Stefano De Pascale and Weiwei Zhang

Scoring with Token-based Models

A Distributional Semantic Replication of Sociolectometric Analyses in Geeraerts, Grondelaers, and Speelman (1999)

Abstract: This paper provides a replication of sociolectometric analyses found in Geeraerts, Grondelaers, and Speelman (1999) with the help of distributional semantic modelling. We selected 14 concepts from the lexical field of football in Dutch and Chinese respectively. Instead of manually disambiguating the corpus occurrences, we explored a semi-automatic procedure based on token-based vector space models and cluster analysis. The experiments show that our workflow is efficient for detecting regional lexical variation in large-scale corpora. More specifically, the results revealed that removing semantic clusters whose most central members are tokens referring to other senses rather than the intended concept's sense, does have an impact on the sociolectometric distances. Furthermore, discarding entire clusters has consequences for the total concept frequency.

Keywords: sociolectometry, distributional semantics, football concepts, Dutch, Chinese

Acknowledgement: This project received funding from a Marie Skłodowska-Curie Individual Fellowship of the EU's Horizon 2020 research and innovation programme (No. 793920), and from the KU Leuven Research Fund C1 (No. 3H150305).

1 Introduction

In 1999, Dirk Geeraerts, Stefan Grondelaers, and Dirk Speelman published *Convergentie en divergentie in de Nederlandse woordenschat: een onderzoek naar kleding- en voetbaltermen* (henceforth GGS1999), which effectively launched the

Stefano De Pascale, KU Leuven, Blijde-Inkomststraat 21 box 3308, Leuven, Belgium, e-mail: stefano.depascale@kuleuven.be
Weiwei Zhang, KU Leuven & Shanghai International Studies University, e-mail: weiwei.zhang@kuleuven.be

https://doi.org/10.1515/9783110733945-016

field of "sociolectometry". In that work they laid out both the foundation for the approach and conducted the first comprehensive corpus-based quantification of synchronic and diachronic distances between Dutch varieties over the past 50 years. Although the lexical sociolectometric enterprise started in GGS1999 predates the institutionalization of Cognitive Sociolinguistics, its focus on corpus-based empiricism and its concept-centered perspective allow for a natural embedding within the scope of this volume's framework.

The novelty of GGS1999 was the aggregation of multiple lexical variables, as opposed to impressionistic observations of single variables. Those lexical variables can be thought of as sociolinguistic variables in the Labovian sense. Concretely, they looked at 30 concepts from 2 lexical fields, clothing and football, and for each concept the several near-synonyms found to refer to those concepts. The quantification of distances between Dutch varieties was then operationalized as the differences between the frequency distributions of the near-synonyms in one variety compared to the other. Such frequency distributions of near-synonyms have been called "onomasiological profiles". For broader overviews and studies of the various X-lectometric approaches see Wieling and Nerbonne (2015) and contributions in Rosseel, Franco, and Röthlisberger (2020).

Since the publication of GGS1999, many research projects at the KU Leuven research unit QLVL have been devoted to specifically lexical sociolectometry (Speelman, Grondelaers, and Geeraerts 2003; Peirsman 2010; Ruette 2012; Ruette et al. 2014; Geeraerts 2018; De Pascale 2019). The increasing availability of larger corpora since the seminal study has posed a crucial and recurring challenge; namely, how to introduce more advanced methods in order to scale up and shift the manual treatment of lexical variables to their semi-automatic processing. In early studies, the choice of lexical variants ("types") for a concept and the selection of the occurrences ("tokens") of these variants expressing the given concept, was done by manually scanning the available resources.

In this contribution we report on the most recent methodological advances made to tackle these lexical-semantic issues in sociolectometric studies. We specifically address the problem of identifying corpus occurrences of lexical variants that instantiate the chosen concepts, and, conversely, discarding occurrences that express other senses, by making use of *token-based vector space models*. We explore the relevance of token-based models by looking back at a historic dataset, that is, the football concepts first investigated in GGS1999.

2 Token-based Vector Space Models in Lexical Sociolectometry

Vector-based models have already been introduced in sociolectometric research by the dissertations of Peirsman (2010) and Ruette (2012) for the retrieval of near-synonymous variants. The models employed in said work were so-called *type-based models*, because they formalized the meaning of a lemma (i.e. a type) as a vector of the co-occurrence frequencies between that lemma and all its context words. Given that words with similar context distributions share similar meanings, this was the go-to method for the semi-automatic retrieval of near-synonymous variants (e.g.: *tv, television, tube*). The growing appeal of vector-based, distributed meaning representations is evident from recent overview articles, such as Lenci (2018) and Boleda (2020).

Yet the one-vector-per-word formalization is unable to represent the full scope of semantic variation inherent to most words. Clearly, for near-synonyms that are highly polysemous, which means that they are used to express other senses (labeled "out-of-concept", e.g.: *tube* as 'long, hollow cylinder') next to the one intended by the given concept (labeled "in-concept", e.g.: *tube* as 'television'), the removal of large numbers of tokens that are out-of-concept is necessary in order to arrive at correct sociolectometric distances, where the frequency distributions are not polluted by token counts of multiple senses. For the task of disambiguating near-synonymous tokens a more fine-grained semantic representation is needed. *Token-based vector space models* serve precisely this goal.

Token-based models also make use of the surrounding context of an individual token, but instead of counting the frequency of those target-context co-occurrence events and storing them in a vector, they rely on the type-based vectors of those first-order context words. For the single representation of a token's meaning, the context word vectors are eventually summed or averaged. An extensive discussion of token-based models can be found in Heylen et al. (2015) and De Pascale (2019).

Our token-based replication will in some respects be faithful to the original design and in others show differences. First, we will not only look at the lexical field in Dutch, but also the corresponding concepts in Mandarin Chinese. Since sociolectometry was conceived as a way to gain insights into pluricentric languages (see also Soares da Silva (2014) on Portuguese), an incursion into yet another pluricentric landscape is therefore important for the validity of the framework. For the sake of comparison with Dutch we will focus on two standard varieties of Mandarin: Mainland *Pǔtōnghuà* and Taiwan *Guóyǔ*.

Second, the underlying materials are different: while in GGS1999 counts were taken from sports magazines published around 1990, here we make use of large-scale corpora. For Dutch we rely on two subsets of the Leuven News Corpus (Ruette 2012) and the Twente News Corpus (Ordelman et al. 2007), as representative corpora for the Belgian Dutch (BE) variety and the Netherlandic Dutch (NL) variety respectively (totaling about 520 million tokens, from 1999 to 2005). For Chinese, we took subsets of the Mainland Chinese (ML) and Taiwan Chinese (TW) newspaper sections of the Tagged Chinese Gigaword Version 2.0 corpus (Huang 2009) to ensure a regionally balanced corpus of 250 million words for each language variety, sampled between 1990 and 2004. For the time being we will assume that all results pertain to one single synchronic period (i.e. the 1990s), even in the absence of full temporal overlap between all the corpus sources, and differences in sociolectometric distances will not be interpreted as real-time changes.

3 Visualizing and Partitioning the Semantic Structure of Concepts

To get an idea of the pervasive, but problematic role of polysemy, we will take a look at the token-based models for the concept COUNTERATTACK in Dutch and Chinese[1]. As single token vectors do not mean much to a human eye, the meaning captured by a token vector can only be determined by virtue of its relation to the vector of other tokens. For that purpose, we calculate pairwise cosine similarity scores between all tokens, apply a dimensionality reduction technique like t-SNE (Van Der Maaten and Hinton 2008) on the resulting similarity matrix, and visualize that reduced token space. The original token space counts as many dimensions as there are tokens, and reducing the dimensions to a number that is conceivable by the human mind has the benefit that one can readily see which semantic structure emerges from the token model.

The last step before the actual sociolectometric calculations is to partition that token space, provided that the regions in the token cloud correspond to meaningful groups of tokens. For this purpose, we carry out a cluster analysis, and choose to divide up the space in 8 clouds, irrespective of the concept, the language, or the amount of tokens. Even though this might sound as an unjustly

1 The whole list of concepts and their variants in Dutch and Chinese can be found at: https://osf.io/2e4bd/.

coarse practice, the primary goal of the use of token-based models for sociolectometric analyses is to simply discard out-of-concept tokens, not to arrive at the most precise semantic classification. By setting the number of clusters high enough, we aim at a relatively granular partitioning of the token space so that we minimize the risk of ending up with a larger cluster that might lump together otherwise separated regions of out-of-concept and in-concept tokens.

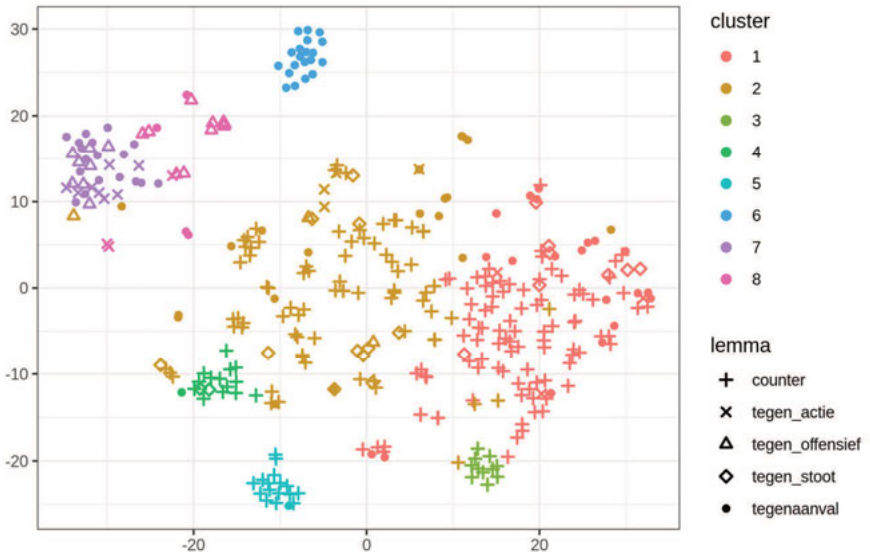

Fig. 1: Token cloud of Dutch COUNTERATTACK

Figure 1 and 2 show the t-SNE generated token clouds for the concept COUNTERATTACK in both languages. The shapes code the different near-synonymous variants and the colors the corresponding clusters in which the tokens have been classified. The clusters are unequal in composition and in size: for example, in the Dutch token clouds clusters 3, 4, 5 and 6 are small and have a predominant or exclusive presence of just one variant, clusters 1 and 2 are large and lexically heterogenous, and cluster 7 is small but contains several near-synonyms as well. The Chinese token cloud in turn distinguishes roughly two core semantic regions, one with clusters 1, 4 and 6 and the other one with clusters 2, 5, 7 and 8.

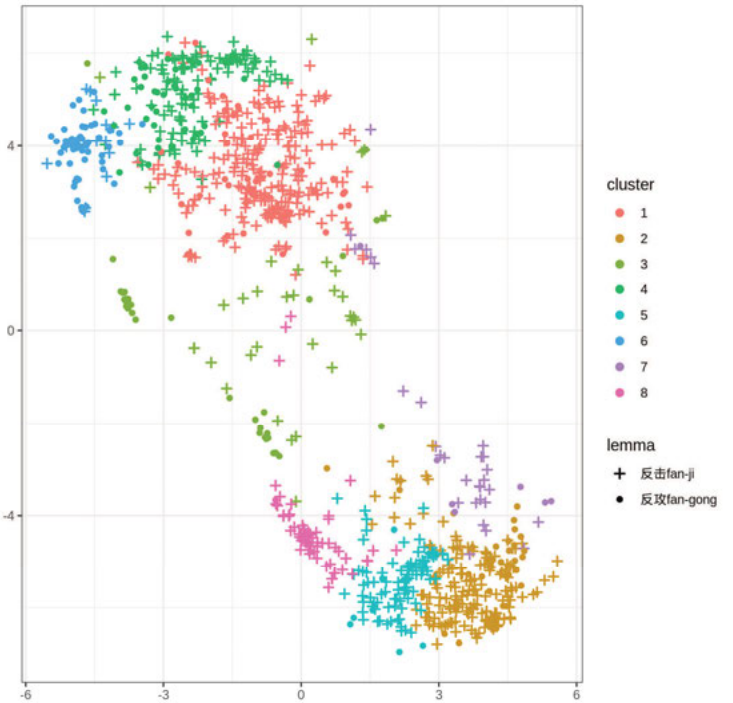

Fig. 2: Token cloud for Mandarin Chinese COUNTERATTACK

For each cluster we manually checked 10% of its tokens, thereby reducing the workload associated with manual disambiguation. Furthermore, we first sorted the tokens so that that 10% would correspond to the most central members of the clusters, using the silhouette width (Rousseeuw 1987) as a measure of centrality of a token in a cluster. Clusters with more than 20% out-of-concept tokens among the ones checked were discarded for the sociolectometric analysis. Inspection of those clusters reveals that monolexical clusters, where we do not observe variation between variants and therefore expect an out-of-concept sense, do not always pose problems. For the Dutch data, clusters 3, 4 and 5 all contain *counter-* tokens instantiating the concept's meaning, but monolexical cluster 6 also contains many non-football related *tegenaanval* tokens and was therefore discarded. It is also risky to assume that clusters with several variants automatically coincide with the COUNTERATTACK sense: cluster 7, which has *tegenaanval*, *tegenoffensief* and *tegenactie* contains too many out-of-concept tokens, primarily military counteroffensive tokens. Regarding the Chinese data, clusters 1 and 4 are kept as valid clusters, i.e. more than 80% of *fan-gong* and *fan-ji* tokens refer to the

concept's meaning, while clusters 2, 5, 7 and 8 contain *fan-gong* and *fan-ji* tokens pertaining to the domain of "war" or "argument" instead of "football". Cluster 6 mainly contains non-football related *fan-gong* tokens referring to counteroffensive actions in baseball or basketball contexts. Manual inspection therefore remains important, but by applying a cluster analysis and taking into account the structure of a cluster we can dramatically decrease this time-consuming task.

4 Sociolectometric Analyses Based on Token-based Models

After the selection of the in-concept clusters, the sociolectometric distance indices can be calculated. The main index is the *external uniformity value*, which quantifies the difference between onomasiological profiles in two lects. The specific formula can be found in Ruette et al. (2014), but now it suffices to know that the higher the values the higher the similarity in near-synonyms usage between the lects.

The bar plots for the Dutch data in Figure 3 show, for each concept, three quantities: the left black bars refer to the external uniformity values between BE and NL as calculated on the data provided in GGS1999. The middle grey bars show the uniformity values based on the selected token clusters and the right light grey bars quantify the uniformity values based on all tokens, as if we had not done any semantic disambiguation at all. For the Chinese concepts in Figure 4 two uniformity values between ML and TW are shown per concept: the left one on all tokens without any semantic control and the right one on the selected in-concept clusters after semantic disambiguation.

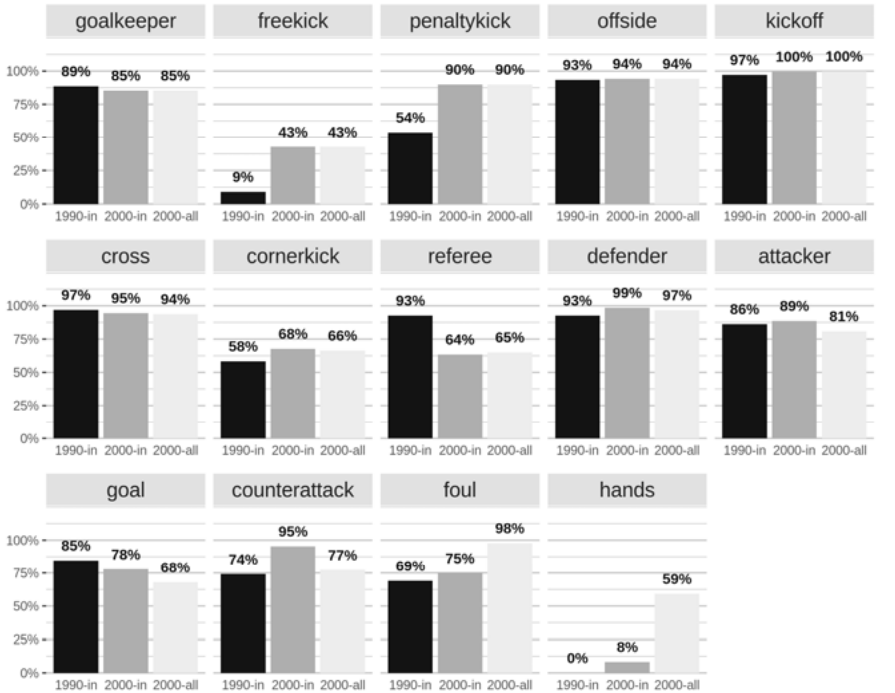

Fig. 3: External uniformity values between Belgian Dutch and Netherlandic Dutch

For a majority of those concepts, no large differences are recorded between semantically "responsible" and semantically naive calculations on the newspaper corpora. For some concepts, like GOAL, COUNTERATTACK, FOUL and HANDS in the Dutch data, and KICKOFF and GOAL in the Chinese data, the discrepancies do exceed the 10%, which we consider a significant deviation between the two calculations. The question now arises as to how exactly, for the abovementioned concepts, the differences in external uniformity values are a consequence of the token-based disambiguation. The scenarios boil down to two factors: the presence of a dominant out-of-concept sense in one variant (in perhaps one lect) and the different lectal frequency distribution of a shared out-of-concept sense. Let us look at the concept HANDS (Table 1): in clusters 1, 2, 3 and 4 we find the variant *handbal* only occurring in its sense 'team game in which the ball is thrown or hit with the hands rather than kicked' and not in the in-concept sense 'touching the ball with the hand or arm, constituting a foul'. As these clusters of *handbal* were large in size, they took up a large portion of the relative token mass for the non-disambiguated profile of HANDS. Removing them caused the relative token mass

in the profile to redistribute to the exclusively Belgian Dutch variant *handspel* and this ultimately led to a very low external uniformity value. In this case we were confronted with an out-of-concept sense that was present in both varieties. Differential token frequency of senses per lect is observed in the case of COUNTERATTACK (Table 2). For that concept, the military or at least non-soccer uses of *tegenaanval*, *tegenoffensief* and *tegenactie* are much more dominant in the Dutch newspaper corpora than in the Belgian ones. Removing these out-of-concept tokens shifts again token mass to another variant, in that case *counter*, giving rise to the observed discrepancy.

Tab. 1: Onomasiological profiles for the concept HANDS in Dutch

	Belgian Dutch before/after disambiguation	**Netherlandic Dutch before/after disambiguation**
handfout	1 (1%) / 1 (2%)	0 (0%) / 0 (0%)
handspel	43 (40%) / 43 (90%)	0 (0%) / 0 (0%)
handbal	59 (55%) / 3 (6%)	75 (73%) / 26 (55%)
hands	5 (4%) / 1 (2%)	28 (27%) / 21 (45%)
	100%	100%

Tab. 2: Onomasiological profiles for the concept COUNTERATTACK in Dutch

	Belgian Dutch before/after disambiguation	**Netherlandic Dutch before/after disambiguation**
counter	130 (70%) / 90 (80%)	81 (50%) / 45 (76%)
tegenactie	5 (3%) / 3 (3%)	13 (7%) / 0 (0%)
tegenstoot	9 (5%) / 5 (4%)	12 (7%) / 5 (9%)
tegenaanval	38 (20%) / 15 (13%)	47 (26%) / 8 (14%)
tegenoffensief	4 (2%) / 0 (0%)	18 (10%) / 0 (0%)
	100%	100%

Similarly, we can deduce the underlying causes for the observed discrepancies between "in" and "all" uniformity values for the Chinese KICKOFF and GOAL in Figure 4. For the concept GOAL (Table 3), cluster 8 contains tokens of the variant *defen* in its TW-specific sense 'scoring more points (runs) than the other team in baseball'. A lectally shared out-of-concept sense is also observed for this concept,

but its frequency in TW and ML is quite different: clusters 1 and 7 contain the variant *de-fen*, which mainly occurs in its non-football sense of 'scores or marks' and this out-of-concept sense is mainly attested in TW; in clusters 2 and 4, we find the variant occurring in the context of basketball instead of football. These clusters of *de-fen* are also large in size. Therefore, like the abovementioned *hand-bal* example in Dutch, discarding them from the profile of GOAL caused a significant change in the external uniformity value. The same can be said for the concept KICKOFF.

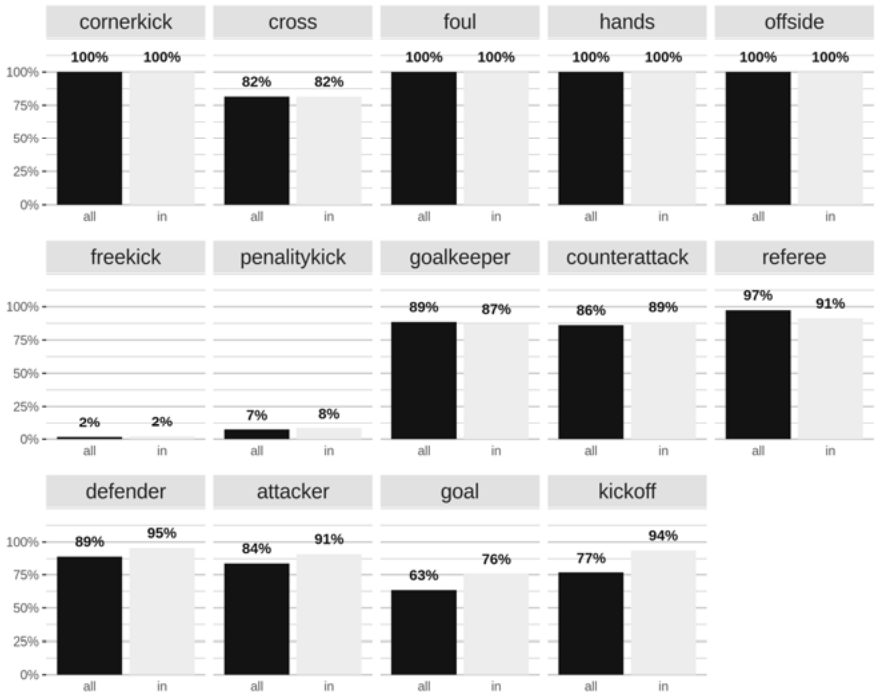

Fig. 4: External uniformity values between Mainland Chinese and Taiwan Chinese

A different, but more relevant way for judging the impact of the cluster-based disambiguation step is provided when considering the reduction of the total mass of tokens in a concept. There are two benefits involved: first, it gives a reasonable indication of the effort that one saves by discarding clusters that do not instantiate the concept's sense, instead of having to look at each token individually. As avoiding time-consuming manual disambiguation was one of the primary drivers for the use of token-based models, this is an important check. Second, a lack of

differences between uniformity values calculated before and after disambiguation does not necessarily imply that we have not been successful in discarding probable out-of-concept tokens. The removal of all tokens of a variant in both lects might accidentally have no impact at all on the relative frequency distributions of the remaining variants in the profiles.

Tab. 3: Onomasiological profiles for the concept GOAL in Chinese

	Mainland Chinese before/after disambiguation	**Taiwan Chinese before/after disambiguation**
de-fen	219 (45%) / 125 (33%)	435 (85%) / 63 (57%)
po-men	99 (20%) / 96 (25%)	9 (2%) / 6 (5%)
jin-qiu	172 (35%) / 156 (41%)	66 (13%) / 41 (37%)
	100%	100%

Tab. 4: Absolute and relative reduction in total token mass after cluster-based disambiguation in the Dutch data

concept (Dutch)	**reduction in total token mass (relative/absolute)**	**final concept frequency**
ATTACKER	- 91%/- 4190	401
FOUL	- 89%/- 4199	511
HANDS	- 55%/- 116	95
COUNTERATTACK	- 52%/- 186	171
GOAL	- 40%/- 3944	6013
REFEREE	- 18%/-410	1833
CROSS	- 18%/- 104	476
KICKOFF	- 13%/- 27	177
DEFENDER	- 11%/- 212	1079
OFFSIDE	- 5%/- 5	99
CORNER KICK	- 2%/- 7	289
GOALKEEPER	/	2731
FREEKICK	/	559
PENALTY KICK	/	1278

Tab. 5: Absolute and relative reduction in total token mass after cluster-based disambiguation in the Chinese data

concept (Chinese)	reduction in total token mass (relative/absolute)	final concept frequency
REFEREE	- 98%/- 5436	88
KICKOFF	- 95%/- 4872	236
HANDS	- 93%/- 598	46
COUNTERATTACK	- 91%/- 4204	420
GOAL	- 91%/- 5088	487
FOUL	- 89%/- 988	127
DEFENDER	- 86%/- 2674	432
ATTACKER	- 81%/- 2864	652
PENALITYKICK	- 33%/- 120	241
GOALKEEPER	- 13%/- 120	797
FREEKICK	- 10%/- 22	208
OFFSIDE	- 6%/- 5	83
CORNERKICK	/	84
CROSS	/	67

Tables 4 and 5 rank concepts by the relative reduction in their total token mass after disambiguation, and gives, in addition, the final concept frequency. As can be seen, there is a huge variation in the amount of tokens discarded: for the Dutch data, concepts like ATTACK and FOUL lose nearly all tokens, while others like GOALKEEPER, FREEKICK and PENALTY KICK do not lose any tokens. The reason for such a drop in tokens in ATTACKER is due to the frequent out-of-concept use of the variant spits which is highly polysemous, whereas in FOUL it is mostly attributable to the out-of-concept sense of both variants involved, i.e. *fout* and *overtreding*. In the case of ATTACK and FOUL we can see that such a removal also has an impact on the external uniformity value, as there is a decrease of 8% for ATTACKER and an increase of 24% for FOUL. At the same time, we see that the conspicuous number of tokens discarded for REFEREE does not have an impact on the uniformity value. For the Chinese data, concepts like REFEREE, KICKOFF, HANDS, COUNTERATTACK, and GOAL lose more than 90% of their tokens. As we explained above, the reason for such a loss in tokens is due to the frequent out-of-concept use of at least one of the variants of those concepts. For instance, the dominant use of the COUNTERATTACK variants *fan-gong* and *fan-ji* in the domains of war or argument or the pervasiveness of the GOAL variant *de-fen* in the contexts of basketball or baseball. For

the concepts at the bottom of the tables we can safely assume that their variants do not show polysemy in the corpora, and that therefore these concepts are not in need of being disambiguated.

5 Conclusion

In this contribution we have attempted a replication of the sociolectometric analyses carried out in GGS1999, by focusing on the football concepts in the dataset. This time, however, we replaced the manual disambiguation of the corpus occurrences with a semi-automatic procedure based on token-based vector space models and cluster analysis, and we looked at two pluricentric languages simultaneously, i.e. Dutch and Chinese. The results revealed that removing semantic clusters whose most central members are considered out-of-concept tokens, does have an impact on the sociolectometric distances. Furthermore, discarding entire clusters has consequences for the total concept frequency, which in turn is important when the concepts are weighted before the final aggregation step (which was not carried out here).

This paper has only scratched the surface of the opportunities that vector space models offer for sociolectometric research, and much more is being investigated in the project "Nephological Semantics" QLVL (3H150305). Apart from their practical utility it will be possible to explore whether the cluster structure derived from tokens models can provide information about the influence of semantic subcomponents on lexical variation. In this project, headed by Dirk Geeraerts, it has become clear that the success of such models strongly depends on the judicious choice of parameter values underlying their construction. In any case, these techniques have proven to be indispensable for any kind of corpus investigation that requires semantic control.

References

Boleda, Gemma. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics* 6(1). 213–234.

De Pascale, Stefano. 2019. *Token-based vector space models as semantic control in lexical lectometry*. Leuven: KU Leuven dissertation.

Geeraerts, Dirk. 2018. A lectometric definition of lexical destandardization. In Stefan Engelberg, Henning Lobin, Kathrin Steyer & Sascha Wolfer (eds.), *Wortschätze: dynamik, muster, komplexität* [Lexicons: dynamics, patterns, and complexity], 233–244. Berlin & Boston: Mouton de Gruyter.

Geeraerts, Dirk, Stefan Grondelaers & Dirk Speelman. 1999. *Convergentie en divergentie in de Nederlandse woordenschat: een onderzoek naar kleding- en voetbaltermen* [Convergence and divergence in the Dutch vocabulary: An investigation into clothing and football terms]. Amsterdam: P.J. Meertens-Instituut.

Heylen, Kris, Thomas Wielfaert, Dirk Speelman & Dirk Geeraerts. 2015. Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua* 157. 153–172.

Huang, Chu-Ren. 2009. Tagged Chinese Gigaword Version 2.0 LDC2009T14. Philadelphia: Linguistic Data Consortium.

Lenci, Alessandro. 2018. Distributional models of word meaning. *Annual Review of Linguistics* 4(1). 151–171.

Ordelman, Roeland, Franciska de Jong, Arjan van Hessen & Hendri Hondorp. 2007. TwNC: a multifaceted Dutch news corpus. *ELRA Newsletter* 12(3/4). ELRA. 4–7.

Peirsman, Yves. 2010. *Crossing corpora. Modelling semantic similarity across languages and lects*. Leuven: KU Leuven dissertation.

Rosseel, Laura, Karlien Franco & Melanie Röthlisberger. 2020. Extending the scope of lectometry. *Zeitschrift für Dialektologie und Linguistik* 87(2). 131–143.

Rousseeuw, Peter J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20. 53–65.

Ruette, Tom. 2012. *Aggregating lexical variation: towards large-scale lexical lectometry.* Leuven: KU Leuven dissertation.

Ruette, Tom, Dirk Geeraerts, Yves Peirsman & Dirk Speelman. 2014. Semantic weighting mechanisms in scalable lexical sociolectometry. In Benedikt Szmrecsanyi & Bernhard Wälchli (eds.), *Aggregating dialectology, typology, and register. Analysis linguistic variation in text and speech*, 205–230. Berlin: Mouton de Gruyter.

Soares da Silva, Augusto. 2014. The pluricentricity of Portuguese: A sociolectometrical approach to divergence between European and Brazilian Portuguese. In Augusto Soares da Silva (ed.), *Pluricentricity*, 143–188. Berlin, Boston: De Gruyter Mouton.

Speelman, Dirk, Stefan Grondelaers & Dirk Geeraerts. 2003. Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities* 37(3). 317–337.

Van der Maaten, Laurens & Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*. 9(86). 2579–2605.

Wieling, Martijn & John Nerbonne. 2015. Advances in dialectometry. *Annual Review of Linguistics* 1. 243–264.