Semantic Proximity Annotation for Lexical Change and Word Sense Discovery

Andreas Johansson
Language Technology Resources HT24 Course Paper
Code: github.com/ajejohansson/LTRproject

Abstract

The paper reports the findings of a semantic-proximity annotation project meant for word-sense and lexical-change discovery. It annotates individual uses of *crippled*, *disabled*, *lame*, *depressed* for semantic similarity and finds sense clusters based on these annotations. Previous annotations for *stab* and *record* are additionally reannotated for validation. The study finds the following: decreased *crippled* for human disability alongside the inverse for *disabled*; loss of physical *depressed*, continued use of its abstract sense, and increased reference to mental depression; increase of abstract *lame* both in a general sense and to describe impotence.

1 Introduction

Word sense disambiguation and similar tasks have traditionally involved the prior categorisation of a word into senses, into which any given use of that word must fit (McCarthy et al., 2016). In other words: senses are predefined, at least for any given task, and language use must fit inside. In the last decade, however, there has been an increasing trend to let language use dictate the lexical categories we define (e.g., ibid., Erk et al. 2009, Schlechtweg et al. 2021). This is of particular benefit to lexical change detection, since it fundamentally requires the assumption that word senses are not fixed. One approach to discover use-guided word senses is the Word Usage Graph (WUG) annotation paradigm (Schlechtweg et al. 2020 and other publications), which employs semantic proximity judgments between word uses to statistically infer word senses of variant discreteness. WUGs can additionally be utilised comparatively across time periods to create diachronic WUGs (DWUGs, Schlechtweg et al. 2021.)

This project aims to carry out a small-scale study on word-sense and lexical-change discovery by using WUG annotation and cluster generation for the following adjectives: *crippled*, *disabled*, *lame*, *depressed*. It will attempt to answer these research questions:

- What senses are discovered by WUG clustering?
- How neatly do the uses fit into the cluster-defined senses, and what annotation choices might affect this outcome?
- What lexical change can be gleaned from the diachronic comparison of the WUGs?

2 Materials and resources

2.1 Data

The data for the target words and their contexts is from Cleaned Corpus of American English (CCOHA, Alatrash et al. 2020), which is based on Corpus of American English (COHA, Davis 2012). COHA is a ~406 million word diachronic corpus spanning publications from 1810 to 2009 of fiction, nonfiction, magazines, and newspapers from a variety of subgenres and domains. It is balanced across decades to include a stable proportion of tokens from its genres, subgenres, and domains. CCOHA is a cleaned version of COHA, which has remedied some problems with e.g., malformed and inconsistent tokens, lemmas, and pos-tags. The full content of COHA (and CCOHA, which is under the same license) is not open-source and only freely-accessible as a limited web-interface, not as a downloadable and thus easily processed dataset, so this project uses a subset of CCOHA that was originally utilised in SemEval 2020 Task 1 (Schlechtweg et al. 2020). The sample covers two time periods spanning 50 years each, separated by 100 years: 1810-1860 and 1960-2010, henceforth *period 1*, and *period 2*, respectively. Each period is represented by 6 million tokens, which are line-separated and exist in both raw-token and lemmatised forms. The original target words of the SemEval task retains its pos-tag, but the sample is otherwise untagged.

To validate the annotation of the target words of the present paper, annotations and contexts for WUG generation in Schlechtweg et al. (2021) are also used. These more or less adhere to the format for the DURel tool (Schlechtweg et al. 2024), with some additional optional columns, but had to be processed to produce new files for DURel interfacing. The full file structure can be found in the accompanying repository and a variant of the structure is described somewhat in section [insert preprocessing section]. This data was originally sourced from the same SemEval CCOHA sample as described for the target words above.

Naturally, the pre-specified periods of the sample limits the interval that can be investigated for lexical change. It also plays a somewhat restrictive role in target-word selection since ideally words that are hypothesised to have undergone lexical change across or between the periods should be chosen. However, because the employed annotation style is quite labour intensive (see section 3.2) the smaller corpus should not be a bottleneck.

2.2 Tools

Annotation is carried out on the DURel (Diachronic Usage Relatedness) platform (Schlechtweg et al. 2024a). The tool facilitates a semantic proximity annotation task where annotators are presented with two text contexts, each with a highlighted target word. They must give a similarity rating between the target words in each given context. The canonical task definition, carried out in e.g. Schlechtweg et al. 2021, 2024b and also employed in the present project, has the target words in each sentence be the same lemma, but this is not a technical restriction. The ratings from which annotators may choose are found in table 1, left column. These ratings are derived from a cognitive semantic proximity scale by Blank (1997), shown in the right column.

Table 1: DURel annotation categories (left) and Blank's semantic proximity scale which the annotation categories analogise (right)

Annotation category	**Blank (1997)**
4. Identical	Identity
3. Closely Related	Context Variance
2. Distantly Related	Polysemy
1. Unrelated	Homonymy
0. *Cannot decide*	-

Following is a brief description of the annotation categories (examples inspired by DURel annotation guidelines, accessed 2025):

4. Semantically identical uses in each context. Does not regard morphology and syntax, e.g. tense of a verb or predicative versus attributive adjectives should not inherently affect the rating, but some such features can be indicative of a semantic difference.
3. Similar usages with minor but identifiable sense differences. E.g., *I saw* ***children*** *running around* and *I saw my* ***child*** *come running*, where the meaning of the former target word is roughly *young person*, whereas the latter has a familial relation component between child and speaker.
2. Not unrelated senses but notably different meanings. A typical example is the comparison of a metaphorical use with a nonmetaphorical one, e.g., a literal *crossroads* versus a decision-point referred to as *crossroads*.
1. Semantically unrelated, e.g., a *riverbank* versus a *financial bank*.
0. To be selected if the similarity can not be determined, e.g., if the meaning of the target in one context cannot be determined.

DURel outputs WUG directories, which are filesystems that can be used with another resource employed in the project: the WUGs GitHub repository (Schlechtweg et al. 2021, 2024a, Schlechtweg 2023). The repository implements a variety of sense-clustering algorithms. The one used in the present project is a simple unsupervised algorithm that produces a graph **G = (U, E, W)**. A graph node $u \in U$ is an individual use (context with a marked target token or tokens). $e \in E$ is a pair of uses (u_i, u_j), called an *edge*, and $w \in W$ is a weight that represents the similarity score of an edge (Schlechtweg et al. 2021, p. 7081). The algorithm searches for a clustering $C : U \rightarrow L$ (each use assigned a label) that minimises the sum of *positive inter-cluster* edge weights clusters and *negative intra-cluster* edge weights (Schlechtweg 2024a, p. 14381). Simpler put: C is penalised for putting similar U elements in different clusters and dissimilar elements in the same cluster. To define positive and negative *w*, a threshold parameter *h* must be defined. *h* should be within the 1-4 annotation scale and effectively determines the threshold for what should be considered a different sense. The default 2.5 (ratings 1 and 2 = negative, 3 and 4 = positive) is retained for the present project. Finally, the number of clusters is not predetermined, but a maximum number needs to be set for the algorithm to eventually stop searching. This parameter is set high enough for it not to be expected to have an impact on the clustering.

3 Methodology

The primary focus of this paper is the annotation of new words and the utilisation of those annotations. Given that these annotations will only be carried out by one annotator, a preliminary validation step is carried out, wherein previously annotated edges are reannotated, and the resulting judgments are correlation-tested with the mean and median previous annotations. Next, new words and edges are annotated. Both rounds of annotations use the DURel tool and scale, and both require individual preprocessing (section 3.1) and data exploration (3.2) steps. These steps can be viewed in full detail in the accompanying code repository. Finally, the new annotations are input into the clustering algorithm described in section 2.2.

3.1 Preprocessing

The data from the previous round of annotations is in tsv files of two types: *usefiles* and *judgmentfiles*. Each investigated word has one file of each type. Some words had files that were not easily loadable due to formatting issues, and since the objective of this step is to select only a few words, not utilise all the data, these were simply filtered out. Usefiles contain this data per row: lemma, the corpus grouping (a number identifying which period the use is from), a unique identifier for the use, the text context, start-and-stop character indices for the context (part or whole of the provided text), start-and-stop indices for the target word, and some additional categories that are not used in the present project. Judgment file rows have the lemma, two identifiers corresponding to the unique identifiers in the usefile, annotator id, judgment (semantic proximity annotation), and some additional information. Since each row of the judgment file represents only one annotation, the annotation data per edge is spread out. Given the purpose to use these annotations for a correlation metric, the judgment data was aggregated per edge into a format that retained the mentioned data (as these are required for DURel interfacing), discarded unused data, discarded the judgment column in favour of mean and median judgment, and added annotation count data. An annotation of 0 is non-rating, categorically different from the rest; it is not just one less than 1 in the way 2 is one less than 3. Therefore, rows with a 0-rating are not used in any mentioned metrics. Each row in the new format contained a unique edge, and these edges were sorted by annotation count such that if a word were selected for reannotation, the *n* edges that would be selected are those with the most annotations. These edges end up with only 3-5 annotations, which is still better for correlation testing than many other edges.

Original usefiles are already unique per row as they are basically lookup-tables for the judgment files, and they are processed only to remove superfluous (for the project) data and to filter out uses that are not among those in selected judgment files.

For the new annotations, lines containing candidate target lemmas are identified in the lemma-version of the corpus (see 2.1). The corresponding raw line is tokenised and pos-tagged with nltk, and any (token, tag) pair that does not match both prespecified token forms and nltk tags are filtered out. For surviving uses, values for previously mentioned usefile data are retrieved and stored. Unlike for the reannotations, the new uses do not require an accompanying edge file, as DURel can produce the edges automatically if no specific pairings are sought.

The pre-annotation stage also included some amount of data exploration, code for which can be found in the repository.

3.2 Annotation

Reannotation is done for *record*, and *stab* (noun pos for both). New annotation is as noted carried out for the adjectives *crippled*, *disabled*, *lame*, and *depressed*. These four words were not the original targets since previous annotation data only exists for verbs and nouns; new targets were aimed to be as analogous to the validation examples as possible. A preliminary round of annotation was meant to be carried out for *strike* (with the intention to do both noun and verb) but was interrupted due to notable pos-mistagging. The switch to the adjectives resulted in very good pos-tagging, likely in large part due to lack of pos-ambiguity of the lemmas; for example, *disabled* in the lemmatised version of the corpus can only be an adjective (assuming good lemmatisation) even if the form can also be a verb. The result was (qualitatively assessed) good precision as there were no pos-mistakes among the target word samples. Whether recall was as good cannot be assessed with the present data. However, precision is the much more important metric given the definition of the annotation task, where only a small number of quality contexts are required.

For the reannotation words, 25 edges each are annotated. Many more annotations are done for the new words, as these are the focus of the study. In its ideal form, this is a quite annotation hungry task. To make the clusters as reliable as possible, as many combinations as possible should be annotated, and to do all edges, $n*(n-1)/2$ (combinations, without replacement) are required. Moreover, a properly representative sample requires a large number of uses. Finally, while not in the scope of the present project, multiple annotations from different people per edge is ideal. On the basis that spurious edge correlations would make analysis less feasible, the chosen method is to include a smaller number of uses but with full edge annotation. This means the results will be on something akin to a toy-sample of language, but it will make the project more self-contained. For a more representative project, some edge sampling would have to be made, since use combinations increase nonlinearly.

10 words are extracted from the corpus per word per period. Each use is combined with each other use (both cross- and within period). This means 190 (20*(20-1)/2) annotations per word, except for *crippled*, which only occurs 8 times in each period for 120 annotations. Combined with the reannotations, this is 740 annotations total.

3.3 Clustering

Clustering is carried out by reassembling some of the pipeline of the mentioned WUGs repository and plugging in the DURel output directory for each word into it, and then follows the procedure outlined in 2.2. The outcome are three cluster plots per word (one per period and one combined) where the cluster-labels (senses) are clearly demarcated. The semantic change detection lies in the difference between the two periods.

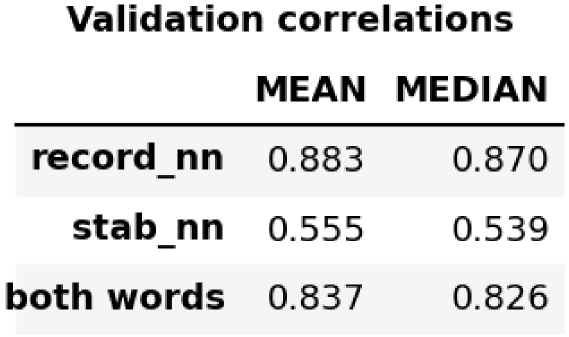
4 Results and analysis

This section displays the results in the form of correlation metrics (4.1) and cluster plots (4.2). It also includes analysis and discussion in order to keep the analysis less separated from its subject.

4.1 Reannotation

Table 2 shows the Spearman ranked correlation between the reannotations and previous annotations. *Record* has strong correlation whereas *stab* is more lacking. However, upon some post-hoc reflection, the chosen correlation metric seems somewhat too punishing, at least when reannotations are put against aggregations. On a small scale with discrete values (1,2,3,4), any deviation can greatly affect the correlation, and the aggregation metrics being non-discrete can cause such deviation just by virtue of their construction. For example, 2 or 3 is as close to perfect as possible for a reannotation of an edge originally labelled

Table 2: Correlation between reannotations and mean and median previous annotations for *record* (noun), *stab* (noun), and both combined.



Validation correlations

	MEAN	MEDIAN
record_nn	0.883	0.870
stab_nn	0.555	0.539
both words	0.837	0.826

(2,2,3,3) by four annotators, but it will still be .5 off both the mean and median.

Schlechtweg et al. (2021) compare annotators pairwise for how often and by how much they diverge. This cannot be replicated at this stage; the closest thing is to check how much and often the reannotations diverge from the median. Discounting edges where the median ends with a .5, *stab* reannotations have full agreement 82% of the time, is one off 18%, and is never off by more. Including the .5 medians, this changes to 56/44%, suggesting the correlation with an aggregate has a large effect.

4.2 New annotations

Figures 1–4 show WUGs for each word. Each figure has three plots: an overview plot with all use-nodes (top) to determine the discovered senses, a plot for period 1 (middle) and one for period 2 (bottom). All nodes in the overview are represented in the period plots, and cluster membership is based on the overall data, only visualised separately in the period plots. Positive edges (3, 4 rating) are represented by bolded lines, and negative edges (1, 2) by faded lines. Different node colours (and integer label) represent different clusters/proposed senses. The same plots in better resolution and larger size are available as html files in the repository. Opened in browser, they are also interactive, where mousing over a node will show the represented use. In text, the clusters will be referred to by their colour, as the node labels are quite small in the provided plots.

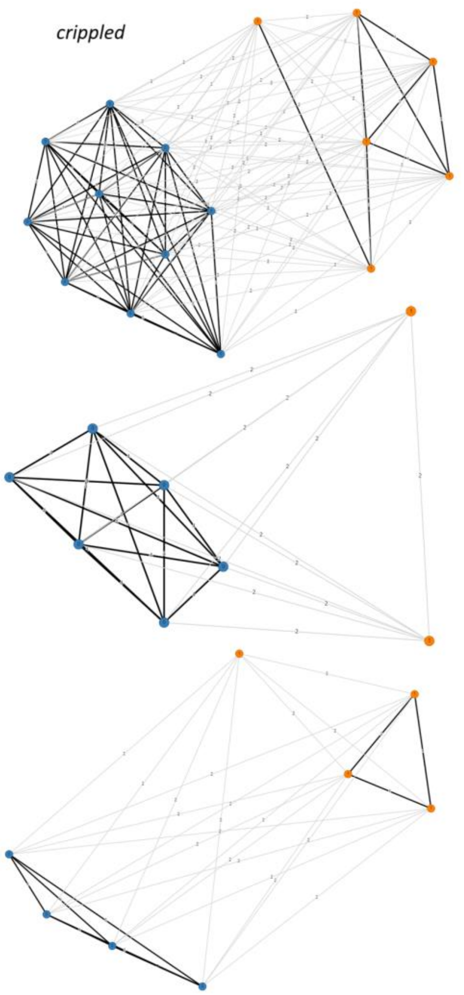


Figure 1: WUGs for *crippled*. Overview (top), period 1 (middle), period 2 (bottom).

Crippled has two clusters. First is one referring in some way to physical disability (blue, example 1). The other refers to a crippled state of something inanimate (orange, 2). The latter is perhaps a metaphorical application of the former that can be paraphrased as *broken* without losing much meaning, which is not the case for the blue sense.

1) …as a box boy in supermarkets, as a chauffeur for a **crippled** woman, and even served for a time as a G.1.
2) …certain that he could have run the distance faster than the **crippled** little aircraft was flying.

There is also a pair of orange nodes (with a 4 rated edge between them) tied together to the cluster by only one positive edge with the rest of the cluster. These two are somewhat more abstract (*crippled bank* and *crippled rhythm*), and would likely receive its own cluster with slightly different annotation. The tie to the rest of the cluster is only due to an edgecase *crippled antigravity*, which was deemed similar enough to both the rest of the cluster (physically broken objects) and the crippled bank.

Between time periods, crippled gets proportionally less likely to be used in the physical-disability sense and moves towards more inanimate/abstract use, although the human use does not disappear.

Disabled has the following clusters: one best summarised as *human disability* (blue, 3), one *nonfunctional object* (orange, 4) and a one-off node referring to children mentally freezing up out of fear (green).

3) …Social Security Act was temporarily amended to allow the **disabled** to earn more than $ 280 a month without forfeiting their government assistance.
4) …when Perry left his almost conquered and **disabled** ship…

Across periods, only the human-disability cluster remains in the second, suggesting semantic change toward *disabled* being the preferred choice for human reference. Paired with the decline of *crippled* for human reference, this indicates that pejoration of *crippled* made *disabled* take its place. However, note that the data sampling in this project (find *n* uses, then stop) inherently ignores raw frequency data, which would be required to make this conclusion better grounded.

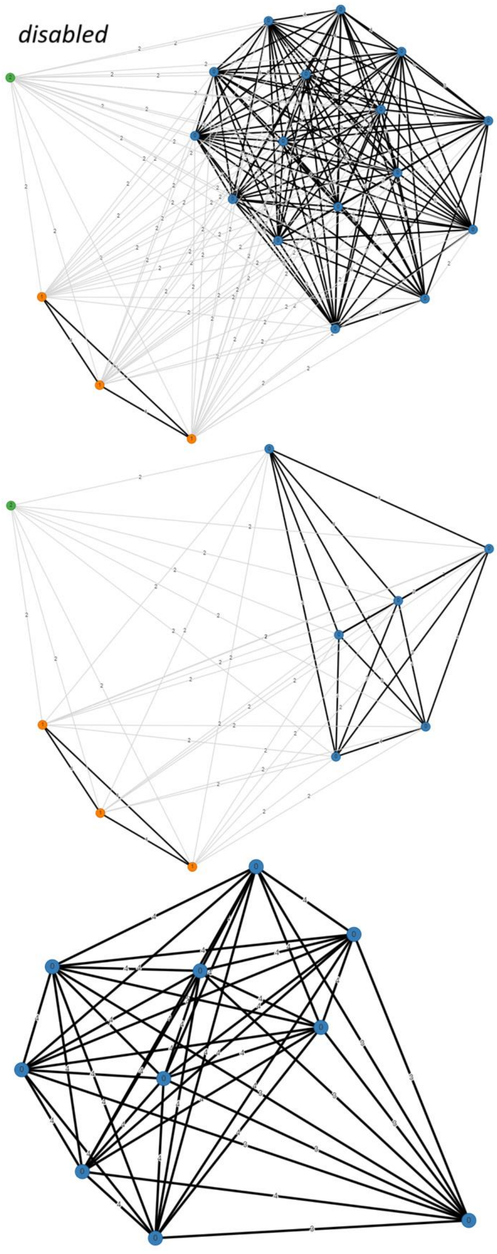


Figure 1: WUGs for *disabled*. Overview (top), period 1 (middle), period 2 (bottom).

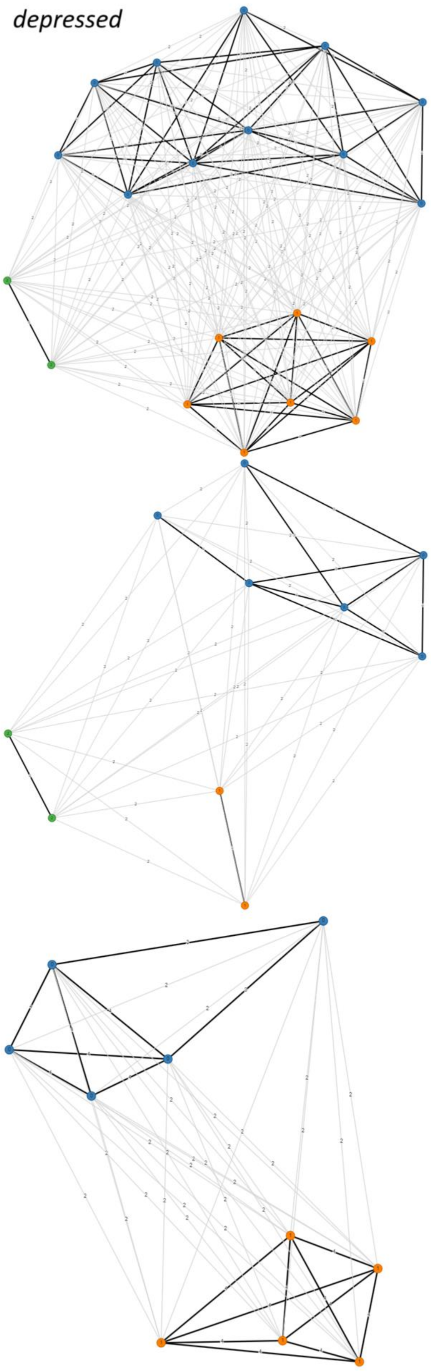


Figure 3: WUGs for *depressed*. Overview (top), period 1 (middle), period 2 (bottom).

Also of note is that other annotation heuristics might split the human-disability cluster. Currently, permanent and (the relatively few) temporary disability uses are clustered together. For example, a reference to the *disabled list* (a list of currently injured athletes) could be argued to be closer to the *nonfunctional* sense, just with a human referent; the referent has a function (as a boat or as an athlete) that it currently is disabled from performing. This is not quite how permanent disability is usually conceptualised. Whether this is enough to warrant a 2 (different sense) instead of a 3 (same sense, broad enough to cover the context variance) is an open question. The threshold parameter *h* could also be adjusted to change this grouping, but this would likely have a notable impact on the rest of the clustering.

Depressed has three clusters: economic depression (blue, 5), mental depression (clinical or mood, orange, 6), and physical depression (green, 7).

5) They offer tax incentives to encourage development in **depressed** neighborhoods.
6) Kidman 's postmarital roles have included a clinically **depressed** writer…
7) … between two waves, we shall have from twenty-five to thirty feet as the utmost altitude which any swell of water can have, reckoning from the most **depressed** portions of the surface near it.

The physical-depression sense has disappeared between periods, the economic sense is close to the same, and the mental sense has notably increased. *Economic* is arguably a misnomer for its cluster, as it also contains some abstract uses that could be paraphrased as *lowered*, i.e., a metaphorical application of the physical sense. However, the economic usage could likely be considered a type of that same metaphorical use that has since become more conventionalised as its own distinct (sub)sense. As in the case with *disabled* above, it is open to interpretation whether this distinction should warrant its own sense.

Finally, the algorithm does not find more than one cluster for *lame*, although it does separate the nodes into three rough groupings: physical lameness (e.g., ‘a **lame** leg’, top right of the overview plot), the modern abstract usage that can (at least sometimes) be paraphrased as ‘uncool’ (e.g., ‘a **lame** skateboard trick’, left in overview plot), and political impotence (e.g., ‘**lame** duck’, bottom right). These groups are largely separated by negative edges, with some exceptions. One is an

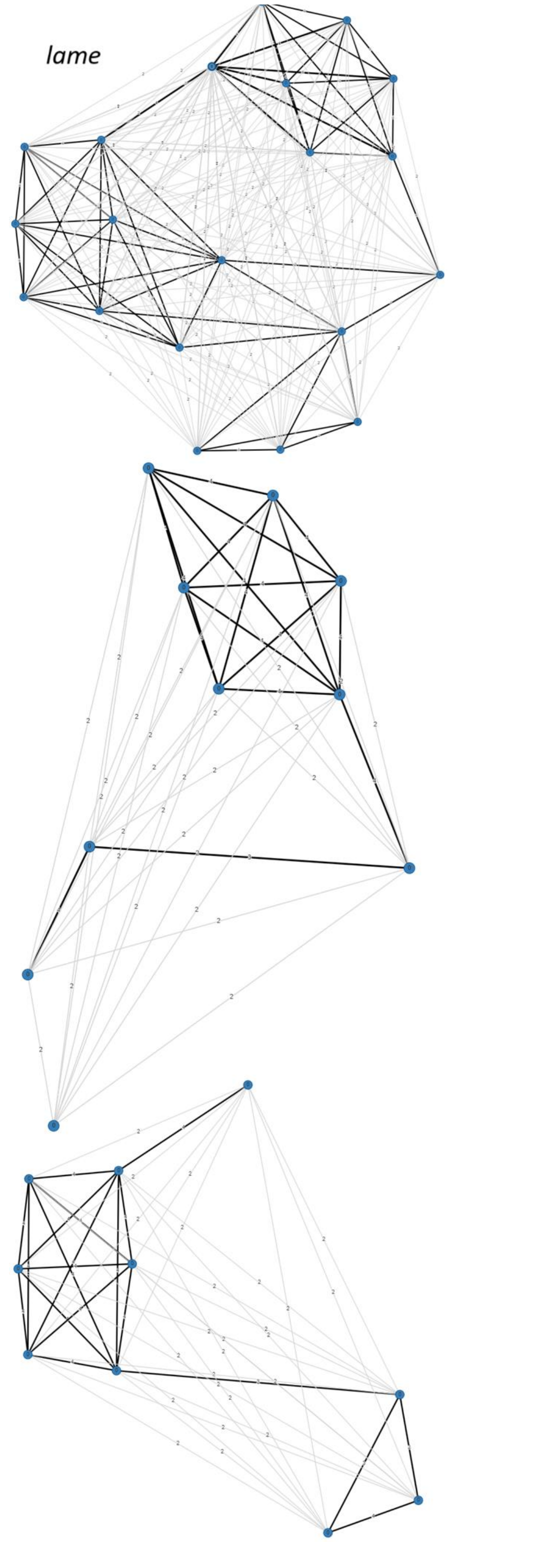


Figure 4: WUGs for *lame*. Overview (top), period 1 (middle), period 2 (bottom).

annotation error in the top left that should not have a 4 rating. The rest is a number of nodes in the middle/bottom right that have positive edges with only a few members of other groups. What these have in common is a surviving connotation of impotence that seems to have survived the metaphorical application of the physical sense. Viewed across periods, this is the group of abstract uses that existed at all in period 1, which was otherwise dominated by physical *lame*. Both the general abstract sense and the political impotence sense seem to have developed since period 1.

5 Ethics

The most notable ethics consideration for this study is the inclusion of some target words that can be used derogatorily. This has indirect ties with the structure of the used data. While COHA's genre balance across time is a positive for any diachronic study, it must be borne in mind that this means more modern data sources (e.g., web-based platforms) are unutilised. The project's results might thus give a false impression that, for example, *crippled* is no longer used to describe disability, when this is likely false on certain corners of the internet. It is stressed here that the findings of this paper only reflect the used data.

6 Conclusion

This paper has presented annotation-based WUGs for four adjectives and used these to discover sense clusters. These graphs have been used to infer sense development across time. Some limitations of the applied approach should be noted. First, the means of ascertaining diachronic development is simply a snapshot each of two time periods, with semantic change only noted as a different between the periods. This removes any granularity of e.g. how fast any change occurred. The inclusion of more periods could remedy this, at the cost of heavier annotation burden. Second is a restatement of the reality that this study has too few uses to properly represent the data, let alone English more generally. Third is a reiteration of the paper's ethics section, with the addendum that the lack of modern data sources might also reflect a more general lack of representative data. Finally, while the annotation scale used in the study served to create relatively noiseless clusters, part of this might be attributable in part to the low number of ratings. A more granular annotation scale could be considered, although this might require a different clustering algorithm.

Future work could seek to carry out research that applies the same processes on entirely different data. While internet sources do not have the large timescale differences we see in the present dataset, certain internet language domain are likely to change fast enough for these diachronic techniques to be valuable.

References

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2020. CCOHA: Clean Corpus of Historical American English. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6958–6966, Marseille, France. European Language Resources Association. Subset accessed from University of Stuttgart: https://www.ims.uni-stuttgart.de/en/research/resources/corpora/sem-eval-ulscd-eng/

Andreas Blank. 1997. *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*. Niemeyer, Tübingen.*

*(did not read Blank directly since I do not speak or read German, but I still wanted to include the original source)

Mark Davies 2012. Expanding Horizons in Historical Linguistics with the 400-Million Word Corpus of Historical American English. *Corpora.* Vol. 7 (2): 121–157 Edinburgh University Press.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2009. Investigations on Word Senses and Word Usages. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 10–18, Suntec, Singapore. Association for Computational Linguistics.

Diana McCarthy, Marianna Apidianaki, and Katrin Erk. 2016. Word Sense Clustering and Clusterability. *Computational Linguistics*, 42(2):245–275.

Dominik Schlechtweg. 2023. Human and Computational Measurement of Lexical Semantic Change. PhD thesis. University of Stuttgart.

Dominik Schlechtweg, Pierluigi Cassotti, Bill Noble, David Alfter, Sabine Schulte Im Walde, and Nina Tahmasebi. 2024a. More DWUGs: Extending and Evaluating Word Usage Graph Datasets in Multiple Languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14379–14393, Miami, Florida, USA. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Dominik Schlechtweg, Shafqat Mumtaz Virk, Pauline Sander, Emma Sköldberg, Lukas Theuer Linke, Tuo Zhang, Nina Tahmasebi, Jonas Kuhn, and Sabine Schulte Im Walde. 2024b. The DURel Annotation Tool: Human and Computational Measurement of Semantic Proximity, Sense Clusters and Semantic Change. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 137–149, St. Julians, Malta. Association for Computational Linguistics.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.