**	10.8
	2*	HW	-	-	-	-	-	-	-	-
	2*	HW+	66.7	**34.1**	81.8	25	100	9.1	85.7	31.8

*high polysemy means the term has min. [THRESHOLD] total synsets / total parts of speech
*high centrality means the term has the intended concept as synset number [THRESHOLD] at most

Table 7: WordShiftEval results broken down by the frequency, centrality, and polysemy of the terms that make up the entries.

five senses per part of speech. Depending on the threshold T, the most-observed type of word pair amongst the HiT results is that of two polysemous terms (45.6 % of the result set, $T = 3$), two relatively unpolysemous terms (45.3%, $T = 5$), or equal proportions of the two ($T = 4$).

The proportion of correctly classified term pairs is unequally distributed across polysemy classes, in particular for HiT. Word pairs with two non-polysemous (i.e. relatively unambiguous) terms are consistently more likely to see their shift direction assessed correctly (64-66% correct, depending on the polysemy threshold) than word pairs with two polysemous terms (46-53% correct), which have an almost equal chance of getting correctly or incorrectly classified. Entries with a single polysemous term consistently fall somewhere between these two trends. Compared to HiT, HW and HW+ have notably smaller proportions of polysemous term pairs, with as little as 9.1% polysemous pairs (under $T = 3$) for HW and 14.8% for HW+, as opposed to 45.6% for HiT.

A low corpus **frequency** was expected to negatively impact the results. With a small number of occurrences used to collect (train) the vector representations, these risk being less stable and reliable. We take the frequencies underlying the 1990s eng-all vector corpus as a proxy for the overall frequencies of the terms and use several frequency cut-offs (5k, 10k and 100k). HiT clearly displays more lower-frequent terms than HW and HW+. For instance, under a cutoff value of 100k, HiT has about the same proportion of low-frequent (33.9%) and high-frequent pairs (31.8%), while HW has clearly more high-frequent (48.1%) than low-frequent pairs (18.5%). Also, HiT is the only set that contains entries made up of terms with frequencies under 5k and 10k.

Looking at the results on HiT when both terms were very sparse (under 5k) the assessment is just ad random (52.5%), but with a higher threshold of 10k the sparse pairs were more likely to be correctly classified (65.1%). At the same time, high-frequent term pairs with target and reference term both over 10k instances showed to be difficult to classify (52.4%). Taken together, these findings suggest that while higher-frequency terms are not always more suitable, a minimum number of instances is indispensable for reliable results.

By **centrality** we mean how good a contemporary example of the intended concept a term is. To this end we look at the synset that connects target and reference term. If the (target or reference) term has this sense listed among

its top senses in WordNet, we assume it is exemplary. For the target term this is also an assessment of whether the change took place in (what is now) its primary, second, or in a more distant sense. For instance, we assess the target term *shrewd* in its currently most prevalent sense 'marked by practical hardheaded intelligence' (synset `astute.s.01`). While the reference term *astute* is central to this concept, *sharp* is not, as it is only the sixth sense listed for this term. Hence *shrewd-astute* might be a better example of the shift than *shrewd-sharp*. HW was excluded from the analysis, as the terms in it are not related through WordNet synsets.

We consider two cut-off points: one that examines just the first sense and a less strict one that includes the second listed sense. For the former, the (rare) word pairs in HiT that were made up of two strong terms (4.2%) surprisingly had the same proportion correct (56.5%) as the much larger group of word pairs (75.8%) with two weak terms (56.4% correct). This might be an artefact due the small sample size, as the groups with a single strong term did show higher accuracies (62.4% and 66.9%) than those with none. Under the looser definition of centrality, the accuracy of the shift assessment on HiT increases with the centrality of the terms involved, from 54.5% on weak pairs up to 63.9% on strong pairs. HW+ displays the same trend. However, with a much higher proportion of weak term pairs and a lower proportion of strong pairs than HW+, the HiT results are more at risk of centrality effects.

7 Conclusion

This work offers the largest and most realistic dataset for assessing sense change to date, HiT, which provides 3624 English word pairs and 979 word-synset pairs. HiT is made available along with this publication (click here or look for SenseShiftEval on GitHub) and can be automatically extended with more entries. Our experiments have given a number of insights. Firstly, they show how brittle the state-of-the-art method really is. When applied to HiT rather than to small sets of handcrafted examples, the state-of-the-art performance drops dramatically. The error analysis shows in what way existing evaluation data are privileged, if not to say biased: they contain fewer polysemous terms, fewer terms that are less exemplary for the intended concept, and fewer terms modelled on a low number of examples in the corpus. All of these are factors inherent to natural language, which a robust model of sense change will need to handle. The analysis showed that these factors indeed hindered our ability to assess shift direction. For this reason, the two corpus-independent factors, polysemy and centrality, will be incorporated as features in the dataset, to be able to select more or less challenging entries and to assess the effect of these factors on the outcomes.

Complementary to the findings above, several studies have demonstrated that noise is inherent to distributional approaches and stems from factors both computational - e.g. cross-temporal vector alignment (Dubossarsky et al., 2017) - and fundamental, by the mere variance found in natural text corpora (Hellrich and Hahn, 2016). Experimental validation was not the focus of this paper, but we would encourage follow-up work with more rigid experimental checks, including control conditions and non-aligned (e.g. see Dubossarsky et al. (2019)) or count-based vectors.

Given the presence of noise, it is crucial to cross-check findings. HiT is unique in that it caters for this with multiple synonymous entries per target term. We have presented a number of ways to derive holistic, sense-level insights. Some aggregations were more promising than others. The term pair with the largest and most significant cosine trend often displayed the predicted trend. However, averaging the vector representations of all synonyms did not sufficiently cancel out noise.

A logical next step would be to exploit lexical factors for sense-level evaluations, i.e., to select the most representative term pair of a synset based on its centrality to the concept and its (lack of) ambiguity. A preliminary experiment on HiT showed that selection by centrality outperforms some other evaluation techniques. This will be the topic of follow-up work.

Acknowledgements

This work was supported in part by VRE4EIC under the Horizon 2020 research and innovation program, grant agreement No 676247, and by the Netherlands Organization of Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWO) as part of VENI grant 275-89-029 awarded to Antske Fokkens. We thank Pia Sommerauer, Aysenur Bilgin, and the anonymous reviewers for their invaluable feedback.

References

Pierpaolo Basile and Barbara McGillivray. 2018. Exploiting the web for semantic change detection. In *International Conference on Discovery Science*, pages 194–208. Springer.

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM.

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

Mark Davies. 2015. Corpus of Historical American English (COHA).

Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. *arXiv preprint arXiv:1906.01688*.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1136–1145.

Christiane Fellbaum. 2012. Wordnet. *The Encyclopedia of Applied Linguistics*.

John Rupert Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*, pages 67–71. Association for Computational Linguistics.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Johannes Hellrich and Udo Hahn. 2016. Bad companyneighborhoods in neural embedding spaces considered harmful. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2785–2796.

Johannes Hellrich and Udo Hahn. 2016a. Bad companyneighborhoods in neural embedding spaces considered harmful. In *COLING (16)*, page 27852796.

Adam Jatowt and Kevin Duh. 2014. A framework for analyzing semantic change of words across time. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 229–238. IEEE Press.

Christian Kay, Jane Roberts, Michael Samuels, Iren Wotherspoon, and Marc Alexander. 2019. The historical thesaurus of english.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. *ACL 2014*, page 61.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. ACM.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. *arXiv preprint arXiv:1806.03537*.

Andrey Borisovich Kutuzov, Elizaveta Kuzmenko, and Anna Marakasova. 2016. Exploration of register-dependent lexical semantics using word embeddings. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 26–34.

Carlos Martinez-Ortiz, Tom Kenter, Melvin Wevers, Pim Huijnen, Jaap Verheul, and Joris Van Eijnatten. 2016. Design and implementation of shico: Visualising shifting concepts over time. In *HistoInformatics 2016*, volume 1632, pages 11–19.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. 2014. Thats sick dude!: Automatic identification of word sense change across different timescales. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1020–1029.

Matthias Orlikowski, Matthias Hartung, and Philipp Cimiano. 2018. Learning diachronic analogies to

analyze concept change. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 1–11.

Yael Ravin and Claudia Leacock. 2000. Polysemy: an overview. *Polysemy: Theoretical and computational approaches*, pages 1–29.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of computational approaches to diachronic conceptual change. *arXiv preprint arXiv:1811.06278*.

Nina Tahmasebi and Thomas Risse. 2017. Word sense change testset. This work has been funded in parts by the project "Towards a knowledge-based culturomics" supported by a framework grant from the Swedish Research Council (2012–2016; dnr 2012-5738). This work is also in parts funded by the European Research Council under Alexandria (ERC 339233) and the European Community's H2020 Program under SoBigData (RIA 654024).

Nina Tahmasebi, Thomas Risse, and Stefan Dietze. 2011. Towards automatic language evolution tracking, a study on word sense tracking. In *Joint Workshop on Knowledge Evolution and Ontology Dynamics*.

Xuri Tang. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering*, 24(5):649–676.

Xuerui Wang and Andrew McCallum. 2006. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433. ACM.

Gerhard Wohlgenannt, Ariadna Barinova, Dmitry Ilvovsky, and Ekaterina Chernyak. 2019. Creation and evaluation of datasets for distributional semantics tasks in the digital humanities domain. *arXiv preprint arXiv:1903.02671*.