# Visualising Changes in Semantic Neighbourhoods of English Noun Compounds over Time

**Malak Rassem, Myrto Tsigkouli, Chris Jenkins, Filip Miletić, Sabine Schulte im Walde**
Institute for Natural Language Processing, University of Stuttgart, Germany
{malak.rassem, myrto.tsigkouli}@outlook.com
{christopher.jenkins, filip.miletic, schulte}@ims.uni-stuttgart.de

## Abstract

This paper provides a framework and tool set for computing and visualising dynamic, time-specific semantic neighbourhoods of English noun-noun compounds and their constituents over time. Our framework not only identifies salient vector-space dimensions and neighbours in notoriously sparse data: we specifically bring together changes in meaning aspects and degrees of (non-)compositionality.

## 1 Introduction

Noun compounds (NCs) such as *fairy tale* and *gold mine* represent a type of multiword expressions (MWEs) whose meanings are semantically idiosyncratic to some degree, i.e., their meanings are not necessarily fully predictable from the meanings of their parts (Partee, 1984; Sag et al., 2002; Baldwin and Kim, 2010). While the restricted compositionality of NCs has been explored extensively and across research disciplines from synchronic perspectives (Mitchell and Lapata, 2010; Reddy et al., 2011; Schulte im Walde et al., 2013, 2016; Baroni et al., 2014; Cordeiro et al., 2019; Garcia et al., 2021; Miletić and Schulte im Walde, 2023, i.a.), the field is still lacking an adequate amount of empirical large-scale approaches towards diachronic models, in order to explore the emergence and changes of compound meanings over time. Up to date, Dhar et al. (2019) and Dhar and van der Plas (2019) exploited the Google $n$-gram corpus and information-theoretic as well as cosine distance measures to predict the compositionality of the compounds in Reddy et al. (2011), and to detect novel compounds, respectively. Maurer et al. (2023) investigated whether diachronic developments of the frequencies and productivities of the compounds and their constituents in Cordeiro et al. (2019) are salient indicators of the NCs' present-day degrees of compositionality, and Mahdizadeh Sani et al. (2024) applied standard cosine-based measures of (dis)similarity to the same compounds and constituents over time.

The current study contributes to the so far limited computational models and insights on the diachronic development of NC compositionality. We provide a framework and a tool set for computing and visualising semantic neighbourhoods of English NCs over time. By focusing on semantic neighbours we explicitly target changes in the salient meaning aspects of NCs; more specifically, we bring together semantic neighbourhoods of NCs and their constituents, thus aligning changes in meaning aspects and degrees of compound-constituent (non-)compositionality. A major side-challenge is concerned with identifying an appropriate set of vector-space dimensions, both regarding the semantic interpretations of the dimensions and regarding the notorious sparse-data problem in historical corpus data that strongly affects compound representations. The contributions of this paper are the following.

- **Semantic Space:** A carefully crafted semantic vector space to represent those 195 noun-noun compounds and their constituents from Cordeiro et al. (2019) that occurred in all time slices of the cleaned corpus of historical American English CCOHA (Davies, 2012; Alatrash et al., 2020).
- **Semantic Neighbours:** Semantic neighbourhoods for compounds and their constituents, both (i) time-specific and dynamic as well as (ii) static present-day representations.
- **Temporal Compound-Constituent Visualisation Tool:** An adaptation of a deterministic approach to multi-dimensional scaling and two-dimensional plotting (Hilpert, 2016; Tsigkouli, 2021) to the vector-space representations of compounds, constituents and semantic neighbourhoods.



*Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 240–246

The semantic spaces and neighbours of our English compounds and constituents, as well as the visualisation tool which is applicable to also further compound and constituent targets in English and additional languages, are publicly available from https://www.ims.uni-stuttgart.de/data/dia-neighbour-nn.

## 2 Data

### 2.1 Corpus: CCOHA

As our diachronic text corpus resource, we used the clean version of the Corpus of Historical American English, referred to as CCOHA (Davies, 2012; Alatrash et al., 2020), in order to ensure that the dataset is free from inconsistent lemmas, malformed tokens and other anomalies that could potentially affect the analyses. We then reduced the fine-grained part-of-speech tags in CCOHA to a coarser-grain set of tags, for example, collapsing all variants of nouns like singular common noun (NN1), plural common noun (NN2), singular locative noun (NNL1), etc. under a single broad noun tag NN to generalise the tokens' part-of-speech (POS) tags. A full list of the mapping of the POS tags can be found in Appendix A. To analyse changes over time, the data was segmented into specific timeslices. The selected timeslices are: 1810–1829, 1830–1859, 1860–1889, 1890–1919, 1920–1949, 1950–1979, and 1980–2009, with each range being inclusive.

### 2.2 Noun Compound (NC) Targets

Our goal is to investigate the semantic evolution of noun compounds across different historical periods, focusing specifically on the noun-noun compounds identified by Cordeiro et al. (2019). Out of the 210 noun-noun[1] compounds mentioned in their work, 195 are present in our corpus. We consider both space-separated and dash-separated compounds, treating equivalents like *credit card* and *credit-card* as identical entities for our analytical purposes. In order to exclude compounds with more than two constituents, we imposed a restriction on the POS tag patterns; namely, the tokens immediately preceding and succeeding a noun-noun target compound must not be tagged as nouns (NN) for the sequence to qualify as a two-part noun-noun compound.

[1]We disregarded noun compounds with other than nominal modifiers (such as adjective-noun compounds).

## 3 Semantic Space and Neighbours

### 3.1 Semantic Vector-Space Creation

As the backbone of our semantic space for plotting compounds as well as their constituents and semantic neighbours over time, we identified a set of semantic space points (SSPs). These SSPs were defined as nouns appearing with a frequency $>$500 in the entirety of the CCOHA, i.e., not just within individual timeslices. The threshold was set to ensure a substantial enough occurrence for meaningful semantic analysis. Then the top 50 most frequent nouns were excluded from the SSPs to eliminate potential semantic hubs (Radovanović et al., 2010; Dinu et al., 2015) that could dominate the analysis due to their high rate of occurrence, given that they typically represent semantically generic terms. Our criteria resulted in identifying 9,345 unique nouns that served as SSPs for further analysis.

For all noun compounds, their constituents and all SSPs, we computed timeslice-specific co-occurrences (TSCs) within a $\pm$10-word window. These TSCs were further refined by limiting the context words to those tagged with the reduced POS content tags: nouns (NN), verbs (VV), adverbs (RR), and adjectives (JJ). The TSCs were then transformed into vectorised formats to enable further processing. This conversion entails mapping the co-occurrence data into numerical vectors, with each dimension corresponding to a specific context word. The magnitude in each dimension was determined by the frequency of each context word's co-occurrence with the noun compounds, constituents or SSPs within the defined timeslice.

We chose to use simple frequency counts for co-occurrences rather than alternative association measures (Evert, 2005) due to the complexities and potential mathematical incorrectness involved. Specifically, measures such as variants of mutual information would require division by the total number of all co-occurrences of the targets we are dealing with. In our case, this would mean coalescing the noun compounds, their constituents and the SSPs together. However, doing so would lead to double counting, because constituents may also function as SSPs. Moreover, there is considerable overlap between the co-occurrences of compounds and those of their constituents or SSPs. For instance, the co-occurrences for the compound *wedding day* are essentially the identical subset of its constituents, which are also SSPs, thereby leading to redundancy in our counts.

| Target | Timeslice | 5 Nearest Neighbours |
|---|---|---|
| credit card | 1830–1850 | — |
| | 1920–1940 | rationing, gallon, shuttle, questionnaire, invitation |
| | 1980–2000 | reservation, card, cash, credit, check |
| credit | 1830–1850 | exchange, money, bank, account, circulation |
| | 1920–1940 | loan, bank, account, banker, reserve |
| | 1980–2000 | card, visa, account, cash, greeting |
| card | 1830–1850 | game, paper, trick, minute, stranger |
| | 1920–1940 | paper, game, ball, box, trick |
| | 1980–2000 | check, credit, paper, line, trick |

Table 1: The five nearest neighbours of the compound *credit card* and its constituents, across timeslices.

### 3.2 Semantic Neighbourhoods

Using cosine (dis)similarity, we compared the TSC vector representations of the noun compounds to those of the SSPs within the same timeslice, in order to quantify their semantic proximity. For each time-specific compound, the five most similar neighbours from the pool of SSPs were identified based on the cosine scores. For example, we can see in Table 1 that the compound *credit card* did not appear in the corpora in earlier timeslices, suggesting that there was no established sense for the compound at that time. Subsequently, the neighbours of *credit card* include written documents, reflecting the term's initial use to denote means of payment such as traveller's cheques.[2] The neighbours in more recent periods transition to the modern sense associated with *cash.*

Following Hamilton et al. (2016), we used as a static semantic space the TSC vectors of the last timeslice of these neighbours, and did the same for those of the compounds' constituents. This approach allows us to capture the evolving relationships between words over time while maintaining temporally fixed reference points for comparison.

## 4 Temporal Compound Visualisation

We implemented two methods to visualise time-specific compounds in semantic space.

### 4.1 Own-Vector Approach

In the own-vector method, we created a single matrix for each compound using its TSC vectors at every timeslice and the TSC vectors of its constituents and neighbours only from the last "static" timeslice. We then applied metric multidimensional scaling (MDS) to this matrix, which we prefered over non-deterministic approaches such as t-SNE due to its determinism, and derived two-dimensional vector representations for plotting, as previously done by Hilpert (2016) and Tsigkouli (2021). Although the own-vector approach seems to be the most intuitive, we found it to produce objectively sub-optimal plots, where the compounds tend to cluster together and away from the SSPs regardless of the timeslice.

### 4.2 Projected-Compound Approach

In this refined method, a single matrix is created using the latest (static) timeslice TSC vectors of all compound neighbours and constituents, but excluding the compound's own TSC vectors. As in the own-vector approach, we derived the coordinates of the neighbours and constituents by applying MDS to this matrix. For the compound, however, rather than using the compound's own TSC vectors to determine the time-specific coordinates, these vectors were computed as the weighted averages of the respective five time-specific nearest neighbours' coordinates, with the weights being their cosine scores. The intuition behind this approach was that in the own-vector approach the SSPs' TSC vectors and the compounds' TSC vectors consistently clustered away from each other and could not efficiently be visualised together, which we attribute to the severe sparsity in the compound vector representations. In contrast, our refined approach projects a compound's semantic change over time by reflecting its relative positions to its neighbours' semantic fields, thus improving over the sparsity issue. Consequently, the method produces plots that more distinctly illustrate the temporal semantic shifts of noun compounds. For example, the trend regarding *credit card* and its neighbours that we described

[2] https://www.etymonline.com/word/credit-card

above based on Table 1 is rather clear in the plot in Figure 1. Likewise, in Figure 2 we observe *gold mine* starting from its literal compositional sense in the earlier timeslices (i.e., the actual mine), where it's surrounded by its constituents, and in later times moving towards SSPs such as *money* and *business*. This shift highlights the development of an additional metaphorical sense of *gold mine* in the later timeslices, as a symbol of value.

## 5 Conclusion

This study used a corpus-based computational approach to examine the semantic evolution of noun compounds in historical American English, thus contributing to the field of diachronic computational linguistics by providing a methodologically robust tool set for analyzing temporal changes in compound semantics. Future research could expand upon this foundation by exploring other types of multiword expressions.

## Limitations

We presented experiments on visualising the temporal evolution of noun compound meanings as captured by high-dimensional semantic vectors. The obtained results strongly depend on the choice of vector space representations and dimensionality reduction methods. We opted for interpretable and deterministic approaches given our linguistic motivation, and with this constraint we explored different implementation variants and presented the most robust systems. Some other combination of experimental settings – including non-deterministic methods – may improve on our results.

More generally, our vector space representations are directly dependent on the properties of the underlying corpus, which is additionally affected by sparsity issues (like most diachronic datasets). A different set of texts may capture different aspects of the target words' semantics; a larger corpus may yield more robust vector representations. Moreover, our experiments are limited to the American English data at our disposal. Due to typological differences in the linguistic realisation of multiword expressions such as noun compounds, our method may not produce equivalent results for other languages or language varieties.

## Ethical Considerations

We do not believe that this paper raises ethical issues. We conducted a linguistic analysis of empirically attested data using well-established methods to computationally represent word meaning. Note though that our bottom-up approach automatically induces the semantic neighbours for a specified target word. We therefore cannot exclude the possibility of inadvertently outputting offensive content or depicting societal biases captured by our corpus, which covers American English usage over the course of two centuries. However, we did not encounter these issues in closely inspected results; we also note that they are inherent in any large-scale corpus analysis.

## Acknowledgments

The research presented here was supported by the DFG Research Grant SCHU 2580/5 (*Computational Models of the Emergence and Diachronic Change of Multi-Word Expression Meanings*). We also thank Martin Hilpert for providing the starting point for the MDS/R visualisation.

## References

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2020. CCOHA: Clean Corpus of Historical American English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6958–6966, Marseille, France. European Language Resources Association.

Timothy Baldwin and Su Nam Kim. 2010. Multiword Expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, pages 267–292. CRC Press, Boca Raton, USA.

Marco Baroni, Raffaella Bernardi, and Roberto Zamparelli. 2014. Frege in Space: A Program for Compositional Distributional Semantics. *Linguistic Issues in Language Technologies*, 9(6):5–110.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised Compositionality Prediction of Nominal Compounds. *Computational Linguistics*, 45(1):1–57.

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

Prajit Dhar, Janis Pagel, and Lonneke van der Plas. 2019. Measuring the Compositionality of Noun-Noun Compounds over Time. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 234–239, Florence, Italy. Association for Computational Linguistics.

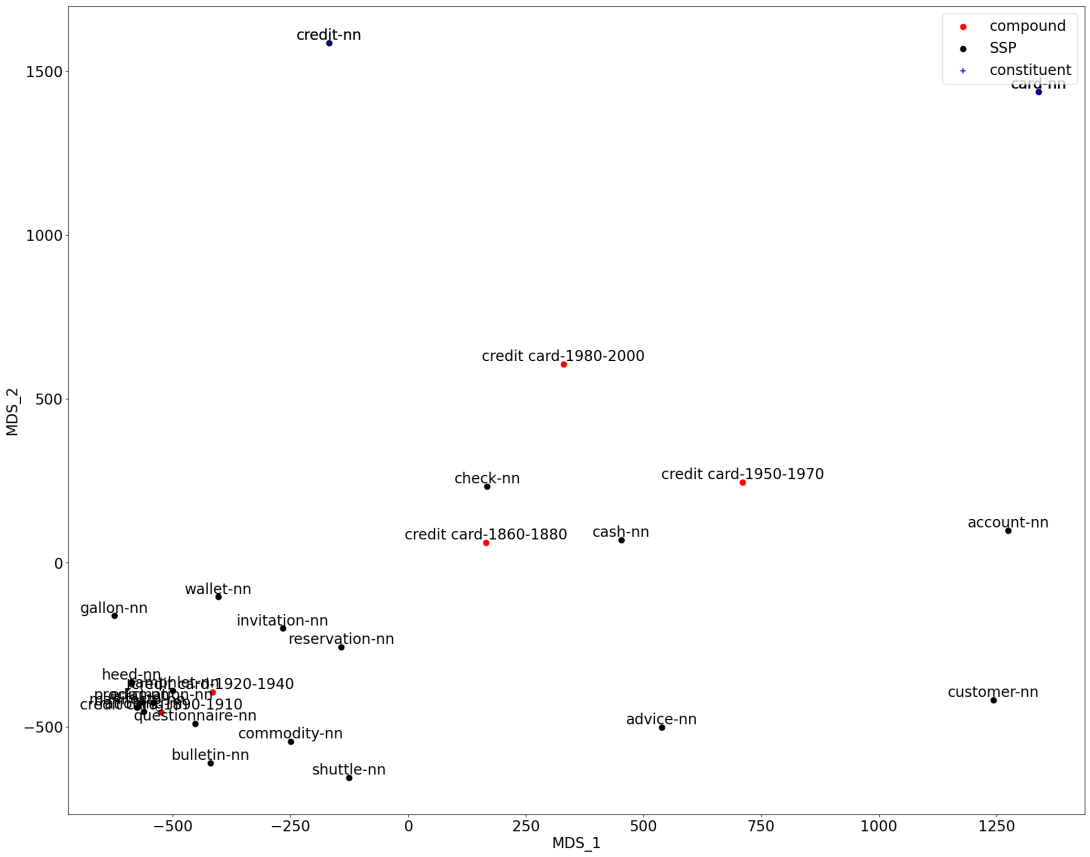


Figure 1: Example of the projected-compound approach on the compound *credit card*.

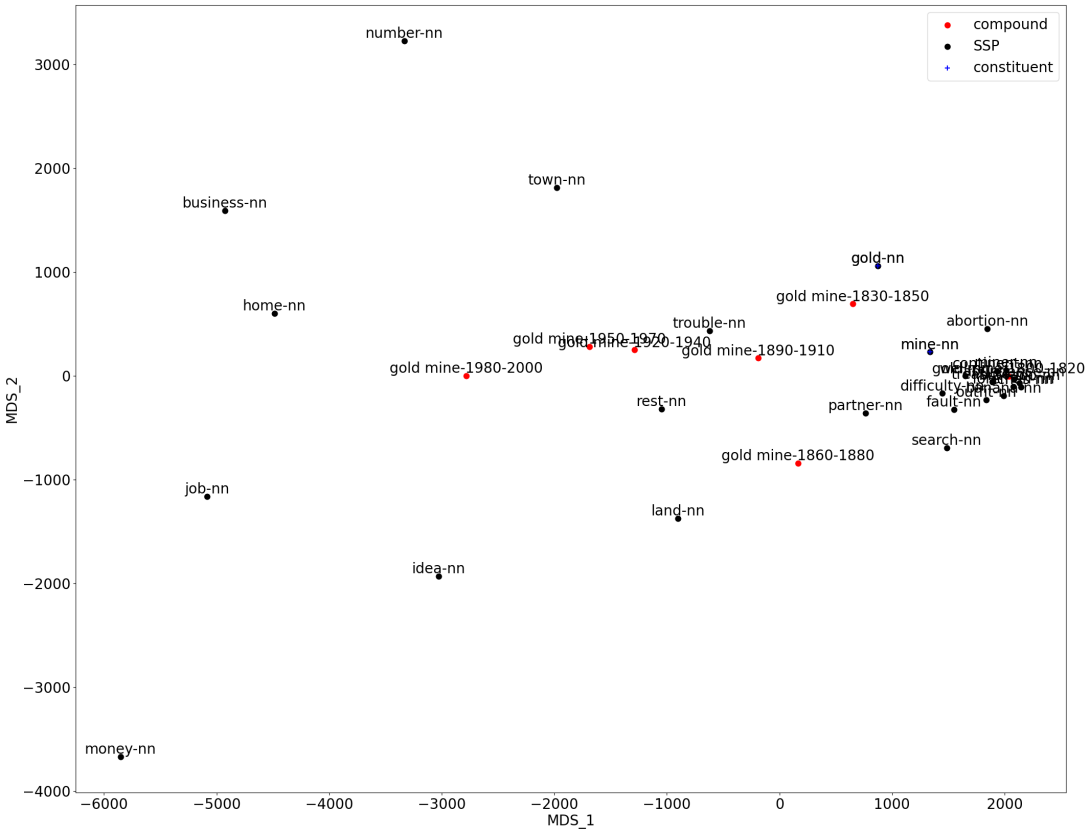


Figure 2: Example of the projected-compound approach on the compound *gold mine*.

Prajit Dhar and Lonneke van der Plas. 2019. Learning to Predict Novel Noun-Noun Compounds. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet*, pages 30–39, Florence, Italy. Association for Computational Linguistics.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving Zero-Shot Learning by Mitigating the Hubness Problem. In *Proceedings of the International Conference on Learning Representations, Workshop Track*, San Diego, CA, USA.

Stefan Evert. 2005. *The Statistics of Word Co-Occurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Germany.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Assessing the Representations of Idiomaticity in Vector Models with a Noun Compound Dataset Labeled at Type and Token Levels. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 2730–2741, Online.

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*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Martin Hilpert. 2016. Change in Modal Meanings: Another Look at the Shifting Collocates of *may*. *Constructions and Frames*, 8(1):66–85.

Samin Mahdizadeh Sani, Malak Rassem, Chris Jenkins, Filip Miletić, and Sabine Schulte im Walde. 2024. What Can Diachronic Contexts and Topics Tell Us About the Present-Day Compositionality of English Noun Compounds? In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 17449–17458, Torino, Italy. European Language Resources Association and International Committee on Computational Linguistics.

Maximilian Maurer, Chris Jenkins, Filip Miletić, and Sabine Schulte im Walde. 2023. Classifying Noun Compounds for Present-Day Compositionality: Contributions of Diachronic Frequency and Productivity Patterns. In *Proceedings of the 19th Conference on Natural Language Processing*, pages 40–51, Ingolstadt, Germany.

Filip Miletić and Sabine Schulte im Walde. 2023. A Systematic Search for Compound Semantics in Pretrained BERT Architectures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1499–1512, Dubrovnik, Croatia. Association for Computational Linguistics.

Jeff Mitchell and Mirella Lapata. 2010. Composition in Distributional Models of Semantics. *Cognitive Science*, 34:1388–1429.

Barbara H. Partee. 1984. Compositionality. In Fred Landman and Frank Veltman, editors, *Varieties of Formal Semantics: Proceedings of the 4th Amsterdam Colloquium*, pages 281–311. Foris Publications.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data. *Journal of Machine Learning Research*, 11:2487–2531.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An Empirical Study on Compositionality in Compound Nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. Association for Computational Linguistics.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, Mexico.

Sabine Schulte im Walde, Anna Hätty, and Stefan Bott. 2016. The Role of Modifier and Head Properties in Predicting the Compositionality of English and German Noun-Noun Compounds: A Vector-Space Perspective. In *Proceedings of the 5th Joint Conference on Lexical and Computational Semantics*, pages 148–158, Berlin, Germany. Association for Computational Linguistics.

Sabine Schulte im Walde, Stefan Müller, and Stephen Roller. 2013. Exploring Vector Space Models to Predict the Compositionality of German Noun-Noun Compounds. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics*, pages 255–265, Atlanta, GA, USA. Association for Computational Linguistics.

Myrto Tsigkouli. 2021. Studying the Diachronic Changes of Constituent Collocates and Meanings in English Noun Compounds. Master's thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Germany.

## A Part-of-Speech Tag Reduction

Table 2 presents the mapping of the original part-of-speech tags into a reduced, coarser set of tags.

| Tag Description | Original Tag | Reduced Tag |
|---|---|---|
| singular noun of direction (e.g. *north, southeast*) | ND1 | NN |
| common noun, neutral for number (e.g. *sheep, cod, headquarters*) | NN | NN |
| singular common noun (e.g. *book, girl*) | NN1 | NN |
| plural common noun (e.g. *books, girls*) | NN2 | NN |
| following noun of title (e.g. *M.A.*) | NNA | NN |
| preceding noun of title (e.g. *Mr., Prof.*) | NNB | NN |
| singular locative noun (e.g. *Island, Street*) | NNL1 | NN |
| plural locative noun (e.g. *Islands, Streets*) | NNL2 | NN |
| numeral noun, neutral for number (e.g. *dozen, hundred*) | NNO | NN |
| numeral noun, plural (e.g. *hundreds, thousands*) | NNO2 | NN |
| temporal noun, singular (e.g. *day, week, year*) | NNT1 | NN |
| temporal noun, plural (e.g. *days, weeks, years*) | NNT2 | NN |
| unit of measurement, neutral for number (e.g. *in, cc*) | NNU | NN |
| singular unit of measurement (e.g. *inch, centimetre*) | NNU1 | NN |
| plural unit of measurement (e.g. *ins., feet*) | NNU2 | NN |
| singular weekday noun (e.g. *Sunday*) | NPD1 | NN |
| plural weekday noun (e.g. *Sundays*) | NPD2 | NN |
| singular month noun (e.g. *October*) | NPM1 | NN |
| plural month noun (e.g. *Octobers*) | NPM2 | NN |
| base form of lexical verb (e.g. *give, work*) | VV0 | VV |
| past tense of lexical verb (e.g. *gave, worked*) | VVD | VV |
| -ing participle of lexical verb (e.g. *giving, working*) | VVG | VV |
| -ing participle catenative (*going* in *be going to*) | VVGK | VV |
| infinitive (e.g. *work* in *It will work*) | VVI | VV |
| past participle of lexical verb (e.g. *given, worked*) | VVN | VV |
| past participle catenative (e.g. *bound* in *be bound to*) | VVNK | VV |
| -s form of lexical verb (e.g. *gives, works*) | VVZ | VV |
| general adjective (e.g. *old, good, strong*) | JJ | JJ |
| general comparative adjective (e.g. *older, better, stronger*) | JJR | JJ |
| general superlative adjective (e.g. *oldest, best, strongest*) | JJT | JJ |
| catenative adjective (*able* in *be able to*) | JK | JJ |
| adverb, after nominal head (e.g. *else, galore*) | RA | RR |
| adverb introducing appositional constructions (e.g. *namely*) | REX | RR |
| degree adverb (e.g. *very, so, too*) | RG | RR |
| wh- degree adverb (*how*) | RGQ | RR |
| wh-ever degree adverb (*however*) | RGQV | RR |
| comparative degree adverb (*more, less*) | RGR | RR |
| superlative degree adverb (*most, least*) | RGT | RR |
| locative adverb (e.g. *alongside, forward*) | RL | RR |
| prep. adverb, particle (e.g *about, in*) | RP | RR |
| prep. adv., catenative (*about* in *be about to*) | RPK | RR |
| general adverb (e.g. *always, typically*) | RR | RR |
| wh- general adverb (*where, when, why, how*) | RRQ | RR |
| wh-ever general adverb (*wherever, whenever*) | RRQV | RR |
| comparative general adverb (e.g. *better, longer*) | RRR | RR |
| superlative general adverb (e.g. *best, longest*) | RRT | RR |
| quasi-nominal adverb of time (e.g. *now, tomorrow*) | RT | RR |

Table 2: Part-of-speech tag reduction mapping.