UNIVERSITÄT
DES
SAARLANDES

UNIVERSITÀ DEGLI STUDI DI TRENTO
CIMeC - Center for Mind/Brain Sciences

Erasmus Mundus Master's Degree in
Language and Communication Technologies

Master's Degree in Cognitive Science, University of Trento (2018-2019)
Master's Degree in Language Science and Technology, Saarland University (2019-2020)

Similar, but Different: Unsupervised Detection of Semantic Shifts in Diachronic Word Embeddings

MSc Thesis | Academic Year 2019-2020

STUDENT:	Simon Philipp Clemens Preißner
SUPERVISOR:	Prof. Dr. Elke Teich (UdS)
CO-SUPERVISOR:	Dr. Yuri Bizzoni (UdS)
CO-SUPERVISOR:	Dr. Aurélie Herbelot (UniTN)

Declaration – Eidesstattliche Erklärung – Declarazione

I hereby confirm that the thesis presented here is my own work, with all assistance acknowledged. I assure that the electronic version is identical in content to the printed version of the Master's thesis.

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe. Ich versichere, dass die gedruckte und die elektronische Version der Masterarbeit inhaltlich übereinstimmen.

Con la presente confermo che la tesi qui presentata è un mio lavoro, con tutta l'assistenza riconosciuta. Assicuro che la versione elettronica ha un contenuto identico a quello della versione stampata della tesi di laurea magistrale.

Saarbrücken, ... September 2020

Signature:

Abstract

Human language varies across many dimensions due to numerous factors, and a large body of research in natural language processing (NLP) aims to identify and measure these factors, often combining insights from linguistics and the ability to leverage large amounts of language data. One of the most studied dimensions of language variation in this field is *diachronic language change*, i.e. change of language over time. In NLP, the de-facto standard representation of lexical meaning are *word embeddings* (Turney and Pantel, 2010), which express the meaning of a lexical item as a vector in a multidimensional space. Intuitively, the angle between two such word vectors can be interpreted as the semantic similarity of the words, and the vectors' difference as 'shift' of meaning from one word to the other.

In this thesis, I delve into the research of diachronic language change on the basis of word embeddings. I aim to contribute to research in two ways. I first outline the many methods involved in and related to diachronic research of language via embeddings in order to provide a solid basis for other researchers from fields outside of NLP. Then, secondly, I introduce and explore a new unsupervised method for the detection of systematic conceptual shifts. For this, I make two assumptions; namely that (1) diachronic language varieties should be treated as distinct despite their obvious similarities, and (2) no assumptions about specific changes should be made *a priori*.

The starting point for the investigations are word embedding models of various time intervals of the *Royal Society Corpus* (RSC, Fischer et al., 2020), a collection of English scientific texts that spans from 1665 to 1929. Previous work on language change within the RSC (cf. Bizzoni et al., 2020) employs both corpus linguistic methods and measurements on these word embedding models (or: *spaces*). Differently to this, I aim to measure shifts across spaces and operate directly on these shifts.

In order to make spaces comparable to each other and work under the assumption of bilinguality (assumption 1), I align pairs of spaces via *Gromov-Wasserstein Optimal Transport* (GWOT, Alvarez-Melis and Jaakkola, 2018). GWOT is an optimization problem which creates a probabilistic association between points across two spaces. It achieves this without any knowledge of language; instead it matches up points across spaces by a measure of how similarly they behave in their own respective spaces. This probabilistic association can then be used to project one space onto the other. In order to preserve the diachronic shifts, the projection considers only the most stable concepts (i.e. those concepts for which the 'spatial profile' changes the least across time).

To detect *systematic conceptual shifts*, I calculate vectors of conceptual shift, apply a clustering algorithm, and approximate the meaning of the shift vector's direction with nearest-neighbor search. The result is a set of clusters with human-interpretable labels which express the components of meaning which change over time.

Thorough investigations of GWOT as well as the experiments with the proposed method show that the assumption of bilinguality is not necessary. The proposed method is promising; it confirms previous findings and is able to identify new dynamics of diachronic change, e.g., that similar words tend to change in similar ways. I provide impulses to improve the proposed method and to apply it to further language resources.

Acknowledgements

This thesis was a long journey. From an academic aspect, it expresses much of that I have learned in the past years. I enjoyed working on it and I'm thankful to everyone who helped me with it. However, this is just the tip of the iceberg. The past two years have been among the most formative of my life so far. During these two years of pursuing the Erasmus Mundus Master's program in Language and Communication Technologies (EMLCT), I enjoyed the company of young people from around the world. I learned a new language, I got to know new places, new food, new ways to see the world and to make life be good, every single day. I learned to appreciate the new, the known, the far-away places, and home. I'm thankful to all these humans for all these encounters.

I'm particularly thankful to the following people:

My main supervisors, Prof. Dr. Elke Teich and Dr. Yuri Bizzoni, who were both open to my ideas, inspired and motivated many parts of this work, supported me with their knowledge and experience, and guided me through the challenging moments.

My secondary supervisor, Dr. Aurélie Herbelot, who guided my first steps in the world of scientific work and writing; and who is inspiring in every possible way.

Bobbye Pernice and Tessa Libowski as well as Dr. Raffaella Bernardi and Dr. Jürgen Trouvain, who guided me through the bureaucratic intricacies that make EMLCT possible in the first place.

Badr M. Abdullah, who listened to my ideas and brought me to my supervisors.

Zuletzt Clemens, Barbara, Stephan und Maria, die nichts im Speziellen, sondern viel mehr das Allermeiste gemacht haben, um mich – egal, wo ich war – dorthin zu bringen, wo ich bin.

Contents

List of Figures

List of Tables

1 Introduction

Language takes on various forms depending on many factors. The most obvious such factor is typology itself; Dutch, Farsi, Russian, Bulgarian, Chinese, Italian, and German are all different from each other both in their surface structures as well as in their potential and ways to express specific concepts. Moreover, any such typologically distinct language itself can take various forms. In the terms of Ferdinand de Saussure, one can distinguish between *langue* (a construct of lexical items connected by rules) and *parole* (a set of linguistic utterances); between language system and language in use. With this distinction, any one language such as the ones above can also be looked at from the aspect of language in use, and clearly, no two instances of *parole* are the same. Languages themselves are organic; they are means to express the thoughts of their speakers. Therefore, differences between and developments of the users of a language lead to differences between the corresponding instances of language in use. These differences and changes can take many forms; two newspapers, two social groups, two generations of users, two topics etc. are all likely to exhibit some sort of difference in the way that language is used. Linguistic research has long shown great interest in characterizing these differences as well as picking apart the factors that influence them. One particularly well-established area of research investigates the *diachronic* perspective: how and why does a language change over time?

An example of research on diachronic language change is the Royal Society Corpus (RSC) (cf. Fischer et al., 2020), which comprises English scientific texts from 1665 to 1929.[1] During those 265 years, scientific English has undergone numerous changes based on advances in the sciences, the humanities, scientific culture, and the English language in general. This is the language resource of interest to this thesis.

There are various methods that can be used to study a collection of *parole* like the RSC; this thesis employs word *embeddings* (Turney and Pantel, 2010; Ling et al., 2015), which are the de-facto standard form of semantic representation in Natural Language Processing (NLP). In principle, the embedding of a linguistic unit (a character, word, lexeme, phrase) is a vector in a multidimensional space which expresses the similarity of any two units by means of spatial distance. Following the distributional hypothesis (cf. Firth, 1961), embeddings are calculated from large text corpora by modeling the contexts in which the linguistic units appear. If two such units appear in very similar contexts, their embeddings are highly similar and will follow similar directions in the vector space. Embeddings, most often word embeddings, form the basis of many applications in NLP, because they express the semantics of the corpus (i.e., the *parole*) on which they are built in a relatively reliable and algorithmically convenient way.

Two of the fields that advanced substantially with the use of embeddings are machine translation and diachronic computational linguistics. In both fields, a central challenge to leveraging the power of embeddings is the *alignment* of two or more embedding spaces; the challenge of making them comparable to each other. While monolingual scenarios of embedding alignment can rely on information from a largely mutual lexicon, the

[1] these numbers refer to version 5.1; the current version 6.0 of the RSC, cited here, comprises more recent texts as well.

bi- and multilingual scenarios most often encountered in machine translation only have few lexicon entries with a shared spelling. Many approaches to bi- and multilingual embedding alignment rely on dictionaries of one-to-one translations which are compiled manually or generated with heuristics.

Nonetheless, there are several successful approaches to multilingual embedding alignment. One of these approaches is described in Alvarez-Melis and Jaakkola (2018). It relies on the observation that the two embeddings of one concept (or: of two very similar concepts) in two distinct embedding spaces tend to have similar constellations with other concepts in their respective spaces, even for rather unrelated language pairs. The method, which I build upon in this thesis, solves an *optimal transport* (OT) problem by minimizing the *Gromov-Wasserstein distance* (GW) between two embedding spaces. A special feature of Gromov-Wasserstein Optimal Transport (henceforth: GWOT) is that it works fully unsupervised; that is, there is no need for manually compiled or generated dictionaries.

Why use an unsupervised bilingual method for alignment if the embedding spaces for these diachronic investigations are all constructed on English text? If investigated with a narrow sense of semantic identity, one might claim that the lexicons of two varieties of language use (i.e., *paroles*) do not match in the monolingual case either. In this sense, it is possible that while a lexicon entry has the same spelling across two language uses, the underlying concept may differ. Indeed, it is not hard to imagine scenarios in which the precise meaning of a word or an utterance depends on its source: the speaker, the situation, the topic — or the time in which it was uttered. Besides the generation of new words, language change happens within words; they are re-purposed, ascribed new meanings, and stripped from old ones without changing the surface form (cf. Traugott, 2017). For this reason, I make the strong assumption that studies of diachronic conceptual change should not fully rely on the largely shared vocabulary across time intervals either. Instead, any two *paroles* of one language from distinct periods in time might as well be regarded as two distinct languages: similar, but different. GWOT as an unsupervised method for alignment satisfies this assumption of bilinguality.

Once two embedding spaces are aligned to each other, it is possible to compare embedding vectors across time. Previous research of word use in the RSC has successfully discovered and confirmed several developments of scientific English over time; additionally, the decade-spanning word embedding spaces have been used for investigations of general dynamics of language change (cf. Bizzoni et al., 2020). For example, Bizzoni et al. (2019a) report that the scientific language within the corpus becomes more specialized over time: clusters of words that are used in similar contexts become tighter, while the clusters grow apart from each other (i.e. become less similar to each other).

In this thesis, I use the embedding spaces to detect conceptual changes on a smaller scale, specific to small groups of concepts. Besides the quantitative aspects of language change (how much language or specific groups of concepts change over time), a particular interest here lies in the qualitative aspects, that is, in which *direction* the change happens.[2] To this end, I propose a novel combination of widely applied techniques to

[2] Here, 'direction' is to be taken in the spatial sense: when the usage of a word changes over time, its vectorial representation also changes; the word has varying positions in space depending on the point

find systematic conceptual shifts and make their nature of change more interpretable.

In addition to the first assumption that one should not rely on similar (i.e. overlapping) vocabularies as an anchor to align word embeddings, I assume that it is desirable to exclude human intuition from the detection of shifts. The main reason for this assumption is human bias and the tendency to miss subtleties: it is simple to come up with obvious candidates for diachronic semantic change of English such as *gay*, *broadcast*, or *cell* (cf. Hamilton et al., 2016b) and investigate their behavior over time. However, it is much less intuitive to think of developments such as the tendency to speak of an *assumption* rather than a *supposition*, and the question of whether this is because *supposition* changed its meaning. Subtle changes like this one are hard to detect with human intuition alone. Therefore, it is desirable to approach the task of detection of diachronic conceptual change as an unsupervised task. The proposed method for shift detection aims to satisfy this by considering every word in the vocabulary and finding groups of concepts which change in a similar way and filtering those groups which show an unusual pattern of change.

What to Expect. With these two assumptions, the application of GWOT to embedding spaces of the RSC, and the proposed method for the detection of shifts, I aim to find answers to the following questions:

1. Is it useful to frame the detection of diachronic conceptual change as bilingual task?
2. How well does the method for unsupervised alignment work with respect to the RSC; is it reliable and feasible?
3. Does the proposed method for the unsupervised detection of shifts work; is it sensible and expressive?
4. Do the insights from the proposed method relate to previous findings?
5. Can the proposed method uncover new (general and/or specific) dynamics of language change within the RSC?

This thesis, seen in the context of current research on the development of scientific English in the past 350 years, aims to establish a basis of knowledge for the investigation of conceptual change via word embeddings. It is addressed to a mixed readership of the fields of linguistics, NLP and history; therefore, it provides overviews and background information on many established methods of NLP for further reference. Besides this channeling of information, the thesis contributes to current research in three ways. First, by proposing and applying a novel method of shift detection. Second, by applying a specific technique for alignment and the proposed method for shift detection to word embeddings of the RSC. Third, by evaluating both, alignment and shift detection, with respect to the RSC. Fourth, by increasing the state of knowledge about the data at hand

in time at which it is observed. The difference between these points is itself a vector that expresses the direction of shift.

through new insights and validation of previous findings. With these contributions, this thesis should give insights about the possibilities and limitations of embedding-focused research on conceptual shifts within diachronic corpora such as the RSC and possibly inspire research on other language resources. The relevant code is available online.[3]

Structure. Chapter 2 gives a frame of reference for related research. Specifically, it gives an intuition and references to techniques for the construction of embedding spaces (§2.1), the alignment of such spaces (§2.2) and the detection of conceptual shifts within them (§2.3). Chapter 3 reports on the text data (§3.1) used to construct embedding spaces, which are described and evaluated in (§3.2). Chapter 4 goes into detail about unsupervised alignment via Gromov-Wasserstein Optimal Transport (GWOT) with theoretical foundations (§4.1) and introduces an adaptation of GWOT for diachronic scenarios (§4.2), followed by applications to the data at hand (§4.3). The insights from Chapter 4 are then used in Chapter 5 to select and align a sub-set of the available embedding spaces of the RSC (§5.1) and apply the method for shift detection presented in Section 5.2. The results, both quantitative and qualitative, are presented and discussed subsequently (§5.3, §5.4). Chapter 6 critically reflects on the methodology, findings and insights from the preceding two chapters, followed by considerations for future investigations. Chapter 7 concludes the thesis.

2 Related Work

This chapter aims at giving a background for the work carried out in this thesis. For this, I will first give an overview of the construction of and work with distributional semantic models in Section 2.1. This is followed by a presentation of previous research on the two main methodological aspects (i.e., word embedding alignment and detection of embedding shifts) in Section 2.2. The chapter concludes with previously established observations about language change over time, both in language in general and an within the language data at hand (Section 2.3).

2.1 Background of Distributional Semantic Models

In order to understand the behavior of alignments for word embeddings, and later on, of shift detection, it is important to have an intuition about the methods with which the word embeddings were constructed in the first place. Therefore I will first briefly describe the broad variety of modeling techniques of Distributional Semantics (DS), followed by intuitions about DS spaces.[4]

[3] https://github.com/SimonPreissner/get-shifty

[4] As for the terminology, I use the term *embedding* more broadly than usual. While the term was coined specifically in the context of predictive modeling, I use it for any vectorial representation of meaning in a distributional semantic model. Similarly informally, I refer to DS models as *spaces*.

2.1.1 Modeling Techniques

Embeddings are the de-facto standard for the representation and processing of semantics in natural language processing (NLP). The numerous approaches mainly vary in two ways: the size of the represented units of meaning and the technique with which the approaches model those units. For a long time, the most commonly used unit of meaning was the word (e.g. Turney and Pantel, 2010; Baroni and Lenci, 2010; Mikolov et al., 2013b; Pennington et al., 2014), but especially recent advances in computational semantics treat larger units such as phrases or sentences (e.g. Peters et al., 2018; Devlin et al., 2019; Yang et al., 2019; Sanh et al., 2020; Brown et al., 2020).[5] In general, the amount of data (and processing power) required for reliable modeling grows with the size of the linguistic units to be modeled. Most diachronic text corpora, especially for less recent times, are not large enough to make approaches with larger units of meaning feasible. Consequently, diachronic studies involving DS representations normally use *word embeddings*; work on larger units (e.g., Martinc et al., 2020) is the exception.

With respect to the modeling technique, the two main paradigms are *count-based* and *predictive modeling*. Going back to the origins of DS, the distributional hypothesis (cf. Firth, 1961) states that the meaning of a concept can be estimated by the set of contexts in which it is used, that is, by its *distribution* in language. In the simplest count-based approach, this means that given a corpus C with v different words (i.e., a vocabulary V of size v), the co-occurrences of words within a certain proximity (e.g., ±5 words apart) can be stored in a matrix $\mathbf{M}^{v \times v}$, and each row of $\mathbf{M}$ is a v-dimensional vector x_w that represents the distribution of a word $w \in V$ in the corpus. These vectors, all of the same size, can be thought of as points in a v-dimensional space. Words with relatively similar contexts correspond to points which are relatively close together in this space. Thus, under the assumption that similar use expresses semantic closeness, the spatial proximity of two word vectors expresses the semantic similarity of their words. Note that the one-to-one correspondence between contexts and vector dimensions in this naive approach is infeasible in reality, because (a) the models do not scale well with vocabulary size and (b) the models value the sheer frequency of a context higher than its actual informativeness. A large body of research has addressed both issues and developed robust and convenient count-based models (e.g. Landauer and Dumais, 1997; Deerwester et al., 1990; Rohde et al., 2006; Levy et al., 2015; cf. Turney and Pantel, 2010 and Bullinaria and Levy, 2007 for overviews).

Predictive methods approach the construction of DS models from a different angle. They use corpora to produce training data for machine learning models which are trained to solve completion tasks. Intuitively, the goal of training is to find features in the observed words which are useful to solve that task. For example, Bengio et al. (2003) train a neural network to predict the next word in a given sequence of words. Within the neural network, the input words are converted to vectors with d features, which are stored in a matrix $\mathbf{M}^{v \times d}$. The values of $\mathbf{M}$ (i.e., the values of the embeddings) are

[5] To this day, the notion of 'word' is debated, and so is its relation to units of meaning (i.e, concepts). The relationships between distinct words (in the sense of letter strings) and concepts are often 1:n (polysemy) or n:1 (synonymy); some concepts can only be expressed with multiple compounds; and so on. In this thesis, the main linguistic unit is the word, in the meaning of letter string.

adjusted by the training algorithm. After training, each row in $\mathbf{M}$ corresponds to the d-dimensional embedding of a specific vocabulary item. There are various tasks used for training word embedding models, for example predicting a word from its surroundings, or inversely, predicting the surroundings of a word (cf. Mikolov et al. (2013b); Ling et al. (2015)).

While distributional semantics has its roots in count-based models (cf. Bullinaria and Levy 2007), there has been a great shift towards predictive modeling with the rise of machine learning. Both approaches do have commonalities; Levy and Goldberg (2014) for example show that skipgram with negative sampling (SGNS), a predictive approach, implicitly uses PMI, which is a common technique in count-based DS models. Still, predictive models tend to perform better on a wide range of tasks (Baroni et al., 2014). The distributional semantics models employed in this thesis are trained as *structured Skipgrams*, a technique proposed by Ling et al., 2015, who extended the SGNS technique from Mikolov et al. (2013b) in order to capture syntactic regularities additionally to purely semantic information. The training of these models will be discussed in detail in Chapter 3.2.

2.1.2 Properties of Vector Space Models

Word embeddings are commonly treated as vectors in a multidimensional (and usually euclidean) space. Two vectors pointing into similar directions correspond to semantically related concepts. The vectors usually are normed to unit length so that the similarity between two vectors is expressed by the angle between them. This measure is called the *cosine similarity* and is arguably among the most widely used (Turney and Pantel, 2010). Values of cosine similarity typically range from -1 (opposite directions) to 1 (identical direction), with a value of 0 expressing no semantic relation at all. Cosine similarity (and its counter-part, *cosine distance*) is frequently used to identify concepts which are most closely related to a given concept, for example to create topical clusters or, if the word embeddings are multilingual, to find translations of words.

Another convenient characteristic of most vector space models is compositionality: basic mathematical operations apply such that adding and subtracting word vectors from each other yields a new vector which is very similar to a concept combining the semantics of the original vectors. The prime example, taken from Mikolov et al. (2013c), is that in a reliable vector space model, the vector resulting from $king - man + woman$ has as nearest neighbor the vector of *queen*. Intuitively, by subtracting the highly male-connoted *man* from *king*, the resulting vector should resemble the concept of *royalty*. Compositionality is often exploited to operate on components of meaning, e.g. by Bolukbasi et al. (2016) in order to mitigate bias in word embeddings.

Beyond compositionality and simple analogies such as the above, Mikolov et al. (2013a) observe that the overall geometric arrangement of concepts in embedding spaces is similar across languages; they are *isomorphic*. Thus, two such vector spaces which are isomorphic to each other can be *aligned* to one another by rotating and scaling one of them so that similar concepts from both spaces lie close together. Chapter 2.2 (and in fact, the entirety of Chapter 4) presents some approaches which exploit the observation of isomorphy.

Despite the many convenient characteristics of vector space models and the fact that DS models nowadays serve as a reliable basis for most semantics-related applications in NLP, there are certain limitations to the modeling capabilities of word embeddings. First, they are known for assigning similar points in space to antonyms (i.e., words with opposite meaning). This happens because antonyms tend to be used in very similar contexts, with the only difference being that one is the negation of the other. Consequently, there has been research specifically on how to distinguish antonymy in word embeddings (e.g. Ono et al., 2015; Nguyen et al., 2016). Second, standard word embedding models which define words as sequences of letters do not account well for polysemy, that is, they model the multiple senses of a word (in the sense of a sequence of letters) as one single embedding. Additional information can help mitigate this problem (cf. Liu et al., 2015; Mancini et al., 2017; Kekeç et al., 2018).[6] Third, word embedding spaces tend to contain embeddings which are popular nearest neighbors of many other words. These *hubs*, discovered by Radovanović et al. (2010), can become problematic in the clustering, comparison and alignment of word embeddings. Fourth, word embedding models, count based as well as predictive, are not necessarily stable. That is, two models which are trained in the same way on very similar data can be relatively different from each other. Wendlandt et al. (2018) find that even relatively frequent words can be unstable, which is counter-intuitive given that they are more likely to be observed often enough to be modeled well. Finally, not all embedding spaces are approximately isomorphic to each other. Especially across languages and even for relatively similar languages (Søgaard et al., 2019), it can not be guaranteed that the geometric constellations of similar concepts are similar across spaces.

These and other issues as well as the intense research on how to mitigate them show that results from experiments directly involving vector semantics are to be taken with a grain of salt. The limitations of word embeddings do not allow to take singular observations out of context. Rather, multiple consistent observations in such experiments can inform about more general tendencies.

2.2 Alignment of Word Embeddings

There are several scenarios in which it is of interest to relate two or more existing embedding spaces to each other; the two most prominent are multilingual and diachronic scenarios. In multilingual scenarios, two embedding spaces from different languages (i.e, with different vocabularies) need to be related to each other in order to find the closest translation of a given word. A secondary interest in this scenario is to compare the behavior of concepts to infer commonalities and differences in conceptualisation across languages. In diachronic scenarios, the embedding spaces' underlying language is roughly the same (i.e., with largely overlapping vocabularies), but stems from different

[6] Note that deep neural models such as in Peters et al. (2018) or Devlin et al. (2019) do solve this issue to a great extent. As they are trained on larger units of meaning, their embedding representations for a single word differ depending on the context in which the word is used.

periods or points in time. Aligning spaces in this scenario allows to compare the behavior and development of single concepts as well as the underlying language in general over time.

There is an extensive body of research about alignment methods for both multilingual and monolingual (e.g., diachronic) embeddings. Approaches range from supervised alignment on shared vocabulary or with translation dictionaries (e.g., Mikolov et al., 2013a; Artetxe et al., 2017; Faruqui and Dyer, 2014; Lazaridou et al., 2015) to fully unsupervised methods (e.g., Conneau et al., 2018; Alvarez-Melis and Jaakkola, 2018; Jawanpuria et al., 2020) and from simple matrix decomposition (cf. Artetxe et al., 2018) to the training of generative adversarial networks (GANs, Goodfellow et al., 2014; e.g., Zhang et al., 2017; Conneau et al., 2018). Not all of these methods are suitable for the task at hand, but in order to give a frame of reference for the task of word embedding alignment in general, I will outline a variety of previously devised methods: mapping-based, second-order and adversarial alignments.

2.2.1 Supervised (and Mapping-Based) Methods

These approaches aim to create a linear transformation in form of a projection matrix $\mathbf{P}$ which projects any point in the source space X either directly onto the target space Y or onto a shared projected space Z. They work supervised, that is, they require a set of words which are known to be 'translations' of each other. The construction of this bilingual signal is crucial to the quality of the resulting alignment. Some of the relevant aspects for the construction of such dictionaries are investigated in Vulić and Korhonen, 2016.

Orthogonal Projection with (Minimal) Supervision. Given a bilingual signal, the simplest solution to finding $\mathbf{P}$ is to solve an orthogonal Procrustes problem, which will be introduced in Chapter 4.1.2. Intuitively, such a $\mathbf{P}$ rotates and scales all points in X at the same rate, thus it works especially well if X and Y are (approximately) isometric.

The challenge for orthogonal methods is to find a set of 'anchor terms' which are very likely corresponding to each other, and to have the bilingual signal be representative of the two word embedding spaces. Of course, it is appealing to achieve a good mapping with as few anchor terms as possible, because the construction of dictionaries is costly and external resources may introduce translation errors. For this, Artetxe et al. (2017) propose an iterative framework for the creation of large bilingual signals: with a small set of at least 25 anchor terms, they create a mapping between source space and target space, perform bilingual lexicon induction, and use the obtained translation pairs as additional anchor terms in the subsequent iteration. In an additional experiment, they use shared numerals as anchor terms, starting with a larger, but trivially constructed set of translation pairs. An equally simply constructed bilingual signal is used by Smith et al. (2017), who base their alignments on shared strings. Differently to this, Zhang et al. (2016) take the 10 most frequent words in the corpus underlying a target space and construct a bilingual signal with these 10 words' Wiktionary translations. The coarse mapping obtained from this weak signal is then used in a part-of-speech (POS) tagger which itself is trained to the bilingual task. This combination allows to restrict the bilingual signal for the mapping to a bare minimum, but Zhang et al. (2016) argue that a bilingual signal of this small size makes the isometric constraint of the projection a necessity.

Regression Methods. As with the orthogonal Procrustes problem (cf. Chapter 4.1.2), regression methods aim to find a projection of one space onto the other that minimizes the distances between certain pairs of points. Here as well, the construction of a reliable bilingual signal is important.

However, instead of using matrix factorization such as in the solution to the Procrustes problem formulated in Schönemann (1966), the projection is optimized via gradient descent, for example with a least squares objective as presented in Mikolov et al. (2013a). In this context the bilingual signal functions as training data and consists of translations from online Google Translate of the 5k most frequent words from the source language. The least squares method has a similar objective as the orthogonal Procrustes problem, with the difference that it can be modified, for example with L2-regularization (cf. Lazaridou et al., 2014). Subsequent research on partial least squares regression as a mapping method has found that the hubness problem (ie., the negative effect some points being the nearest neighbor of many other points) is somewhat unidirectional and that hubness effects in one direction can be mitigated by learning the projection in the reverse direction (e.g., learning $\mathbf{P}\ Y$ to X in order to translate from X to Y; cf. Shigeto et al., 2015).

Margin Methods. In the effort to learn mappings from distributional semantic spaces to the visual domain, Lazaridou et al. (2015) include cross-linguistic experiments. They attribute the increase of hubness effects with the least-squares method (presented above via Mikolov et al., 2013a) specifically to that method and present an alternative objective called *max-margin.*

Just as in other settings, the starting point are two spaces X and Y and a small bilingual signal. The max-margin objective requires a measure of distance $dist()$ that can be applied to a pair of vectors of the same dimensionality. Similar to the regression method, the aim is to find the optimal mapping W from X to Y. This is achieved by minimizing $\Sigma_{i \neq j}^{k} max\{0, \gamma + dist(Wx_i, y_i) - dist(Wx_i, y_j)\}$ for all x and y in the bilingual signal. This means that W should project every x_i onto Y in a way that the projected vector is closer to the true translation y_i than to negative translations y_j. The hyperparameter γ allows the mapping to lead to wrong translations, which is desirable in some cases in order to prevent overfitting. The translation pairs serving as training data are obtained from bilingual, possibly word-aligned data; Lazaridou et al. (2015) use the 5k most frequent word pairs obtained from a Europarl-derived dictionary. In their experiments, the max-margin objective outperforms a regression model similar to the one outlined above.

Canonical Methods. This approach to word embedding alignment, such as Canonical Correlation Analysis (CCA), described in Faruqui and Dyer (2014), establishes two mappings, one for each space X and Y, to project into a common space Z. Here, the vectors of the words in the bilingual signal form sub-spaces X' and Y'. CCA finds two projection vectors v and w that maximize the correlation between projected vectors of translation pairs, $\rho(x_i v, z_i w)$, across X' and Y' respectively. The vectors v and w can then be used to map the remaining words (of X and Y, respectively) into two new spaces of equal dimensionality. To obtain a bilingual signal, Faruqui and Dyer (2014) select words in one language and their most frequently aligned word in the other language as translation pairs using word-aligned corpora. This results in 36k translation pairs (as reported in Lu et al., 2015).

Further work on canonical methods in this field has seen neural approaches as well. Lu et al. (2015), for example, employ neural networks, one for each embedding space, to extract feature vectors which are then used as the starting point for CCA. Using the same embedding spaces and translation dictionary as Faruqui and Dyer (2014), they report better performance on multiple word similarity tasks than conventional CCA.

Artetxe et al. (2018) subsume the four mapping techniques outlined above under a generalized framework. It defines each technique as a linear transformation with partly optional pre- and post-processing steps: normalization, whitening, re-weighting, de-whitening, and dimensionality reduction. All of the presented techniques map two monolingual embedding spaces onto one shared space, using bilingual dictionaries as a basis.

Non-Orthogonal Projections Many mapping-based techniques (e.g. Mikolov et al., 2013a; Miceli Barone, 2016) are based on the assumption that word embeddings are approximately isomorphic, i.e. can be aligned with linear and even orthogonal transformations. However, this assumption has been criticised as not being true for every language pair, especially for farther-apart languages (Søgaard et al., 2018; Vulić et al., 2019).

For this reason, some research has gone into locally sensitive mapping techniques. Nakashole (2018) identify ‘neighborhoods’ of vectors which are close to each other. Each neighborhood has an associated weight vector which is used to ‘enrich’ a word vector with information from its neighborhood before being projected onto the target space. The projection as well as the neighborhood weights are learned jointly. While not leading to state-of-the-art results on closely related languages, this approach proves to be effective for aligning word embedding spaces from more distant languages.

Another locally sensitive approach is presented by Glavaš and Vulić (2020). Here, starting with a bilingual signal, a source space X is first rotated to roughly match the target space Y and then every vector x_i in the bilingual signal is mapped to its corresponding y_i. These mappings are also applied to the closest neighbors of each x_i *individually* in order to induce new translation pairs. These two steps, rotation and instance-based mapping, are repeated in order to extend the bilingual signal.

2.2.2 Unsupervised and Second-Order Methods

In the unsupervised setting, no bilingual signal is given. Thus, an alignment must be built solely on the linguistic commonalities between two word embedding spaces. In other words, instead of comparing vectors directly to one another across spaces, these methods often measure and relate their *behavior*.

Alvarez-Melis and Jaakkola (2018) make use of an abstraction from the points in a vector space to distances between them. The key idea is that although two embedding spaces do not share dimensions, the same metrics (e.g., cosine distance) can be employed within them. This allows to construct metric spaces via within-space measurements which are directly comparable. Solving an optimal transport (OT) problem analytically over these metric spaces minimizes the Gromov-Wasserstein (GW) distance and yields a 'transportation plan' Γ between the points in X and Y, which can be interpreted as a probabilistic translation table – or a bilingual signal (cf. Chapter 4.1.3). This is then used to solve an orthogonal Procrustes problem to obtain a projection $\mathbf{P}$ from X onto Y. This OT approach is generalized in Alvarez-Melis et al. (2019).

Differently to creating a bilingual signal analytically, Grave et al. (2019) learn an orthogonal projection matrix with stochastic gradient descent. Their objective is to minimize the Wasserstein distance between the projected space $X\mathbf{P}$ and the target space Y. The objective function employs $\mathbf{P}$ and the 'transportation plan' Γ between X and Y. Γ is sub-sampled in each iteration, which makes optimization faster than the approach of Alvarez-Melis and Jaakkola (2018) and achieves state-of-the-arts results in bilingual lexicon induction.

Very recently, Jawanpuria et al. (2020) present an approach which also operates on distance matrices such as those in Alvarez-Melis and Jaakkola (2018). It also learns a permutation matrix rather than a projection, but unsupervised alignment is framed as a bidirectional domain adaptation problem between the distance matrices. The permutation matrix is constructed with a manifold optimization algorithm and subsequently used to infer a bilingual signal.

These methods are appealing to the task at hand for multiple reasons.[7] First, by framing the alignment of diachronic word embeddings as a bilingual task, there is very little to no bilingual signal available, as there is little need for diachronic dictionaries. Second, an unsupervised approach seems 'cleaner' than a supervised approach involving heuristically constructed dictionaries such as those based on shared strings. Third, the inherent similarity of the spaces at hand might very well allow for orthogonal projections, despite the fact that embedding spaces are not generally isomorphic (cf. Søgaard et al., 2018). The diachronic embeddings are trained on data of (more or less) the same language and the set of domains is largely consistent (cf. Chapter 3). This justifies the assumption of (close) isomorphism, which is required for successful orthogonal projections.

[7] The choice for Alvarez-Melis and Jaakkola (2018) is was made before Jawanpuria et al. (2020) had been published.

2.2.3 Adversarial Methods

This unsupervised approach employs the encoder-decoder architecture of generative adversarial networks (GANs, Goodfellow et al., 2014). Training a projection $\mathbf{P}$ from a source space X to the target space Y involves an adversarial game: a generator is trained to sample embeddings from X and transforms them to be as close to embeddings in Y as possible. At the same time, a discriminator is trained to detect which embeddings in a given set are originally from Y and which ones are created by the generator. After training, the generator is used as $\mathbf{P}$.

Zhang et al. (2017) perform experiments on three different GAN architectures, outperforming the regression method of Mikolov et al. (2013a). In a combined approach, Conneau et al. (2018) use GANs to learn a rotation which coarsely aligns spaces. Pairs of highly frequent words, one from each of the two spaces, which come close to each other by this rotation are then used as bilingual signal to solve a Procrustes problem. Finally, the authors introduce a new density-sensitive similarity metric, CSLS (cross-domain similarity local scaling) which mitigates hubness effects and shows to be better than cosine similarity with respect to the induction of translations between the less frequent words of the spaces.

To conclude the short overview of approaches to the alignment of word embeddings, Søgaard et al. (2019) contains a comprehensive review on methods for the construction of multilingual word embeddings, including their construction from parallel and comparable data, such as in Yao et al. (2018) or Upadhyay et al. (2016). On the basis of the large variety of methods reviewed, some key insights are: (1) with supervised methods, the choice of bilingual signal is often more important than the underlying method itself; (2) cross-domain similarity local scaling (CSLS) as a measure for spatial proximity is more robust against hubness than cosine similarity; and (3) unsupervised methods often struggle to align embedding spaces from corpora of varying sizes or domains.

2.3 Shift Detection with Word Embeddings: Techniques and Findings

The fact that distributional semantic models can capture subtleties of word use and semantics, combined with the possibility to make such models comparable to one another, has brought about research on semantic change from various areas such as linguistics, computational linguistics, history, political sciences, and social sciences. Two surveys of semantic change should not be missed when delving into this topic: Kutuzov et al. (2018) provide a selection of influential past and current research on language change with word embeddings and outline the open challenges in the field. The survey of Tahmasebi et al. (2018) is more general, additionally describing non-embedding-based approaches. It also includes considerations from the perspective of linguistics. The following presents a selection of past research, describing a variety of methods as well as findings. The aim here is not to give an exhaustive survey; the presented works are selected in order to provide an intuition about the wide range of possibilities of and the limitations to the detection of semantic change in word embeddings. The 'laws' of semantic change formulated in the presented literature are listed in Table 1.

Incrementally Trained Embeddings. A popular technique to compare word embeddings to each other, especially in the diachronic scenario which allows to cleanly partition training data, is to train the embeddings *incrementally*: the embeddings for each time interval (years, decades, centuries) are trained by initializing the model with the embeddings of the previous time interval. This method is used by Kim et al. (2014), who train year-wise word embeddings on the Google Books Ngram corpus (Michel et al., 2011) starting from 1850. They allow for a 'starting phase' of 49 years, comparing word embeddings across years beginning from 1900. Comparisons are carried out by means of cosine similarity between the various vectors of a word across time; lower similarity values are associated with greater shifts. To better characterize the shifts, they also track the development of cosine distance from a word to its 'old' and 'new' nearest neighbors.

The same training paradigm and corpus is used by Dubossarsky et al. (2015), but only every full decade is saved. The study then performs K-Means clustering on the embeddings of the 7000 most frequent words and computes the cosine distances between the words' decade-apart pairs of vectors. The distance of a word (i.e., the amount of change over time) is then correlated to its proximity to the cluster's centroid. They find that words which are closer to the centroid are less likely to change (later called the 'law of prototypicality').

In a non-diachronic study of semantic differences, Cafagna et al. (2019) train word embeddings incrementally on topic-aligned sub-corpora of two Italian newspapers. Here, the conceptual difference of a word between the two spaces is not quantified by cosine distance. Instead, the authors subtract one of the vectors from the other and measure the length of the resulting shift vector. Differences between two different words are also measured, adjusting the length of the resulting vector according to the difference between two words' corpus frequencies.

Training diachronic word embeddings incrementally is a viable approach to circumvent the need for alignment. However, as Kim et al. (2014) point out, the diachronic change of a word can only be characterized if that word appears reasonably often at both points in time. If for example the word's usage suddenly drops, the corresponding vector will stay similar and not be re-trained very much.

Aligned and non-aligned Embeddings. In order to demonstrate the capabilities of varying methods, Kulkarni et al. (2015) employ frequency methods on raw and on POS-tagged tokens and perform experiments with skipgram embeddings which are trained individually for each time period and aligned with least-squares regression. Similarly to Kim et al. (2014), they use series of a word's cosine distances across time to measure how much it shifts. They additionally devise a method for 'change point detection' to identify which of the shifts are statistically significant and show its effectiveness on three large corpora (Google Books Ngram, Amazon Movie reviews, and Twitter data) with varying temporal granularity (every 5 years, yearly, and monthly, respectively).

Hamilton et al. (2016b) investigate the influence of word frequency and degree of polysemy on rates of semantic change. For this, they train word embeddings with count-based as well as predictive methods, decade-wise on various sub-corpora of the Google Books Ngram corpus, and align them with Procrustes. As an alternative to the change point detection of Kulkarni et al. (2015), the time-series of cosine distances are compared by means of Spearman correlation in order to detect words with statistically significant rates of change. The degree of polysemy of a word is approximated by the diversity of its contexts; intuitively, a polysemous word is a member of many otherwise relatively disjoint clusters. Based on their findings, Hamilton et al. (2016b) proclaim two statistical laws of semantic change, namely the 'law of innovation' (degree of polysemy and rate of semantic change are positively correlated)[8] and the 'law of conformity' (corpus frequency and rate of semantic change are negatively correlated).

In a follow-up study on the same embedding spaces, Hamilton et al. (2016a) use two different measures of semantic shift; the 'global' conventional cosine similarity and a 'local' shift measure. The latter first finds the word's k nearest neighbors at both points in time and then constructs two similarity vectors from this joint neighborhood. The similarity vectors consist of the cosine similarities between the neighborhood and the word in question. Finally, the cosine distance between these similarity vectors is taken to be the extent to which the word's neighborhood has changed. By contrasting a word's global and local change, it is possible to distinguish linguistic drift from cultural shift. The experiments also show that nouns tend to shift locally while other classes of content words (verbs, adjectives, adverbs) tend to shift globally.

Another law of semantic change is formulated by Eger and Mehler (2016): the 'law of linear decay' states that the similarity of a word to itself at different points in time decreases linearly with the temporal distance. The embedding models are not aligned as in other studies; instead, words are expressed as vectors of cosine similarities to other words at the same time (similarly to the 'local' measure in Hamilton et al., 2016a). These second-order vectors are then used to construct a graph which represents a word's change in 'spatial role' over time. Indeed, several studies demonstrate that embedding spaces need not be aligned in order to uncover diachronic tendencies. Gonen et al. (2020) propose to quantify conceptual shift of a word by the amount of overlap of large neighborhoods (e.g., 1000 nearest neighbors) at different points in time. Earlier, Xu and Kemp (2015) have used this approach, comparing the changes in neighborhoods of 100 nearest neighbors, and found that synonyms as well as antonyms change at similar rates. They conclude that their experiments are in line with the 'law of parallel change' (Stern, 1921). In the case of near synonyms, this change may also be semantic decay (cf. Kutuzov et al., 2018).

Finally, Schlechtweg et al. (2019) compare a variety of embedding methods and measures of similarity and divergence and conclude that the best performing approach is to first train embedding models with SGNS, align them with orthogonal projections and then use cosine similarity as a measure for shift detection.

[8] From the view of polysemous words being less likely close to the center of a word cluster and more likely at overlapping outskirts of multiple clusters, this is somewhat similar to the 'law of prototypicality' from Dubossarsky et al. (2015).

Developments of Scientific English. Apart from the more recent studies on the RSC which focus on writing styles and purely linguistic developments, the RSC (or parts of it) has previously been subject of research on scientific language, with outreach to the humanities. For example, taking a rather sociolinguistic perspective, Atkinson (1996) finds that scientific writing in the 17th and 18th centuries was greatly influenced by norms of social conduct. Moessner (2009) confirms the hypothesis that the formation of a new, scientific style of writing was influenced by language policies of the *Royal Society*. Clearly, historical, sociological, and normative linguistic developments such as the introduction of a reviewing process in 1751 (Degaetano-Ortlieb and Teich, 2018) all have an influence on the development of the language of the RSC, both in terms of linguistic style (i.e., how science is communicated) and conceptualization of words (i.e., what certain words mean).

To uncover these developments, Degaetano-Ortlieb and Teich (2016) estimate probability distributions over part-of-speech (POS) trigrams for spans of 50 years and measure the Kullback-Leibler divergences (KLD) between consecutive distributions. This allows to identify POS-trigrams which become more typical over time. Additional analyses with the average surprisal of these trigrams inform about the versatility of their contexts. They find that gerund and passive constructions both become more conventionalized over time: they become a more typical part of the syntactic structure of sentences, but at the same time they are not used in a greater variety of contexts.

KLD is also used in Degaetano-Ortlieb and Teich (2018) to characterize the developments of lexis and grammar and to compare them with one another. The KLD models are created in multiple granularities (min. 2-year periods), either on lemmas (for lexis) or POS-trigrams (for grammar). The comparison of models from consecutive periods shows that in the RSC, grammar becomes increasingly consolidated and lexis, in contrast, experiences a back-and-forth between periods of lexical expansion and lexical consolidation. Additionally, the diachronic modeling of lemma usage with KLD allows to identify the words that contribute to lexical changes (e.g., by being newly introduced).

Differently to these information- and frequency-based methods, Bizzoni et al. (2019a) use word embeddings. Specifically, they use structured skipgrams (cf. Chapter 3.2) and measure cosine distances between varying types of part-of-speech. The results suggest that these embedding spaces expand over time, reflecting the law of linear decay. This tendency to drift apart is more pronounced for terms with a specialized meaning; in contrast, highly functional words are less affected. Words in between these two classes, i.e. content-bearing words with grammatical function such as verbs in gerund or past tense forms, also undergo the process of specialization that pure content words experience.

These more recent findings are combined in Bizzoni et al. (2020) to present a comprehensive overview of the dynamics of language change in the RSC, from the perspective of topical, lexical, and grammatical development.

Challenges for Diachronic Research. Research on language change and specifically on diachronic conceptual shifts in word embeddings undoubtedly faces some methodological obstacles. The two most pronounced ones, treated in both Tahmasebi et al. (2018) and Kutuzov et al. (2018), are the quality of the embedding models and the

Source	Name	Description
Dubossarsky et al. (2015)	law of prototypicality	More prototypical concepts are less likely to shift. But: this is much weaker than initially claimed (Dubossarsky et al., 2017)
Hamilton et al. (2016a)	linguistic drift vs. cultural shift	Nouns tend to shift locally, verbs/adjectives/adverbs tend to shift globally.
Hamilton et al. (2016b)	law of conformity	More frequent words change at slower rates. But: this is largely an artefact of word frequency (Dubossarsky et al., 2017)
Hamilton et al. (2016b)	law of innovation	More polysemous words change at faster rates. But: this is largely and artefact of word frequency (Dubossarsky et al., 2017)
Eger and Mehler (2016)	law of linear decay	Self-similarity decreases linearly over time.
Xu and Kemp (2015)	law of parallel change	Synonyms and antonyms have similar rates of change. First mentioned by Stern (1921).
Degaetano-Ortlieb and Teich (2018)	lexical change varies	The rate of lexical change varies around a certain average value.
Degaetano-Ortlieb and Teich (2018)	grammar consolidates	Over time, fewer syntactic patterns are used more often (i.e., grammatical variation declines steadily).
Bizzoni et al. (2019a)	lexical specialization	Words drift apart over time, especially when they are infrequent.
Bizzoni et al. (2019a)	function word stability	Clusters of function words drift away from other clusters, but do not themselves disperse.

Table 1: 'Laws' of semantic change (*upper rows*) and dynamics of language change in the RSC (*lower rows*).

verification of findings.

Tahmasebi et al. (2018) rightfully criticises that many word embedding models do not differentiate the senses of polysemous words, and do not account for orthographic variation over time. Such embedding models thus do not make full use of their potential expressiveness. In larger corpora such as the Google Books Ngram, which itself has frequently been used for diachronic studies (Kutuzov et al., 2018), this is feasible. However, many diachronic (and especially the domain-specific) corpora are relatively small, so that sense-differentiation is often not feasible because of data sparsity. Another issue of quality arises with the inherent role of word frequency in the construction of word embedding models. Dubossarsky et al. (2017) examine three statistical laws of semantic change (prototypicality, conformity, and innovation) and replicate the original experiments on a control condition. In the experiments, the laws also hold for the control conditions, which leads to a theoretical analysis for the involvement of word frequency in measurements of cosine similarity. The authors conclude that the previously established laws of semantic change are mainly artefacts of frequency inherent to the examined count-based embedding models.[9]

The second methodological obstacle is posed by the large diversity of studies on diachronic conceptual change and the lack of human judgements on historical word meaning. Research in the field is carried out on a wide variety of text resources which all differ in size, time of origin, quality, domain, and other aspects. Diachronic develop-

[9] Dubossarsky et al. (2017) also argue that the similarity of SGNS to a factorized PMI matrix, shown by Levy and Goldberg (2014), may allow implications for frequency effects in SGNS, too.

ments observed in one study are thus often not transferable to other studies. In terms of validation of methods, the time- and domain-specific human-rated data to statistically validate how well a method identifies conceptual shifts usually does not exist (cf. Tahmasebi et al., 2018).

Some solutions to circumvent the issue of missing gold standards are proposed by Kutuzov et al. (2018). A straight-forward approach, taken in Rosenfeld and Erk (2018), is to generate synthetic data in which conceptual shifts are introduced in a controlled manner. Ecologically more valid methods are based on historic facts. Yao et al. (2018) use 'diachronic equivalents' in a time-dependent question answering task; here for example, the answers of time-specific embedding models to the question *Who governed [like Obama-2016]?* can be compared to who actually was the US president at the models' respective times. Such use of historic events and facts as an objective point of reference is used by Degaetano-Ortlieb and Teich (2018), who verify a detected increase in the use of 'oxygen' with the fact that at the same time oxygen was discovered. This type of verification might be carried out as a prediction task, similarly to Kutuzov et al. (2017), who use diachronic word embeddings to predict the time at which certain armed conflicts happened.

3 Data and Resources

3.1 The Royal Society Corpus

The language resource underlying the present work is the Royal Society Corpus (RSC), version 5.1; the most recent release 6.0 is described in full detail in Fischer et al., 2020. The RSC is a diachronic corpus of English texts from the Royal Society of London (both from the *Philosophical Transactions* and the *Proceedings*). Its earliest texts date back to 1665, the most recent ones in version 5.1 are from 1929. As such, the RSC covers 265 years of language in science – a field that has constantly been changing throughout time, both in the discussed topics and the style of scientific discourse. The RSC (version 6.0) is publicly available up to the 1920s[10] (the more recent parts are omitted due to copyright restrictions).

The RSC was constructed from raw files taken from JSTOR[11] (years 1665-1869) and files provided by the Royal Society (1870-1929). After OCR-scanning, errors were either corrected with pattern matching (for JSTOR files, Kermes et al., 2016) or with the Noisy-Channel Spell Checker (otherwise; cf. Fischer et al., 2020). The documents were then annotated in varying granularities with the available metadata, which allows first-glance analyses as well as the construction of sub-corpora. In this work, I will make use of word embeddings trained on the decade-wise sub-corpora. On the token level, the corpus is POS-tagged (Knappen et al., 2017) and annotated with normalized forms and other information. The RSC version 5.1 contains about 91.2M tokens which are distributed relatively unevenly across the various decades. Figure 1 shows the decade-wise token count.

[10] https://corpora.clarin-d.uni-saarland.de/cqpweb/

[11] https://www.jstor.org/publisher/rsl

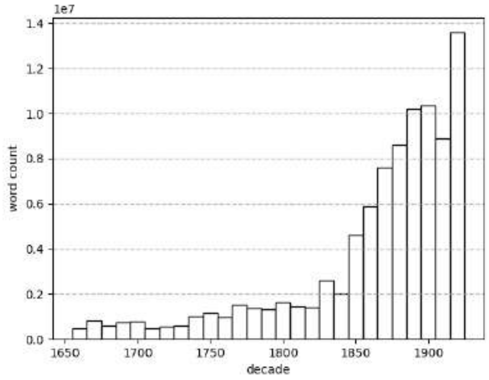
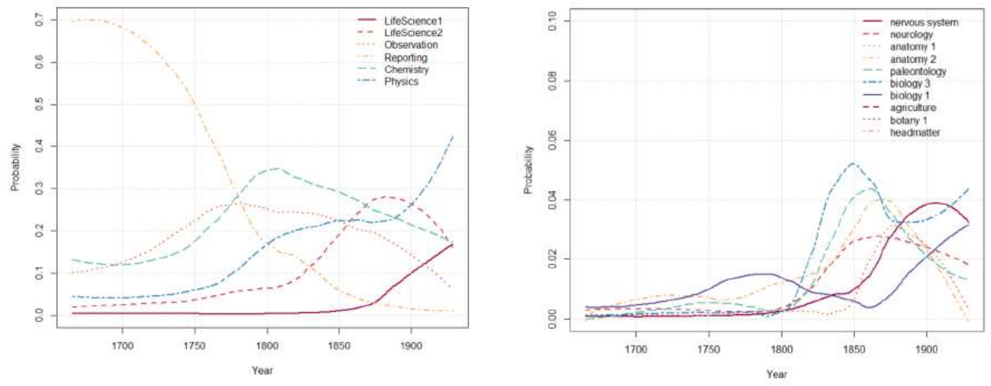


Figure 1: Token counts of the RSC (version 5.1) per decade.



(a) Main topics

(b) Sub-topics of ‘LifeScience2’

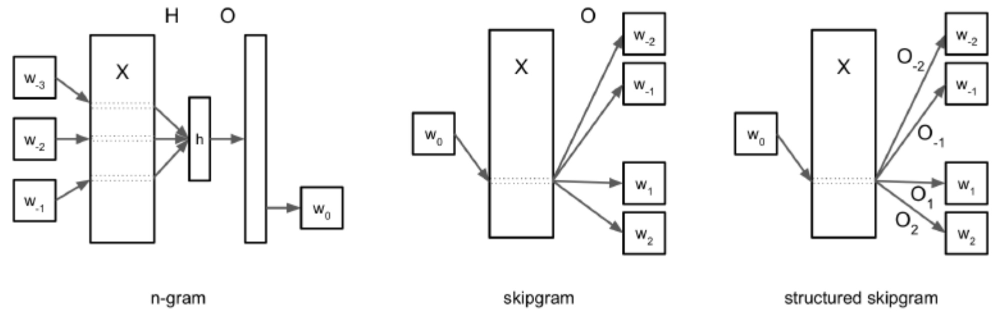
Figure 2: Distributions of main topics in the RSC (version 5.0) over time. Figures taken from Bizzoni et al. (2020).

These metadata at the document level and the token level have made it possible to investigate the scientific English in the RSC in various ways, as discussed in Chapter 2.3. While the corpus is smaller than other diachronic corpora of English such as the COHA (400M tokens, Davies, 2012) or the English Google N-grams corpus (361B tokens, Michel et al., 2011), it is domain specific to science and thus especially well-suited to investigate the development of scientific concepts and scientific discourse. Within the RSC, however, the distribution of scientific topics and sub-topics does change over time, as Figure 2 shows. Especially before and after the 1750s, the majority style of ‘reporting’ gives way to texts which can be attributed to more distinct scientific fields.

3.2 Word Embeddings of the RSC

3.2.1 Background: Model Architectures

Although the word embedding models have been trained prior to this work, I will describe the algorithmic background of their construction in order to give a complete picture of the data at hand.



n-gram

skipgram

structured skipgram

Figure 3: Neural architectures from Bengio et al. (2003), Mikolov et al. (2013b), and Ling et al. (2015). The prediction step after passing **O** is the same in all three architectures; the output representation of size $|V|$ is not shown in the skipgram architectures for better readability.

The word embeddings used here were constructed by means of predictive modeling. That is, the embeddings are the result of a language model that was trained to predict a word, being provided some input (usually one or more words in the near context).

One of the earlier neural language models, introduced by Bengio et al. (2003), achieves this by creating a neural network with a hidden layer that performs n-gram modeling: given n consecutive words $w_{-n}, ..., w_{-1}$, the model predicts the next word w_0 (cf. Figure 3, left). Each of the words in the input is first transformed into a d-dimensional word vector with a matrix $\mathbf{X} \in \mathbb{R}^{d \times |V|}$, where V is the vocabulary of the language model. These vectors are then combined in another (hidden) layer $\mathbf{H} \in \mathbb{R}^{nd \times h}$ which applies a non-linear function (typically *tanh* or *sigmoid*), to form a hidden representation h. For prediction, h is forwarded through an output layer $\mathbf{O} \in \mathbb{R}^{h \times |V|}$, which transforms the hidden representation into a probability distribution over V. All the parameters (i.e. the values in **X**, **H**, and **O**) are adjusted during training with standard stochastic gradient descent and backpropagation. After training, each column x_i in **X** is the embedding of the corresponding word in V.

In their highly influential Word2Vec architecture, Mikolov et al. (2013b) remove **H** and directly forward the word vectors obtained from **X** to make a prediction. This simplification of the architecture makes training considerably faster.[12] However, this is not the only difference; there are three more novelties to Word2Vec. The first one is the prediction task. Bengio et al. (2003) use n-gram prediction (i.e., given n consecutive words, predict the next word); Mikolov et al. (2013b) instead introduce two new tasks as alternatives. The task which is of interest here is called *skipgram*: the model is given one word w_i and has to predict one of the context words within a window of $\pm n$ positions (cf. Figure 3, center). A second novelty from Mikolov et al. (2013b) is subsampling,

[12] In fact, Word2Vec turned out to be influential not only to language modeling (the focus of Bengio et al., 2003), but especially to Distributional Semantics (DS), where it allowed major steps towards the goal of creating one generalized, rather than multiple specialized, models of semantics (cf. Baroni and Lenci, 2010). At its core, Word2Vec is inspired by the notion of DS that context and frequencies matter. This shows (1) in the two prediction tasks introduced in Mikolov et al. (2013b) which both take into account the context to both sides of a word and (2) in the use of downsampling to mitigate frequency effects (both points discussed below).

whereby tokens in the training data are discarded with a certain probability relative to their frequency. This counteracts the negative effects of very frequent words on modeling and also helps to model meaningful relations, as the usually less subsampled content words can take over the context positions of highly frequent words and ‘move into the window’ of other context words (cf. Goldberg and Levy, 2014). The third novelty is the use of negative sampling, which makes training of word embeddings from a certain amount of text more efficient. In principle, negative sampling constructs a set of training instances (i.e., input-target word pairs) which are very likely not in the original corpus. The training objective now becomes slightly different, as the aim is to give a high score v_c to input-context pairs (w, c) which are in the original corpus D and a low score to the constructed pairs in D' (i.e., the negative examples). All put together, the objective is formulated by Goldberg and Levy (2014) as

$$\underset{\theta}{\operatorname{argmax}} \sum_{(w,c)\in D} \log \sigma(v_c \cdot v_w) \; + \sum_{(w,c)\in D'} \log \sigma(-v_c \cdot v_w), \tag{1}$$

where σ is the sigmoid function $\sigma(x) := \frac{1}{1+e^{-x}}$.

Intuitively, word embeddings trained via skipgram with negative sampling (SGNS) represent a word w as a vector x_w which is optimized to describe that w’s textual surroundings in the training data. This is achieved by repeatedly presenting positive as well as negative examples of what such word somewhere in the surrounding of w can be. By discarding many of the high-frequency words before training, frequency effects are mitigated. By constructing a set of negative examples, the available data can be leveraged more effectively.

The architecture of skipgrams is slightly varied by Ling et al. (2015) in order to account for positional information of context words. In their structured skipgrams, there is not just one output layer $\mathbf{O}$. Instead, each of the $\pm n$ context positions in the skipgram model has its own $\mathbf{O}_i$ (cf. Figure 3, right). This allows to make different predictions depending on how far away a context word is and whether it comes before or after the input word. Structured skipgrams have the same complexity as their non-positional predecessor, because each training step only involves one of the $\mathbf{O}_i$. However, each output layer is only involved in a share of the training steps. While Ling et al. (2015) demonstrate a high performance of their structured SGNS on syntax-involving tasks like POS-tagging and dependency parsing, this splitting of prediction layers might lead to slower learning for purely semantic tasks.

3.2.2 Training

The word embeddings used in this work were created following the procedures described in Fankhauser and Kupietz (2017) and first researched in Bizzoni et al. (2019a); they are available online[13]. There are multiple sets of word per-decade embedding models, each set being constructed in a different way. The variables for training one such set of word embedding models are as follows:

[13] http://corpora.ids-mannheim.de/openlab/diaviz1/description.html

- ‘Incremental’ training: This approach follows Dubossarsky et al. (2015), who initialize the training of word embeddings for one time interval with the embedding model from the time interval before. This creates somewhat comparable word embedding spaces. The following will only treat one set of spaces which are not trained in this way; this set is denoted as ‘individual’.

- INIT/NO-INIT: For the incremental approach, the first decade normally has no predecessor on which to base the training. Therefore, the embeddings of the first period (i.e., the 1660s) can be initialized with embeddings which were trained on the entire corpus. The main benefit from corpus-wide initialization is that sparse data problems can be alleviated, which leads to more stable embeddings. The drawback is that the influence of period-specific information may be overshadowed.

- COUNTS/NO-COUNTS: The implementations of Mikolov et al. (2013b) and Ling et al. (2015) use information on token frequency for subsampling, negative sampling, and the adjustment of the learning rate. Token counts are used in any case; here, the distinction is between using each decade’s individual token counts (COUNTS) and using the same, corpus-wide token counts forvevery model (NO-COUNTS). By using the token counts of each sub-corpus when training the corresponding embedding model, these hyperparameters can be set more sensibly.

- STRUCT/NO-STRUCT: Structured SGNS (Ling et al., 2015) captures some syntactic information in addition to the semantics of a term. However, the involvement of multiple prediction layers may lead to slower training and slightly lower scores in purely semantic tasks.

All other hyperparameters are the same for all sets of models: words with less than 5 occurrences are ignored and the threshold for subsampling of frequent words is $1e^{-4}$. The maximum window size is ±5 words. The proportion of negative samples to positive samples is 10. Each model is trained for 25 epochs with the default adaptive learning rate.

3.2.3 Evaluation

To illustrate the effect of the variable hyperparameters, I evaluate the decade-wise embeddings against the MEN similarity test set from Bruni et al. (2014), which contains 3000 word pairs constructed from 751 different words, each labeled with a score between 0 and 50 (Table 2 shows example pairs).[14] In this purely semantic task, word pairs are ranked by the cosine similarity of their embeddings. The model is then evaluated with the non-parametric Spearman correlation ρ between this ranked list and that of the MEN dataset.

[14] The dataset was constructed as follows: pairs were rated by crowdworkers who were presented with two of the pairs at a time and decided which one of the two pairs had a stronger semantic relation. After labeling, each pair was scored by randomly sampling 50 instances of comparison involving the pair in question and counting the number of times that this pair was rated to have the higher semantic relatedness.

I test 5 different sets of models, each with a different configuration of the aforementioned hyperparameters. Four of these sets are trained 'incrementally', i.e. each decade is initialized with the space from the previous decade. INIT-COUNTS-STRUCT is trained by (i) initializing the first decade with a corpus-wide model; (ii) using decade-specific token counts; and (iii) training with structured skipgram. The other three 'incremental' sets of models each ablate one of these options (no corpus-wide initialization; corpus-wide token counts; regular skipgram). For the fifth set, DECADE-INDIVIDUAL, each decade is trained with SGNS solely on its respective sub-corpus. Figure 4 shows the ρ-values across all 27 decades.

The conventional skipgram with corpus-individual counts and corpus-wide initialization achieves the highest performance, with ρ-values between 0.462 (1730s) and 0.539 (1860s). The scores of the structured skipgram variant are similar but slightly lower, between $\rho = 0.425$ (1730s) and 0.491 (1870s). Using corpus-wide token counts instead of decade-specific ones is detrimental to the score (ρ between 0.289 and 0.432), and the embedding models without corpus-wide initialization only achieve reasonable scores ($\rho \geq 0.4$) for the most recent decades. The non-'incremental' variant performs similarly to the non-initialized variant, but the scores of subsequent decades vary more.

The results show that without corpus-wide initialization which leverages all 91.2M tokens, the representations for the early decades cannot express the semantics of their small sub-corpora. With this initialization, however, the models have a decent performance across all decades. To put these scores into perspective: Bruni et al. (2014) report a human performance of $\rho = 0.84$, which could be taken as the upper bound of performance on this test set. Hamilton et al. (2016b) report $\rho = 0.649$ for SGNS embeddings on a 850B tokens sub-corpus of Google books. Levy et al. (2015) report $\rho = 0.723$ of SGNS embeddings on a 1.5B tokens dump of the English Wikipedia. However, these comparisons do not factor in the effects of varying hyperparameters (window size, number of dimensions etc.).

How reliable is this evaluation with the MEN test set? It is important to keep in mind that the word pairs in the MEN test set are labeled by English speakers of the 21st Century while the texts used to construct the embeddings date back several decades and centuries. Additionally, the embeddings from the RSC are domain-specific while the raters of the MEN test set were not, which may result in different conceptual nuances of words. This evaluation will thus not paint a precise absolute picture, but it can validate the quality of the embeddings to a certain extent and indicate differences between decades and training methods relative to each other.

For the work in this thesis, I choose the INIT-COUNTS-STRUCT variant as the default type of embedding model,[15] as it contains relatively reliable word embeddings for all decades — and work with the NO-INIT and NO-COUNTS variants is not feasible across the whole range of the RSC. Another strong reason for this variant and against the conventional skipgram models is continuity of research: previous work on language development in the RSC has used structured skipgrams to investigate syntactic changes additionally to semantic ones. Re-using them here allows to better compare results to

[15] In Chapter 5, I will make use of decade-individual models.

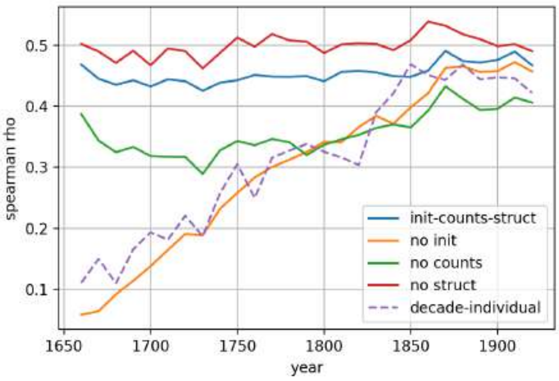


Figure 4: Performance of varying word embedding (WE) models against the MEN test set.

Rank	Word 1	Word 2	Score
1	sun	sunlight	50
5	morning	sunrise	49
10	cat	feline	48
50	flowers	petals	46
100	grapes	vine	44
200	leaf	nature	43
500	bloom	daffodils	39
1000	eat	strawberry	33
2000	dawn	snow	18
3000	bakery	zebra	0

Table 2: Example pairs from the MEN test set.

the previous findings. It also opens up the possibility to investigate semantic changes in words which are important to discourse and syntax. It is true that the INIT-COUNTS-STRUCT embeddings are already somewhat comparable to each other. Still, the post-training alignment of two spaces promises to additionally adjust them more precisely by minimizing shifts between the most stable concepts and at the same time pronouncing the shifts of less diachronically stable concepts.

3.3 Hardware

Experiments are carried out on two machines: my PC and a CPU cluster. The PC has an Intel Core$^{\text{TM}}$ i7-8565U CPU with 4 cores (8 threads) at 1.8 to 4.6GHz and 14.9GB of available RAM. The CPU cluster consists of 4 AMD Opteron$^{\text{TM}}$ 6380 CPUs, each with 16 cores at 2.5 to 2.8GHz and 996GB of available RAM.[16]

4 Alignment of Word Embeddings

This chapter introduces and presents insights about the alignment of word embedding spaces with Optimal Transport via Gromov-Wasserstein distance (GWOT), which was first carried out by Alvarez-Melis and Jaakkola (2018). There are four main questions to be answered:

1. Is it feasible to frame the alignment of diachronic embeddings as a bilingual alignment task?
2. How well does GWOT work with respect to the RSC; is it reliable and feasible?
3. How can embedding spaces be aligned (and concepts be associated diachronically) while preserving diachronic shifts?

[16] While the 64 threads of the CPU cluster seem to allow for fast computations, the optimizations in Chapter 4 take about the same time on both machines. However, the limiting factor for the PC when it comes to larger optimization tasks is RAM. This is no issue with the CPU cluster.

4. Which factors influence the quality of alignment with GWOT, and in what way?

To this end, I will first give the foundations of the relevant techniques in Section 4.1. Section 4.2 aims to answer the question regarding shift-preserving alignment. This is followed by the more practically oriented Section 4.3 which applies the previously introduced techniques to the data at hand, including a coarse analysis of the quality and divergence of the embedding models. The chapter concludes with Section 4.4 which distills the insights. It further presents a selection of pairs of embedding spaces which will be used for further investigation as well as a setting of hyperparameters for their alignment.

4.1 Background

4.1.1 Notation

The following contains an account of definitions and naming conventions which will be used throughout the thesis. It aims to formally connect corpora, embedding spaces, and the main components obtained from GWOT. Table 3 lists the most frequent elements.

The RSC is split into a set of 27 Word2Vec models (or *spaces*), each incorporating one specific decade-long sub-corpus. As the alignment method acts on two spaces at a time, a **pair of spaces** is expressed simply as a tuple of the first years of the spaces' incorporated decades. For example, $\langle 1680, 1770\rangle$ refers to the pairing of the space for the 1680s with the space for the 1770s.

Spaces and **sub-spaces** are denoted as X_{decade} or Y_{decade}; the subscript *decade* denotes the first year of the space's incorporated decade. Formally, a space $X^{m\times d}_{decade}$ is a matrix which is constructed by "stacking" m real-valued vectors of length d. Differently to conventional denotation of matrices as $\mathbf{X}$, the denotation X refers to a matrix of embedding vectors (as opposed to a weight matrix, for example) and is therefore used to refer to spaces. Sub-spaces may only denote the differing m as d is fixed (cf. Section 3.2.1). If not specified, the selection of vectors in a sub-space X^k of X is based on the k most frequent words in the sub-corpus underlying X.

A space has a corresponding **vocabulary** U_X; a set of words u. It has a functional role as well, denoting a mapping from u to the index of the corresponding **vector** (or **point** or **embedding**) x_u in X (analogous with V_Y, v, y_v, and Y). Similarly to the notation of spaces, the embedding vectors are denoted as x rather than $\mathbf{x}$ as their multidimensional nature is irrelevant to most parts of the alignment methods.

The **frequency distribution** f_X maps words $u \in U_X$ onto a number that expresses how often the word appears in the sub-corpus of X. Similarly to this, p and q are **probability distributions** for the vocabularies U_{X^k} and V_{Y^k} of sub-spaces, respectively, normalizing the values of f_X and f_Y over their frequency mass.

Concerning alignment, the central element connecting two sub-spaces X^k and Y^k is a **coupling** $\Gamma^k_{X,Y}$. It is a $k \times k$ matrix of likelihood values. From this coupling, a set T of **translation pairs** $\langle u, v\rangle \in U \times V$ can be constructed. A specific interest lies on the pairs of best mutual translations, therefore T is restricted to those word pairs for which

Component	Definition	Source	Shared	Target
space	$X^{m \times d}_{decade}$	X, m	$d, decade$	Y, n
sub-space	$X^k \subseteq \{x_i \in X\}; \lvert X^k \rvert = k$		k	
vocabulary	$U_X := u \mapsto \mathbb{N}_0^{m-1}; \lvert U_X \rvert = m$	U, u		V, v
word	$u \in U_X$	u		v
vector/point/embedding	$x \in \mathbb{R}^d : X_{U_X(u)} = x_u$	x		y
frequency distribution	$f_X := u \mapsto \mathbb{Q}; u \in U_X$		f	
probability distribution	$p := U_X(u) \mapsto f_X(u) / \sum_{u \in U_X} f_X(u)$	p		q
coupling	$\Gamma^k_{X,Y}$		Γ	
projection matrix	$\mathbf{P}_{X^k Y^l} \in \mathbb{R}^{k \times l}$		$\mathbf{P}$	
translation pairs*	$T_\Gamma := \{\langle u, v \rangle \in U \times V \mid \langle \underset{\Gamma_{V(v),\cdot}}{\text{argmax}}, \underset{\Gamma_{\cdot,U(u)}}{\text{argmax}} \rangle = \langle U(u), V(v) \rangle\}$		T	

Table 3: Overview of frequently used definitions and denotations for the alignment of two embedding spaces ("source" and "target"). *$\Gamma_{i,\cdot}$ and $\Gamma_{\cdot,j}$ are shorthand for selecting row i or column j from Γ.

the coupling maximizes translation likelihood for both words. By solving a Procrustes problem on vectors of such pairs, a **projection $\mathbf{P}^{d \times d}$** from X to Y can be obtained in the form of a matrix. The following sections discuss these processes in more detail.

4.1.2 Orthogonal Procrustes Problem

This mathematical problem (henceforth: *Procrustes*) is a variant of a family of optimization problems in linear algebra called *least square problems*. It was solved by Schönemann (1966) who defined the problem as "transforming a given matrix [$\mathbf{X}$] into a given matrix [$\mathbf{Y}$] by an orthogonal transformation matrix [$\mathbf{P}^*$] so that the sums of squares of the residual matrix [$\mathbf{E} = \mathbf{X}\mathbf{P}^* - \mathbf{Y}$] is a minimum". More formally, we seek to find an orthonormal matrix

$$\mathbf{P}^* = \min_{\mathbf{P}} ||\mathbf{X}\mathbf{P} - \mathbf{Y}||^2_F, \tag{2}$$

subject to $\mathbf{P}^\top \mathbf{P} = \mathbf{P}\mathbf{P}^\top = \mathbf{I}$, where $|| \cdot ||_F$ is the Frobenius norm $||\mathrm{A}||_F = \sqrt{\sum_{i,j} |a_{i,j}|^2}$ (cf. Alvarez-Melis and Jaakkola, 2018). The general solution under these circumstances can be computed efficiently with singular value decomposition (SVD, cf. Klema and Laub, 1980):

$$\mathbf{X}\mathbf{Y}^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \tag{3}$$

where

$$\mathbf{P}^* = \mathbf{U}\mathbf{V}^\top. \tag{4}$$

In the context of the alignment of two word embedding spaces X and Y, the solution of the Procrustes problem is an orthogonal matrix $\mathbf{P}$ that can be used to project any

$x \in X$ onto Y so that $\mathbf{P}x$ is comparable to any $y \in Y$ (and of course, any projected x as well). $\mathbf{P}$ is obtained via SVD on two sub-spaces X^k and Y^k of k vectors where x_i and y_i express similar concepts.[17]

Procrustes is thus only solvable if there is a known set of pairs of points which correspond to one another. In other words, some form of bilingual signal is needed. If only a coarse mapping between X and Y is needed, this can be achieved with small and/or hand-curated data. For example, Zhang et al. (2016) use translations of the ten most frequent words in the target language. In a weakly supervised approach, Smith et al. (2017) propose to construct a 'pseudo-dictionary' of words that appear in both vocabularies. Artetxe et al. (2017) go one step further and construct the initial bilingual signal of their iterative approach purely from shared numerals.

These approaches work reasonably well for dictionary induction and other bilingual tasks and suggest a straight-forward solution for the alignment of RSC spaces: that is, to pair up embeddings of string-matching words and then solve Procrustes. However, the assumption of bilinguality (i.e., two snapshots of a language from different points in time should be treated as two different languages), does not allow to take this simple solution. In principle, there are two reasons which speak for making this assumption. First, it would be wrong to assume that two words with identical spelling always express the same general concept. Although the spaces at hand are all from more or less the same language and domain (namely, British English of scientific texts) and show a large overlap in vocabulary, the lexical and semantic diachronic shifts cause the spaces to diverge from one another. Pairing embeddings by the shared orthography of their labels rather than their position in space would thus introduce errors. Second, in order to detect semantic shifts, it is of interest to align spaces only along a selected subset of embeddings, for example along the most stable concepts. With Procrustes alone, there is no notion of stability by which to filter the bilingual signal. One might prioritize pairs of more frequent words to less frequent ones, but even highly frequent words are often not modeled reliably (cf. Wendlandt et al., 2018) and within a set of frequent and thus relatively stable words, some may experience more diachronic change than others.

With the conservative attitude that assumes bilinguality and for the sake of finding the best possible anchor points, the alignment task here requires an approach to alignment that does not assume a shared vocabulary and incorporates a reliable notion for the diachronic stability of concepts. Procrustes alone thus does not suffice for the task at hand.

4.1.3 Optimal Transport with Gromov-Wasserstein Distance

A solution that satisfies all these requirements is presented by Alvarez-Melis and Jaakkola (2018), who frame the alignment task as an Optimal Transport (OT) problem that relates two embedding spaces to one another by minimizing the Gromov-Wasserstein (GW) distance between their respective metric spaces. OT itself is a very active field and has a broad range of applications, for example in object matching (Mémoli, 2011, Haker et al.,

[17] Ideally, x_i and y_i should express *exactly the same* concept, but this is extremely unlikely. In practice, the approximation with *highly similar* concepts works well enough for most applications.

2004), economics (Koopmans, 1949), or traffic modeling (Helbing, 2013). The following briefly introduces OT and GW with respect to the task at hand. The aim is to provide an intuition, a formal frame and algorithmic solutions for the method. As most of the information presented and several implementations throughout this thesis are adapted from Alvarez-Melis and Jaakkola (2018), I will use their notation. Further details are mostly taken from Peyré and Cuturi (2019), which provides a thorough account of OT.

Monge's OT Problem. Optimal Transport (OT) is a family of optimization problems that aim to find a mapping between two distributions that minimizes a certain cost function. OT is often introduced informally with the example of Gaspard Monge (1746-1818), in which a worker needs to move a pile of sand from one location to another with a shovel.[18] The "source" pile as well as the "target" pile each have a distinct shape which is known (i.e., they are *distributions* of sand). The worker wants to spend as little energy on this task as possible. The aim is to find a *transportation plan* (i.e., a description of how many shovels of sand should be carried from each part of the source pile to each part of the target pile) which minimizes the distance that the worker has to walk between the two locations.

In the context of OT for two word embedding spaces X and Y, the two discrete distributions μ, ν are defined over subsets $X^m \subseteq X, Y^n \subseteq Y$ of embedding vectors:[19]

$$\mu = \sum_{i=1}^{m} p_i \delta_{x^{(i)}}, \quad \nu = \sum_{j=1}^{n} q_j \delta_{y^{(j)}}, \tag{5}$$

where p and q are non-negative weight vectors associated with X^m and Y^n, respectively. The discreteness of the distributions is formally expressed by the Dirac function δ, which allows to distribute portions of probability mass over a finite set of points rather than over a continuous range.

These two distributions need to be associated. In the Monge problem from 1781 (cf. Peyré and Cuturi, 2019), each x_i needs to be assigned to exactly one y_j so that the combined mass of all x_i assigned to a particular y_j is equal to that y_j's mass.[20] The function $T_\sharp$ that verifies this is called a *transportation map* and defined as

$$T : \{x_1, ..., x_m\} \to \{y_1, ..., y_n\} \;\mid\; \forall j \in \{0, ..., m\}, \;\; q_j = \sum_{i:T(x_i)=y_j} p_i \tag{6}$$

There could be multiple such transportation maps for which $T_\sharp\mu = \nu$ (i.e., which transport the mass of μ to ν), but the aim is to find the one which minimizes the *cost*

[18] For this reason, it is also known in the field of computer vision as *earth mover's distance* (cf. Peyré and Cuturi, 2019)

[19] While the formulations of OT problems generally allow $m \neq n$, in practice, many applications set $m = n$, including the computations in this work.

[20] As there is no notion of "mass" for word embeddings in this context, assume that the mass is distributed uniformly across all x_i and y_j, which only leaves p and q for comparison.

$c(x, T(x))$ of transportation:

$$\min_T \left\{ \sum_i c(x_i, T(x_i)) \mid T_\sharp \mu = \nu \right\} \tag{7}$$

Note that this is only possible for discrete measures. For problems involving continuous probability measures, the solution to the Monge problem is

$$\min_T \left\{ \int_X c(x, T(x)) d\mu(x) \mid T_\sharp \mu = \nu \right\}. \tag{8}$$

Before discussing the choice of cost function, the Monge OT problem needs to be modified as it is not suited for the task at hand. This has multiple reasons, but most importantly, it defines $T_\sharp$ to be surjective, meaning that OT between word embedding spaces assignins exactly one target word to a given source word. This is problematic, because words and concepts can have multiple corresponding counterparts across natural languages.

Probabilistic OT: Kantorovich's Relaxation. Instead of the deterministic nature of Monge's OT problem, Kantorovich's approach from 1942 considers OT to be a *probabilistic* problem in which the mass from a certain source point can be split across several target destinations (cf. Peyré and Cuturi, 2019). This leads to several re-formulations.

Instead of the transportation map $T_\sharp$, the connection between two distributions now is a *coupling* in the form of a matrix $\Gamma \in \mathbb{R}_+^{m \times n}$. The set of possible couplings is defined as

$$\Pi(p, q) = \left\{ \Gamma \in \mathbb{R}_+^{m \times n} \mid \Gamma \mathbb{1}_n = p, \ \Gamma^\top \mathbb{1}_m = q \right\}, \tag{9}$$

where $\mathbb{1}_k$ is a k-dimensional vector of ones. In other words, the probability mass of p_i for point x_i is represented as a distribution over the n target points. Conversely, each of the n columns of Γ is a distribution which expresses how much of the mass q_j is received from each of the m source points. Returning to the example of the worker moving sand from one pile to another, the coupling cell Γ_{ij} prescribes how much sand (in terms of grains, shovels, ...) should be moved from point i to point j. In the context of OT for word embedding spaces, Γ_{ij} can be interpreted as the likelihood that x_i is a translation of y_j.[21]

Similarly, the transportation cost in this setting is given as a matrix $\mathbf{C} \in \mathbb{R}_+^{m \times n}$. The optimization problem with Kantorovich's relaxation is thus to determine Γ^*:

$$\Gamma^* = \min_{\Gamma \in \Pi(p,q)} \langle \Gamma, \mathbf{C} \rangle := \sum_{ij} \Gamma_{ij} \mathbf{C}_{ij}. \tag{10}$$

Algorithmic solutions to this problem and further variations are described in more detail in Peyré and Cuturi (2019). There are several programming libraries that implement optimizers for a wide variety of OT problems, for example the POT library

[21] This probabilistic formulation makes any coupling Γ symmetric so that $\Gamma^\top$ can be used analogously for OT from Y to X.

(Flamary and Courty, 2017), which is used by Alvarez-Melis and Jaakkola (2018) and thus in this work as well.

Faster Computation with Regularization and Matrix Scaling. There is one important modification to the objective in (10) which has algorithmic implications, namely regularization. This is commonly done with *entropy regularization*, first introduced by Cuturi (2013), of the coupling:

$$\Gamma^* = \min_{\Gamma \in \Pi(p,q)} \langle \Gamma, \mathbf{C} \rangle - \lambda H(\Gamma), \tag{11}$$

where λ is used for scaling the entropy of the coupling matrix, defined as

$$H(\Gamma) = -\sum_{ij} \Gamma_{ij}(\log(\Gamma_{ij}) - 1). \tag{12}$$

Intuitively, this regularization term allows solutions that make use of the probabilistic nature rather than a deterministic one and leads to couplings with lower sparsity, i.e., overall smoother distributions, which shows in visualizations of such couplings as a blurring effect (cf. Peyré and Cuturi, 2019).

Decreasing the sparsity of couplings has a second effect: it allows for faster computation. Specifically, (11) can be solved efficiently using a Gibbs kernel $\mathbf{K}$ that is associated with the cost matrix $\mathbf{C}$:

$$\mathbf{K} \in \mathbb{R}^{m \times n} := \mathbf{K}_{ij} = e^{-\frac{\mathbf{C}_{ij}}{\lambda}}. \tag{13}$$

In combination with two vectors $\mathbf{a} \in \mathbb{R}_+^m$ and $\mathbf{b} \in \mathbb{R}_+^n$ which function as scaling variables, this allows to re-write the couplings as defined in (9) as

$$\Gamma = \operatorname{diag}(\mathbf{a})\ \mathbf{K}\ \operatorname{diag}(\mathbf{b}). \tag{14}$$

Now, instead of a matrix Γ with mn parameters, the goal is to find the right vectors $\mathbf{a}$ and $\mathbf{b}$, which amounts to optimizing $m + n$ parameters. The two variables are bound to satisfy the mass conservation constraints to couplings (cf. (9)):

$$\operatorname{diag}(\mathbf{a})\ \mathbf{K}\ \operatorname{diag}(\mathbf{b})\mathbb{1}_n = p, \quad \operatorname{diag}(\mathbf{b})\ \mathbf{K}^\top \operatorname{diag}(\mathbf{a})\mathbb{1}_m = q. \tag{15}$$

A solution to (11) with the formulation of Γ as in (14) is possible with the Sinkhorn-Knopp algorithm, which first initializes $\mathbf{K}$ in dependence to $\mathbf{C}$ and $\mathbf{b}$ as a uniform vector (e.g., $\mathbf{b}_0 := \mathbb{1}_n$) and then iteratively updates $\mathbf{a}$ and $\mathbf{b}$ until convergence or a certain number of iterations t is reached:

$$\mathbf{a}_{t+1} := \frac{p}{\mathbf{K}\mathbf{b}_t}, \quad \mathbf{b}_{t+1} := \frac{q}{\mathbf{K}^\top \mathbf{a}_{t+1}} \quad \text{(entry-wise division).} \tag{16}$$

This matrix scaling procedure is much faster than linear programs: while a linear program for (10) has a complexity of $O(n^3 \log(n))$ (Alvarez-Melis and Jaakkola, 2018), the Sinkhorn-Knopp algorithm can approximate (11) to τ in $O(n^2 \log(n)\tau^{-3})$ (Peyré and Cuturi, 2019). Cuturi (2013) demonstrate the acceleration empirically and show that it is “several orders of magnitude faster”.

Cost Functions and Gromov-Wasserstein Problem. A central component to optimization, namely the cost function, has been left out so far for the sake of simplicity, because the choice of cost function depends on the nature of the distributions to be coupled. In the example of the worker re-allocating sand from one pile to another, the cost for transporting some sand from point $x_i \in \mu$ to point $y_j \in \nu$ might be defined as a combination of distance and amount of sand carried from x_i to y_j, or more simply, as just the distance walked from x_i to y_j.

For Monge problems like (8), the cost function c associated with a transportation plan T from μ to ν is typically defined as $c(x, T(x)) = \|x - T(x)\|$. For Kantorovich problems like (10), the cost of transportation between the points in μ and ν is represented as a matrix $\mathbf{C} \in \mathbb{R}^{m \times n}$ and typically defined as $\mathbf{C}_{ij} = \|x_i - y_i\|$. In the context of OT between word embeddings, this translates to the cosine distance d_{cos} between vectors, a metric commonly used to measure the similarity of word embeddings to each other. The cost matrix can thus be defined as

$$\mathbf{C} \in \mathbb{R}^{m,n} := \mathbf{C}_{ij} = d_{cos}(x_i, y_j) = 1 - \frac{x_i \cdot y_j}{\|x_i\|_2 \|y_j\|_2}, \tag{17}$$

where $\| \cdot \|_2$ is the euclidean distance. Such a cost matrix is then used together with a chosen regularization λ to initialize the Gibbs kernel $\mathbf{K}$, and apply the Sinkhorn-Knopp algorithm. This definition of the cost function, however, relies on the assumption that X and Y share the same metric space, i.e., for all dimensions d of X and Y, $d_i(X)$ and $d_i(Y)$ express the same component of meaning. This is not generally the case, especially in bilingual scenarios such as in Alvarez-Melis and Jaakkola (2018) which aim at the coupling of two independently trained embedding spaces.[22]

This complication requires a cost function which abstracts from the individual spaces and allows to compare points across spaces independently from their absolute positions. The key insight is the observation that word embeddings of similar concepts tend to have similar geometric arrangements, independently from the space in which they exist. For example, for two embedding spaces X and Y, the two embedding vectors x_{katze} and y_{gatto} of the concept “cat” each have a similar distance to x_{hund} and y_{cane} (“dog”), respectively, a similar distance to x_{schuh} and y_{scarpa} (“shoe”), and so on. This observation has previously been exploited by Mikolov et al. (2013a) for bilingual lexicon induction. Alvarez-Melis and Jaakkola (2018) use it to relate the points in X and Y in terms of their geometric arrangement (or: *spatial role*), which is defined by their distances to other points within their space.

The Gromov-Wasserstein (GW) distance was introduced by Mémoli (2011). At first glance, the Gromov-Wasserstein problem for spaces X, Y is an OT problem that seeks to find the optimal coupling by means of two distance matrices $\mathbf{C} \in \mathbb{R}^{m,m}$ and $\mathbf{C}' \in \mathbb{R}^{n,n}$ over X and Y, respectively. Intuitively, each row (or column) $\mathbf{C}_i$ can be interpreted as the “profile” of distances of x_i to other points in X. The most straightforward choice

[22] Alvarez-Melis et al. (2019) provide a generalized approach to solve this problem, but the solution from Alvarez-Melis and Jaakkola (2018) is sufficient for the task at hand.

of within-space distance measure for $\mathbf{C}$ and $\mathbf{C}'$ is the cosine distance as defined in (17). These distances serve as inputs for the cost function $l : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, which is typically defined as either the "square loss" $\frac{1}{2}(a-b)^2$ or the Kullback-Leibler divergence $KL(a|b)$ (Alvarez-Melis and Jaakkola, 2018). The computed values can be stored in a 4-th order tensor $\mathbf{L} \in \mathbb{R}^{m \times m \times n \times n}$. Written in length, the square loss function for the GW problem on cosine distance matrices computes

$$\mathbf{L} \in \mathbb{R}^{m \times m \times n \times n} := \mathbf{L}_{ijkl} = l(\mathbf{C}_{ik}, \mathbf{C}'_{jl}) = \frac{1}{2} \left(d_{cos}(x_i, x_k) - d_{cos}(y_j, y_l) \right)^2 . \tag{18}$$

$\mathbf{L}_{ijkl}$ can be interpreted as the share of the cost of coupling x_i with y_j which is incurred by associating two within-space distances of their profiles (namely those distances to x_k and y_l, respectively) to one another. Another interpretation is that of the cost of pairing up the start points and the end points of the two distances (cf. Alvarez-Melis and Jaakkola, 2018). By minimizing this cost, the solution to the GW problem prescribes how the points in X should be associated to those in Y so that their profiles of distances match optimally. The discrete version of the GW problem is thus formally defined as

$$\Gamma^* = \mathrm{GW}(\mathbf{C}, \mathbf{C}', p, q) := \min_{\Gamma \in \Pi(p,q)} \sum_{i,j,k,l} \mathbf{L}_{ijkl} \Gamma_{ij} \Gamma_{kl}. \tag{19}$$

This problem is non-linear and non-convex and thus mathematically much harder than the Kantorovich OT problem in (10). It is also computationally more complex, because it requires operations on the 4-th order tensor $\mathbf{L} \in \mathbb{R}^{m \times m \times n \times n}$, which leads to a complexity of $O(m^2 n^2)$. Peyré et al. (2016) propose to solve (19) with projected gradient descent, which involves solving a common OT problem in each iteration. They additionally show that the initialization of the Gibbs kernel $\mathbf{K}$ for these iterations can be accelerated for certain cost functions: If the loss can be written as $l(a, b) = f_1(a) + f_2(b) - h_1(a)h_2(b)$ for functions (f_1, f_2, h_1, h_2), then for all $\Gamma \in \Pi_{(p,q)}$ a pseudo-cost matrix $\hat{\mathbf{C}}_\Gamma$ can be computed, which has the form

$$\hat{\mathbf{C}}_\Gamma(\mathbf{C}, \mathbf{C}', \Gamma) = \mathbf{C}_{xy} - h_1(\mathbf{C}) \; \Gamma \; h_2(\mathbf{C}')^\top, \tag{20}$$

where $\mathbf{C}_{xy}$ is a matrix of fixed values that incorporates p and q:

$$\mathbf{C}_{xy} = f_1(\mathbf{C}) p \mathbb{1}_n^\top + \mathbb{1}_m q^\top f_2(\mathbf{C}')^\top . \tag{21}$$

Such a decomposition is possible for example for the square loss function from (18)[23] and allows to solve the problem in $O(m^2 n + mn^2)$ instead of $O(m^2 n^2)$.

In the context of OT for word embeddings, the solution to the GW problem on X and Y is a coupling Γ that provides a direct probabilistic alignment between the two coupled spaces. That means that each cell Γ_{ij} can be interpreted as the likelihood that the

[23] The functions used in the implementation of Alvarez-Melis and Jaakkola (2018) are $f_1(a) = \frac{a^2}{2}$, $f_2(b) = \frac{b^2}{2}$, $h_1(a) = a$, and $h_2(b) = b$. Note that their example algorithm uses a slightly different cost function than (18) which decomposes to $f_1(a) = a^2$, $f_2(b) = b^2$, $h_1(a) = 2a$, and $h_2(b) = b$.

words of x_i and y_j are translations of each other (Alvarez-Melis and Jaakkola, 2018). I will henceforth refer to the optimization of a coupling by solving the GW problem as "GWOT". With the intuitions and formal background given in this section, the next steps are to adapt this alignment method to the diachronic scenario and to gain empirical insights with respect to the available word embedding models.

4.2 Two Connections Between Spaces

In the context of this thesis, the difference between two spaces X and Y is always *diachronic*, that is, X and Y express concepts on the bases of text dating to different points in time. By convention, X is the 'earlier' space and Y is the 'later', 'more recent' space.

Performing GWOT on two word embedding spaces X and Y (or: sub-spaces) results in a coupling Γ which can be interpreted as a table of translation likelihood between all possible word pairs. This allows to create a dictionary T of bidirectional translations: if a coupling cell Γ_{ij} holds the maximum value for both row i and column j, then the corresponding words $u \in U_X$ and $v \in V_Y$ form a pair (u, v) of *mutual best translations.*

This is the first of two connections between the two spaces: for each of these translation pairs, Γ claims that across the two (sub-) spaces, the behavior of u in X is most similar to the behavior of v in Y and vice-versa. This approach is clearly more complex and time-consuming than the usual approach to diachronic lexical semantic research, in which the two embeddings x_w and y_w of the same word w (i.e., the same string) are compared. However, a key point of reasoning here is that the language modeled by two embedding spaces should *not* the assumed to be the same, even if they both express a widely similar language and the only major difference seems to be the time in which their underlying corpora were produced. Orthographic reforms and other dynamics of lexical change affect the surface form of concepts and may also affect the concepts themselves, but in these cases, it is difficult to pair up the two concepts, because their labels do not match. GWOT is agnostic of surface forms; it relates embeddings based on how similarly they behave within their space rather than whether their labels spell the same. As shown in Section 4.3.7, this can overcome effects of orthographic change and hint towards lexical change. The first connection between two spaces, namely the set T of translation pairs, is thus a lexical connection obtained from a coupling, based on conceptual similarity.

The second connection is that between the word embeddings themselves; it is the projection matrix $\mathbf{P}$. Alvarez-Melis and Jaakkola (2018) obtain this simply by solving a Procrustes problem on the translation pairs obtained from Γ. Their motivation for this step is scaling: for bilingual lexicon induction (BLI), a coupling of the entirety of X and Y would suffice, but GWOT becomes increasingly complex for larger sub-spaces, making a full-fledged optimization on entire embedding spaces unfeasible (cf. Section 4.3.2). A reasonably large bilingual signal obtained from a coupling between sub-spaces X^k and Y^k provides enough information to create a projection matrix that precisely maps between the two spaces so that BLI can be carried out via nearest-neighbor search. The motivation for the task of conceptual semantic shift detection is that embeddings can be compared across spaces even if their labels are not contained in both spaces' vocabularies. In this case, one can project the embedding of u from X to Y and then compare: $\mathbf{P}x_u$ to y_v.

The crucial requirement for $\mathbf{P}$ if it is used for the detection of shifts is *restricted optimality*: $\mathbf{P}$ should not align spaces perfectly, but rather project X along those points which change the least with respect to Y. By restricting the bilingual signal for Procrustes to the word pairs of the most stable concepts, the differences between other, less stable concepts are preserved. The selection of these "anchors" is straightforward, as the translation likelihood values in Γ provide a direct measure for the stability of translation pairs. The intuition here is that two vectors of a stable concept have relatively similar profiles of distances to other vectors, even if those other vectors are less stable. GWOT thus assigns a large translation likelihood score to the two vectors in question (i.e., declares that a lot of probability mass should be moved exclusively between these two points). Note that this notion of stability promises to be more robust than the assumption that corpus frequency correlates with embedding stability. While Hamilton et al. (2016b) demonstrate that more frequent words show less conceptual change, Wendlandt et al. (2018) claim that the frequency of a word is not a major factor regarding the reliability of its vector representation. This means that embeddings of highly frequent and thus semantically rather stable concepts can still have unstable embeddings. The scores of Γ take into account both sources of divergence – the actual semantic shift of a concept as well as the potential inaccuracy of its embedding.

4.3 Application of GWOT to the RSC Embedding Spaces

With the two connections between X and Y, namely the set T of translation pairs and the projection matrix $\mathbf{P}$ which is non-optimal on purpose, there are various possibilities for diachronic investigations. Before turning to such experiments, however, there are multiple considerations to be made regarding the parameters for GWOT as well as certain hyperparameters such as the number of anchor terms for $\mathbf{P}$. As the particular embedding spaces certainly have an influence on the feasibility of the previously discussed theory, these considerations are best made after preliminary investigations, which will be the matter of the following sections.

Based on the insights from Alvarez-Melis and Jaakkola (2018) on GWOT for word embeddings, the following experiments are carried out with normalized embeddings. The intra-space distance matrices, computed at the initial step of GWOT, are normalized as well. The entropy regularization parameter λ, however, will differ.

4.3.1 Use of String Matching among Translation Pairs

While alignment of embedding spaces is being treated as a bilingual task, the fact that both parts of the alignments in question stem from largely the same language is convenient for practical reasons: there is a very large overlap in vocabulary. A qualitatively good coupling of two spaces is bound to relate embeddings with the same label to one another, leading to many translation pairs which are string-matching. Moreover, a translation pair's score can be interpreted as the confidence with which the coupling

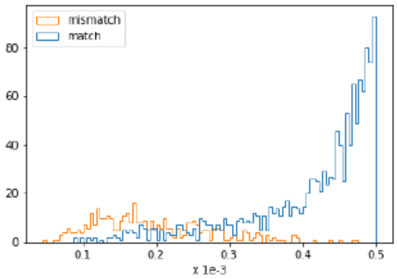


Figure 5: Example histogram of string-matching (blue) and mismatching (orange) pairs by coupling score (x-axis).

rank	1800s	1830s	rank	1800s	1830s
400	current	stream	1024	materially	greatly
448	lately	recently	1034	constituent	elementary
636	tympani	marrow	1040	broad	wide
739	happens	occurs	1062	native	british
859	hon.	rev.	1063	annual	diurnal
954	related	detailed	1082	acquainted	aware
957	velocities	intensities	1088	arguments	data
975	resemble	constitute	1089	electrified	charged
978	firmly	closely	1090	signs	sign
983	smallest	slightest	1107	emitted	excited

Table 4: Highest-scoring mutual translation pairs with string mismatches obtained from a coupling of $\langle 1800, 1830 \rangle$ (uniform p/q, k=2000).

relates two words. A reliable coupling should thus yield matching pairs with high scores and mismatching ones with lower scores, leading to a distribution that can be nicely partitioned by means of their score.

In Section 4.3.3 I will use this to carry out sanity checks for couplings of various space pairs. As a simple form of visualization, Figure 5 depicts a histogram of the number of matching and mismatching translation pairs as a function of their translation score.

The opportunity to use surface forms as a source of information leads to a second possible use valuable to the discovery of language changes: string-mismatching pairs with a high translation likelihood can be seen as indicating lexical shift. In these cases, two concepts are coupled because they take on a very similar role in their respective embedding spaces; however, their labels are not the same. In other words, the referring expression of a concept changes while the concept itself does not. Table 4 illustrates this by showing the highest-scoring mutual translation pairs from a small coupling of ⟨1800, 1830⟩. Mismatches that are of interest occur predominantly for pairs which are (1) semantically similar (*current* and *stream*), (2) semantically related (*annual* and *diurnal*), (3) morphologically different (*signs* and *sign*), or (4) orthographically different (*connexion* and *connection*, not listed). Out of these, semantically similar pairs (2) can inform about possible lexical shifts. At the same time, it is important to recognize especially the semantically close word pairs as not necessarily shift-indicating and in some cases (such as for antonyms) to treat them as noise (Section 4.3.7 discusses this in more detail).

Bear in mind that although comparing the string representation of words to one another can lead to various insights, it is not a valid source of information under the assumption of bilinguality. On the one hand, such insights about the general behavior of GWOT can influence the decisions involved in the alignment process and further steps, e.g. choices of hyperparameters. On the other hand, string comparisons should not directly inform about the bilingual signal T, conceptual shifts, or alignments.

4.3.2 Complexity Constraints of GWOT

Optimizing distances between distances (cf. Alvarez-Melis and Jaakkola, 2018), GWOT is a quadratic optimization problem. Although it can be approximated efficiently, the amount of memory and especially computations needed to align any pair of word embedding spaces with reasonably sized vocabularies in full is very high.

With the available hardware, the GWOT on sub-spaces of size k=20000 requires about 21 hours (Table 16 in the Appendix reports the times taken for the optimization of couplings of various sizes). This shows that GWOT does not scale well to complete word embedding spaces which usually comprise embeddings for well above 30K words. A coupling between two spaces will thus necessarily always be optimized on sub-spaces of a certain size k.

This leads to the question of what will be a reasonable coupling size k for the task at hand. For their bilingual lexicon induction (BLI) task, Alvarez-Melis and Jaakkola (2018) optimize couplings of 20K words. Differently to BLI, the aim here is not to induce a set T of translation pairs, but rather to directly use the pairs obtained from the coupling as bilingual signal for shift detection. The coupling size thus also puts an upper limit to the number of concepts that can be investigated for shifts. Nonetheless, 20K appears to be a sensible size for the task at hand. With the available hardware, the 21h taken for an optimization with 300 iterations is acceptable. The number of translation pairs retrieved from such a coupling varies and depends on the spaces which are coupled. Preliminary tests on various pairs of RSC spaces showed that this number ranges from 6.6K to 16.4K. Assuming that most of the interesting (i.e., relevant and non-arbitrary) shifts occur for moderately frequent concepts, $|T| \geq$ 10K should include most cases of interesting shifts. Therefore, the default size of a "large coupling" will henceforth be 20K.

4.3.3 Divergence of RSC Spaces

Intuitively, the differences between two languages (or language uses) tend to increase along with their temporal distance. In context of the RSC, it is reasonable to assume that two spaces (more specifically: their vocabularies and the embeddings of shared words) from far-apart decades differ more than a space pair of neighboring or close-by decades.

There are 27 decade-spanning embedding spaces at hand, or 351 possible space pairs. A thorough analysis of all of these pairs would exceed the scope of this thesis; a selection of a few pairs of spaces is required. Alvarez-Melis and Jaakkola (2018) use the objective of alignment, the Gromov-Wasserstein distance D_{GW}, as a metric for the similarity of two spaces: highly similar spaces tend to align well, which is captured by a small D_{GW}. However, a low D_{GW} does not necessarily imply a large, high-quality set of translation pairs which is needed as one of the two connections between two spaces.

Conceptual divergence. Therefore, I investigated the feasibility of alignment by means of small-scale couplings, evaluating both by D_{GW} of each coupling (i.e., the overall cost of transportation) as well as string-matching of translation pairs. For each pair of spaces, I optimized two couplings of the 1K[24] most frequent *shared* words; one of the two couplings was initialized with uniform p/q, the other one with distributions drawn from log-flattened corpus frequencies (cf. Section 4.3.4). The quality of a coupling is then estimated by the number of string-matching translation pairs relative to the coupling size. As the underlying vocabularies for both sides of the coupling are the same, a truly optimal coupling will lead to 1K matching translation pairs, while couplings of lower quality will lead to fewer matching pairs and/or fewer pairs in general.

What is the intuition behind the number of string-matching pairs as estimator for a coupling's quality? By optimizing on a shared vocabulary, it is certain that both subspaces X and Y comprise the embeddings of the same 1K words. In the case that X and Y belong to close-by decades with only small differences in language, the profiles of distances of x_w in X and of y_w in Y for one and the same word w will tend to be similar to each other, leading to a higher translation score and a greater chance to become a translation pair. If X and Y belong to decades with greater linguistic differences, these profiles will diverge more, making it more difficult for GWOT to associate x_w in X with y_w in Y. This decreases the chances of string-matching translation pairs as well as the overall chances for finding translation pairs. Both consequences are captured by the number of string-matching pairs relative to the coupling size.

Figures 6a and c show the quality estimations for 1K-couplings of all combinations of the RSC's decade-wise word embeddings, for optimization once without and once with frequency information. This confirms the intuition: couplings for pairs of decades close to each other (depicted along the diagonal) allow for near-to-perfect sets of translation pairs. In contrast, the couplings for pairs of spaces which are more than 5 decades apart yield fewer matching translation pairs, with exception of the five latest decades, which appear to be slightly easier to align to one another in all settings (Figures 6a – e).

Spaces between 1690 and 1750 tend to be more difficult to align to spaces between 1810 and 1850 than other pairs, clearly indicated by the blue patch in Figure 6c. This may indicate that the 'mid-section' of the RSC experiences more language change than the early and the late periods, and that within this section, farther-apart spaces are naturally more different than close-by pairs. At the same time, the lighter colors in the bottom-left corners of the two heat maps indicate that couplings to the recent spaces (1850s or later) work better for the 17th century spaces than for spaces of the 18th century. I attribute this observation in part to the incremental training method of the embedding models: the 1660s space is initialized on a corpus-wide model which is conceptually biased towards the most recent decades, because these contribute the most text to the model. The re-training of the 1660s space and subsequent spaces does adjust the embedding vectors; however, it is possible that the 'legacy of initialization' is still present in these early spaces and allows GWOT to find mutual translation pairs more reliably than for spaces of the mid-section of the RSC. Lastly, the differences in scores between the two figures underlines the slight negative effect of frequency information on medium-to-large-sized couplings (cf. Section 4.3.4).

[24] The choice of 1K as size for the couplings is based on considerations of complexity and expressiveness. While larger couplings operate for more words than smaller ones and are thus challenged more to pick the correct (i.e., string-matching) word pairs, they are considerably more expensive to optimize.

These observations on translation pairs are supported by Figure 6e, which shows the D_{GW} values associated with the couplings from Figure 6a. However, the two reported measures do not correlate perfectly. For example, some of the couplings of relatively difficult (i.e., distant) pairs such as from the first half of the 18th century to the 20th century yield a reasonable number of matching translation pairs.

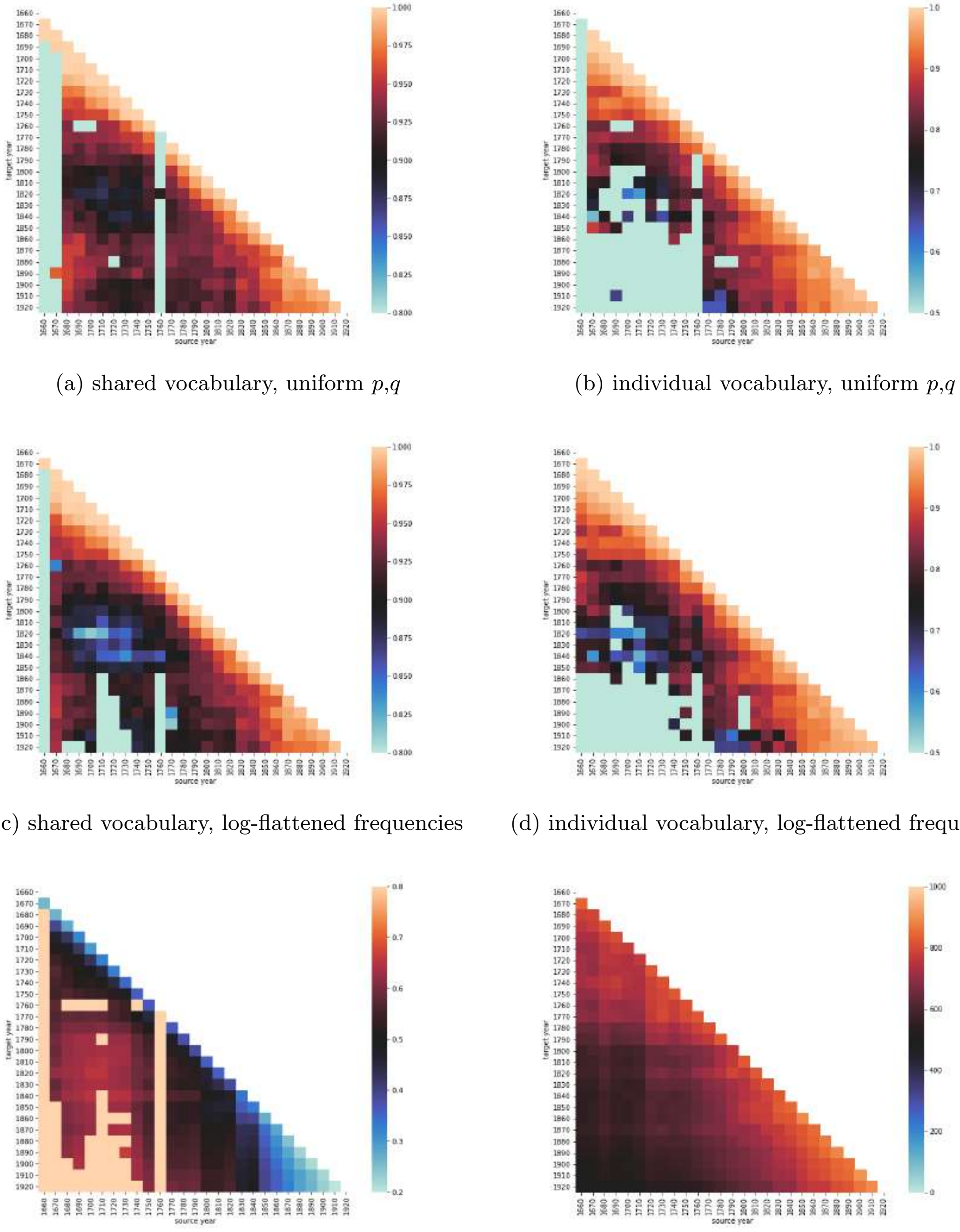
To conclude, these experiments on sub-spaces with identical vocabulary inform about the conceptual divergence of pairs of spaces, which can be used to select a balanced range of pairs of spaces for further investigation.

Lexical Divergence. As much as the optimization of couplings on shared words looks promising and can inform about the overall potential to relate two spaces, the reality is different: the distribution of words varies from decade to decade, and so does the set of the more frequent words. It is thus very unlikely that the vocabularies of any two sub-spaces are the same. For this reason, I repeated the experiments from above, but without the requirement that the most frequent words be mutual. At a size of 1K, the two sub-vocabularies are large enough to not only contain function words and small enough to only capture words in a relevant range of frequencies. The estimator of quality in this experiment is the number of string-matching pairs relative to the vocabulary overlap. Figures 6b, d, and f show the results and the corresponding vocabulary overlap.

As before, it is apparent that while couplings of the given size for temporally close spaces still work well, spaces which are farther apart in time are harder to couple. This is expressed by tiles in light-blue, indicating that less than 50% of the shared vocabulary can be coupled in the form of string-matching translation pairs. Specifically, almost all couplings of spaces for decades before 1770 with spaces of decades after 1850 show a systematically low performance.

One could argue that this is due to a decrease in vocabulary overlap, as each point in either sub-space that does not correspond to a mutual word acts as a distractor during the optimization of the coupling. However, the vocabulary overlap decreases smoothly in correlation to two spaces' distance in time. This smooth development cannot account for the sudden drop in performance for couplings of spaces from before the 1770s to after the 1850s. Also, differently to the couplings with shared vocabulary, the earliest spaces in this setup are not coupled to the recent spaces more reliably. The hypothesis of the 'legacy of initialization' cannot explain the light-blue zone in the bottom left of Figures 6b and d.

A possible external factor is the introduction of a reviewing process for the Proceedings of the Royal Society in the 1750s, bringing about "a major period of change around the 1750s [...] related to conventionalized style of writing" (Degaetano-Ortlieb and Teich, 2018, p.29). The reviewing process requires authors to communicate findings clearly and to refer to concepts without causing ambiguities; it sets standards for scientific communications. To meet these standards, authors started using the same syntactic constructions in similar situations more and more. To avoid lexical ambiguities, the scientific vocabulary, already specialized, was further developed to give each term a unique meaning. These changes introduce differences between the embedding spaces before and after the 1750s.



(a) shared vocabulary, uniform p,q

(b) individual vocabulary, uniform p,q

(c) shared vocabulary, log-flattened frequencies

(d) individual vocabulary, log-flattened frequencies

(e) GW-distance (shared vocabulary, uniform p,q)

(f) vocabulary overlap

Figure 6: Semantic and lexical divergence in terms of string-matching translation pairs obtained from couplings of RSC decade-wise embedding space pairs, optimizing for the 1K most frequent words (**a**, **b**, **c**, **d**). *Left*: coupling on *shared* words; score relative to the coupling size; range for color-coding: 0.8–1.0. *Right*: no shared vocabulary assumed; score relative to vocabulary overlap; range for color-coding: 0.5–1.0. *Top row*: couplings initialized with uniform distributions. *Middle row*: couplings initialized with log-flattened frequency information (cf. Section 4.3.4). **(e)**: GW distance of couplings optimized as in (a) (×10^2, range for color-coding: 0.2-0.8). **(f)**: vocabulary overlap for (c) and (d).

Lastly, the experiments uncover that certain spaces and space pairs are consistently difficult to couple, often leading to extremely poor results. This is the case for spaces of the 1660s and 1760s as well as for several pairs involving a space from the 1710s, 1720s, or 1730s. This low performance could be due to "bad luck" in the optimization process, which is non-convex and therefore does not guarantee to converge to the optimum or at all. This could be further investigated with more repetitions of the experiment.[25] However, the chances of no convergence to the optimum happening as selectively as depicted in Figure 6a can be deemed very small. A suitable explanation for these observations has still to be found.

To conclude, the vocabulary overlap of two sub-spaces influences the quality of their coupling. For pairs of decade-wise sub-corpora from the RSC, the overlap of the 1K most frequent words in each sub-corpus ranges from 0.48% to 0.91%. As more reliable models allow for better couplings irrespective of their conceptual divergences, couplings of space pairs of the decades 1860 or later perform well. This is underlined by the major coupling difficulties for spaces from before the 1770s to after the 1850s, which are likely to arise as a combination of differences in training data and increases conceptual divergence.

How will larger and smaller couplings perform based on changes in vocabulary overlap? For larger couplings, it is likely that the relative vocabulary overlap will decrease slightly once the coupling size exceeds the number of words which occur moderately frequently. At the same time, the absolute number of correct translation pairs obtained from a reliable large coupling will increase. For very small couplings, it is likely that the relative overlap increases, because the distribution of the utmost frequent words does not change a lot. If this is the case, smaller couplings will yield fewer string-mismatching translation pairs, but also fewer 'non-trivial', interesting pairs.

String Matches and Translation Scores. Following the intuitions from Section 4.3.1, Figure 7 shows a more detailed view of sets T of translation pairs, obtained from 2K-sized couplings of a range of space pairs. In general, string-matching pairs tend to have high translation scores, while mismatching pairs are assigned lower scores. Figures 7b, c, j, and l depict extremely unreliable couplings judging by their T; the reasons for this difference in performance are not completely clear. Note, however, that the temporal distance is is not the deciding factor *per se*: the spaces of the T shown in Figures 7a, d, and k are the same number of decades apart as b, c, and j, respectively, (0, 2 and 9 decades, respectively), but these sets contain mostly string-matching translation pairs.

There is a development observable for Figures 7f, g, h, and i, which form a 'chain' of subsequent space pairs, each 2 decades apart: the number of mismatching translation pairs decreases and the number of matching pairs as well as their median score increases. These developments lead to the assumption that it becomes easier over the course of this period to couple two spaces. This may indicate that language change slows down over time between the 1770s and the 1890s. However, on the basis of Figure 7 alone, these are just speculations; further work is needed to clarify such developments and their

[25] Interestingly, couplings of pairs with the 1660s space do reach reasonable performance if frequency information is provided for optimization (cf. 6d).

driving forces.

Lastly, it remains to mention for further reference Figures 27 and 28 in the Appendix which show the same analysis for couplings of the same space pairs, with the difference that those couplings were provided frequency information for optimization.

4.3.4 Effects of Frequency Information on Optimization

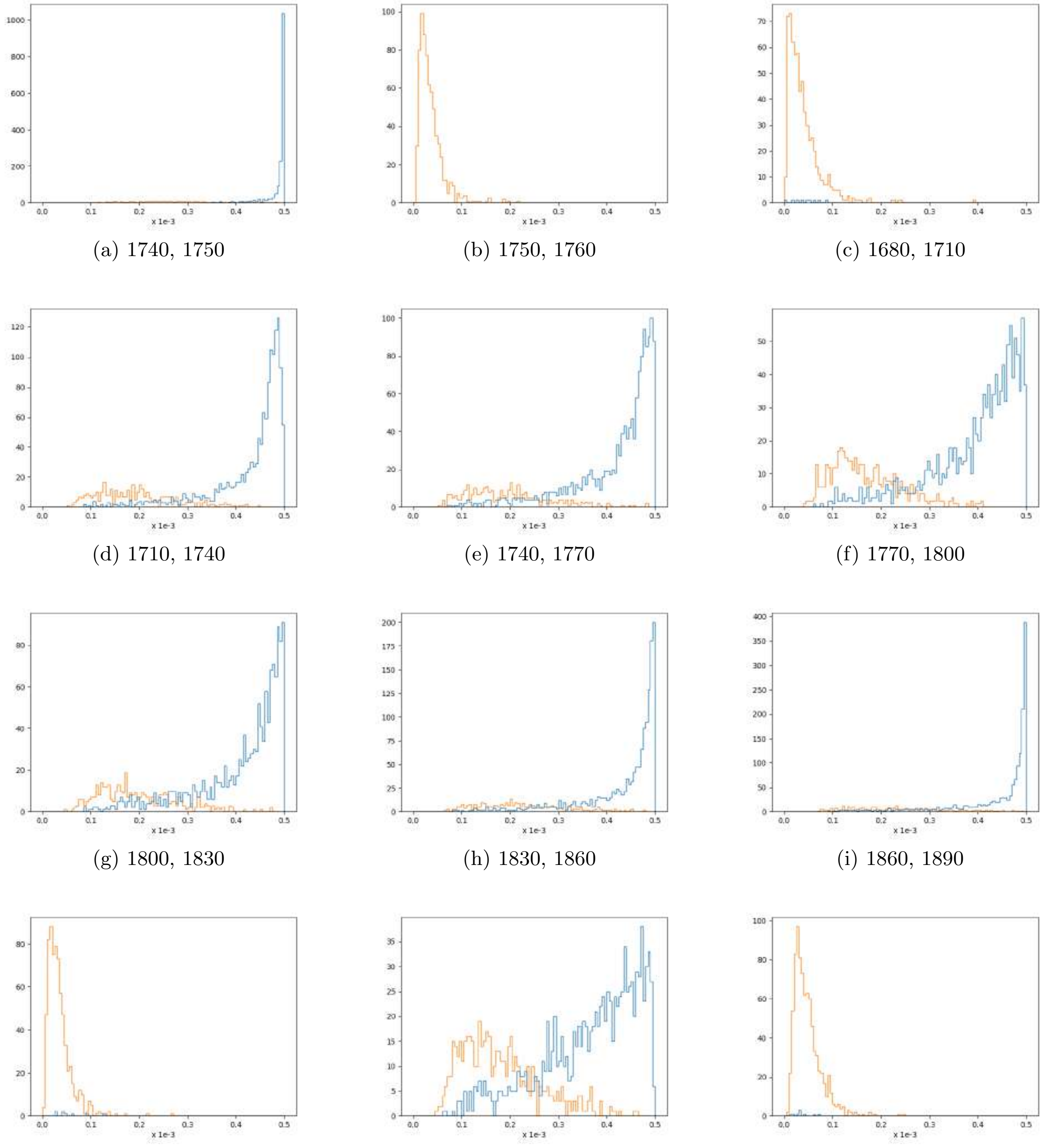
Alvarez-Melis and Jaakkola (2018) mention that frequency information is helpful for solving the optimal transport problem. That is, the information about relative token frequencies in the corpora underlying two spaces can be passed to the optimization algorithm in the form of two probability distributions p and q. The coupling is then initialized as the outer product of these two distributions rather than as a uniform matrix.

Foreshadowing the results of the following investigations, Section 4.3.3 presented couplings optimized both with and without frequency information and discloses the overall tendency that additional frequency information does not improve couplings *per se*. Here, I present two variants of providing frequency information: RAW and FLATTENED. Given two sub-spaces X^k and Y^k, the probability distributions p and q are estimated from the frequency distributions f_{X^k} and f_{Y^k}, respectively. The RAW variant of f maps words to their token counts in the space's sub-corpus; the FLATTENED variant of f maps to the logarithm of these counts. Figure 8 shows the number of string-matching and mismatching translation pairs obtained from couplings of selected pairs of $k = 2\text{K}$ sub-spaces, optimized with raw, flattened or uniform p and q.[26]

There are several observations to be made. First, the scores for couplings optimized with both variants of frequency information reach scores of up to 10^{-3}, while those of uniformly initialized couplings are at most 0.5×10^{-3} (i.e., $\frac{1}{k}$). The abrupt cut of scores for the latter is likely due to the fact that the initial couplings are uniform. These uniform distributions greatly underestimate the higher translation probabilities of the more frequent words. Apparently, the optimization process does not tend to assign a lot of additional probability mass to the very frequent words and therefore does not react to this initial underestimation as much as to the overestimation for less frequent words.

Second, couplings of the RAW variant have a considerably low performance with respect to both the number of string-matching translation pairs as well as the correlation between translation score and error rate. Even for couplings that lead to a reasonable T, the translation scores do not allow for a clear partition into string-matching and mismatching pairs. Compared to that, couplings of the FLATTENED variant perform reasonably well. What could explain these observations? One likely reason is the highly uneven distribution of words in natural language ('Zipfian' distribution, cf. (Zipf, 1932)). As the frequency information is used to initialize the coupling, any one translation pair's score correlates with the respective token frequencies of the pair's words. It seems likely that these distributions are skewed too much in order to be useful for the optimization algorithm. When the differences between frequencies are reduced by applying a simple logarithmic function, the distributions are better suited to initialize couplings, providing

[26] Figures 27 and 28 in the Appendix show further examples.



(a) 1740, 1750

(b) 1750, 1760

(c) 1680, 1710

(d) 1710, 1740

(e) 1740, 1770

(f) 1770, 1800

(g) 1800, 1830

(h) 1830, 1860

(i) 1860, 1890

(j) 1700, 1800

(k) 1800, 1900

(l) 1700, 1900

Figure 7: Distributions of translation pairs by coupling score (range: $[0, 0.5 \times 10^{-3}]$). Couplings optimized on sub-spaces with k=2000, initialized with uniform p/q. Matches in blue, mismatches in orange.

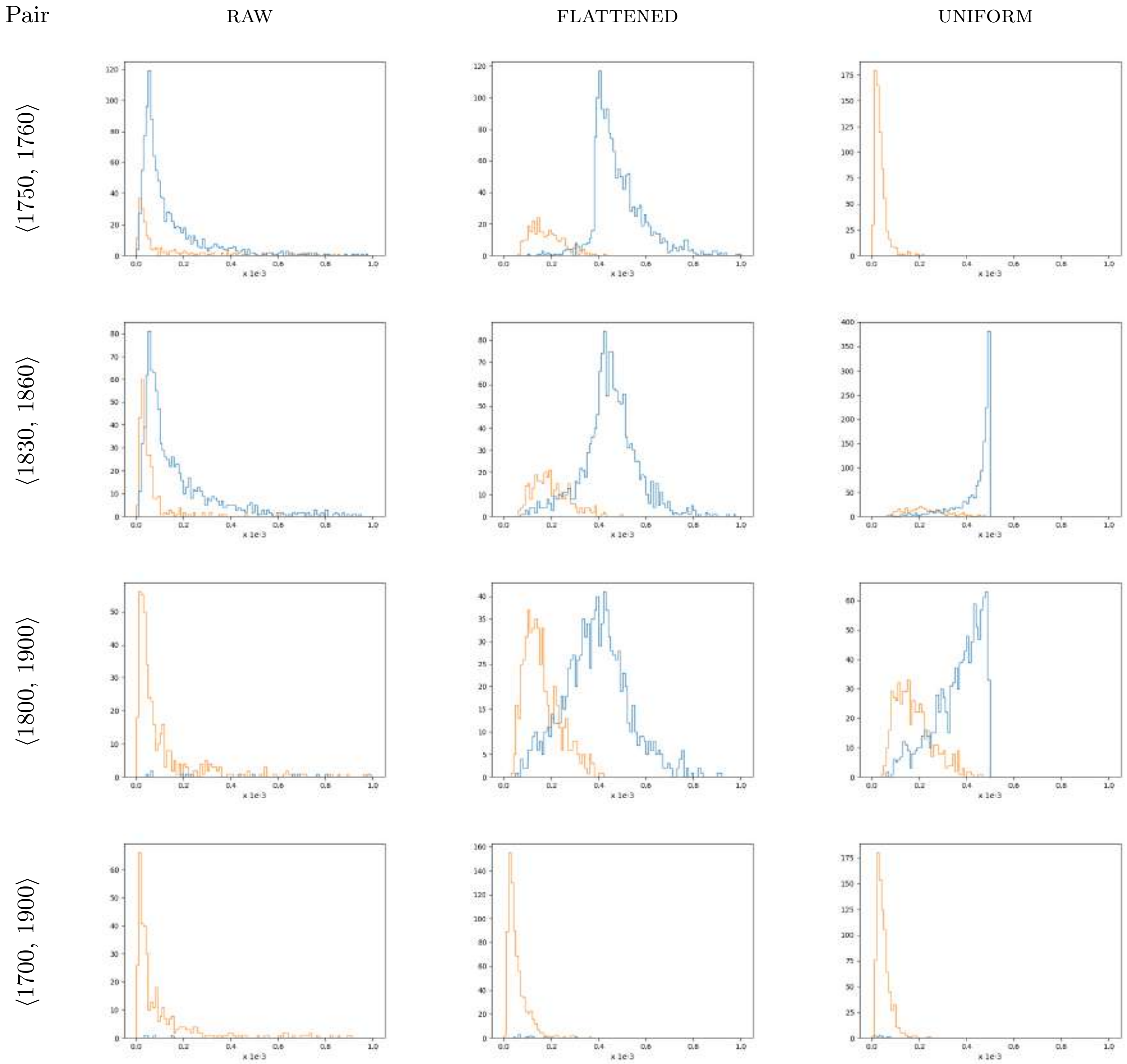


Figure 8: Distributions of translation pairs by coupling score (range: $[0, 10^{-3}]$). Couplings optimized on sub-spaces with $k = 2000$, initialized with varying distributions p and q. Matches in blue, mismatches in orange. Scales of y-axes vary.

FLATTENED					UNIFORM				
Top 10	Rank	1830s word	1860s word	Score*	Top 10	Rank	1830s word	1860s word	Score*
the	446	founded	based	0.417	them	211	founded	based	0.461
of	478	detached	isolated	0.405	which	220	detached	isolated	0.459
to	584	ounces	c.c.	0.377	such	393	tend	tends	0.426
a	596	ascertained	determined	0.374	to	423	crystallized	crystalline	0.419
in	598	happens	occurs	0.374	been	446	besides	including	0.415
and	613	supposition	assumption	0.369	be	478	happens	occurs	0.406
by	615	platina	platinum	0.368	place	483	ingredients	constituents	0.405
that	627	foregoing	previous	0.363	remaining	557	lately	recently	0.389
as	636	grains	grammes	0.361	form	559	signs	sign	0.389
be	640	besides	including	0.36	by	571	supposition	assumption	0.386

Table 5: Overall highest-scoring translation pairs (all string-matching) and highest-scoring translation pairs with string mismatches. Pairs obtained from couplings (k = 2K) with varying initialization. $^{*}\times 10^{-3}$

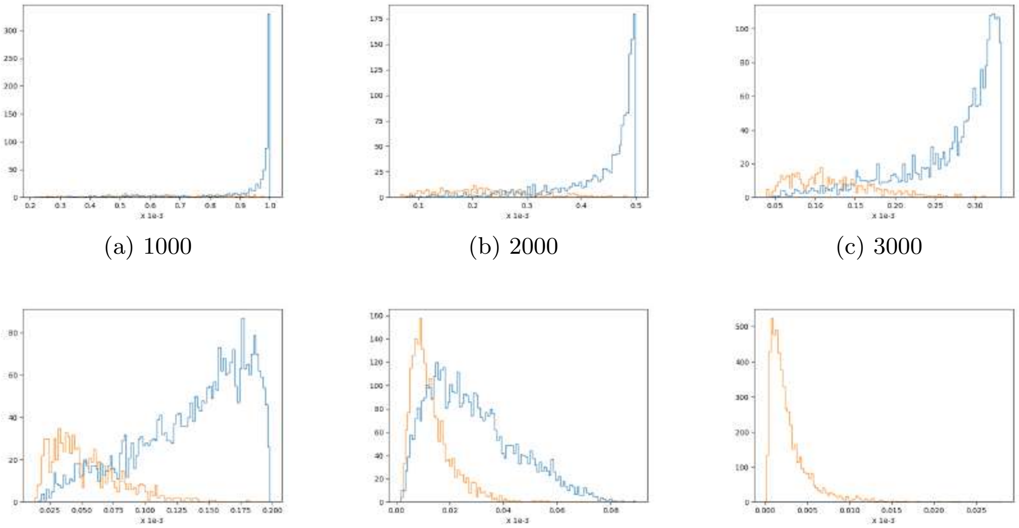
less extreme initial estimations for translation likelihood.

Third, couplings of the FLATTENED variant, while leading to comparable results, are in general slightly outperformed by uniformly initialized couplings in terms of the number of retrievable string-matching translation pairs. However, the FLATTENED-couplings tend to be more consistent in producing reasonable results (consider also the differences between Figures 6b and d, with 6d showing fewer low-performance couplings). This indicates that frequency information, if used with caution, can indeed assist in optimizing a coupling, especially in cases in which the spatial constellations of points alone would not suffice.

Lastly, Table 5 presents a comparison of the highest-scoring string-matching as well as mismatching translation pairs obtained from the FLATTENED variant and the uniformly initialized coupling. For both couplings, most of the high-scoring mismatches occur between semantically similar or related words. With respect to matching translation pairs, the FLATTENED variant behaves differently to the uniform variant; here, the highest-scoring words are not just mainly, but exclusively function words. As shown by Wendlandt et al. (2018), a high corpus frequency of a word does not necessarily lead to high embedding stability (i.e. a reliable vector representation of one and the same concept). Uniformly initialized couplings operate solely on the similarity of spatial constellations between points.[27] For a uniformly initialized coupling, the highest-scoring translation pairs can thus be considered to be the most stable concepts across spaces, regardless of their corpus frequency. Conversely, it is possible that the highest-scoring pairs from frequency-informed couplings do not belong to the most stable embeddings.

To conclude, information on token frequencies in the spaces' underlying corpora can help to optimize couplings with GWOT if the embeddings alone do not allow for good optimization. This is subject to the condition that the naturally extreme skewedness of the frequency distribution is first *flattened*, for example by applying a logarithmic

[27] Uniformly initialized couplings are still influenced by frequency information, because the points in the two sub-spaces to be coupled are selected on the basis of corpus frequency. This, however, does not affect which score is assigned to one word pair or another; it only affects which word pairs are considered by GWOT in the first place.



(a) 1000

(b) 2000

(c) 3000

(d) 5000

(e) 10000

(f) 20000

Figure 9: Distribution of string-matching mutual translation pairs by translation likelihood $\times 10^{-3}$. Optimal coupling computed for the k most frequent words in $\langle X^k_{1830}, Y^k_{1860}\rangle$ (not necessarily mutual), optimized without frequency information. Matches in blue, mismatches in orange. X-axes scaled individually.

function to the raw frequencies. While the couplings optimized in this way tend to hold fewer mutual translation pairs, these pairs have a wider range of scores.

4.3.5 Effects of Coupling Size on Mutual Translation Pairs

Section 4.3.2 discussed the consequences of coupling size for the time required for optimization. The following showcases the influence of the size parameter on the contents of couplings.

To this end, I optimized couplings of various sizes k of $\langle 1830, 1860\rangle$ which has previously shown to allow a good coupling at k = 1K and k = 2K (cf. Figures 6b and 7h, respectively). This also leads to the assumption that the couplings do not require frequency information for successful optimization. They were thus initialized with uniform p/q. Table 9 shows the distributions of translation pairs obtained for the various couplings.

There are two tendencies observable with increasing coupling size. Firstly, the distribution of coupling scores for mutual translation pairs shifts from a long left-tailed distribution with mostly high-scoring pairs to a right-tailed distribution with a large number of low-scoring pairs. One possible explanation is that larger couplings have to relate embeddings of words which are less frequent in the underlying corpus.[28] As the quality of a model depends, among other factors, on the amount of available data (cf. Wendlandt et al., 2018), two embeddings x and y of less frequent words tend to be less stable across spaces, even if the x and y are supposed to express the same concept. In

[28] As before, this is because sub-spaces are compiled by picking the vectors based on token frequency.

these cases, GWOT does couple x and y, but assigns a relatively low score because the spatial constellations of x and y do not match as neatly as would be the case for more stable embeddings.

Secondly, the ratio of matches to mismatches decreases. As larger couplings have to relate more points to one another, there is a higher chance that two unrelated points have deceptively similar spatial constellations. This makes it harder to achieve a reliable coupling. The number of string-matching mutual translation pairs is still large for $k =$ 5K and $k =$ 10K, but drops to 0 for $k =$ 20K. This last result is unexpected and does not fall in line with the other couplings of $\langle 1830, 1860 \rangle$ or with other couplings of size 20K. The reasons for this drop are unclear and as with the other cases in Section 4.3.3, further investigations are required to shed light on these anomalies.

The unexpected result for $k =$ 20K stresses that GWOT behaves differently with each space pair and the predictions are not to be trusted blindly. Still, there are general tendencies observable for the alignment of spaces with GWOT.

4.3.6 Dynamics for Small Couplings

As presented in Section 4.2, a space pair $\langle X, Y \rangle$ is aligned by solving Procrustes on a small bilingual signal, namely the most stable concepts from both spaces. There are two options to obtain this bilingual signal. One could simply re-use the list of translation pairs from the large coupling $\Gamma_{X,Y}^{k>15000}$ and take the highest-scoring such pairs. A cleaner option, however, is to optimize a new coupling for small sub-spaces $\langle X^k, Y^k \rangle$ which comprise only the most stable concepts from both spaces. This is computationally easy to achieve for $k \leqslant 500$ and minimizes the number of points that might act as distractors during optimization.

As previously shown, the coupling size and the availability of frequency information during optimization influence a coupling's quality. These factors are investigated in the following for ⟨1830, 1860⟩ for various $k \leqslant 500$. Figure 10 shows the numbers of mutual translation pairs obtained from these couplings as well as the numbers of string-matching mutual translation pairs.

The main factor of comparison is the availability of frequency information ('flattened' in Figure 10). Frequency-informed couplings of a size up to $k = 250$ slightly outperform uniformly initialized couplings with respect to the overall number of mutual translation pairs, and they clearly yield more matching mutual translation pairs than the uniform variant. If the latter is taken as an indicator of coupling quality, this suggests that couplings of size $k \leqslant 250$ should be initialized with frequency information. The second factor, size, plays a minor role to the quality of a coupling at this magnitude. Note, however, that frequency-informed couplings tend to yield fewer mutual translation pairs than the uniform variant for $k \geqslant 300$. The size for couplings Γ_P used to create a projection P is therefore set to 300. The bilingual signal obtained from such a coupling can then be further reduced, for example by solving Procrustes on the embeddings of only the, say, 100 highest-scoring mutual translation pairs.

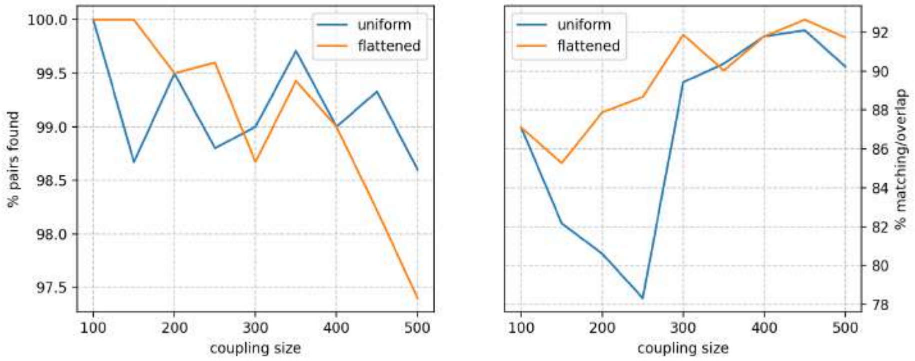


Figure 10: Performance of couplings of ⟨1830, 1860⟩ as a function of coupling size, both with and without frequency information. *Left:* number of mutual translation pairs in proportion to coupling size. *Right:* number of string-matching pairs relative to vocabulary overlap.

4.3.7 Evaluation of the String-Match Estimator

The previous sections operationalized the number of string-matching and of string-mismatching translation pairs as a rough estimator for the quality of a coupling between two word embedding spaces. This *string-match estimator* is used because it provides a more intuitive interpretation of how similar two spaces are to each other than the GW-distance. It can be used to a certain extent because the diachronic scenario at hand is monolingual and the vocabularies of a pair of sub-spaces have a large overlap. However, as mentioned previously, the unsupervised nature of GWOT can be used to view diachronic investigations as a bilingual task. While under the assumption of monolinguality, a string-mismatching translation pair can be treated as an erroneously aligned item, this is not necessarily the case from the bilingual point of view. For example, one and the same concept might be referred to with varying expressions (i.e., show lexical change) or simply be spelled differently, perhaps due to orthographic reforms. As Bizzoni et al. (2019a) point out, in the RSC there is a tendency for "lexical specialization" over time (e.g., ⟨*detached*, *isolated*⟩, cf. Table 5). These cases of string-mismatching translation pairs should not be seen as false alignments *per se*, as their two corresponding vectors may very well be the most similar to each other in terms of their intra-spatial roles.

Annotation Experiment. In order to gain empirical insights about the reliability of the string-match estimator, a small set of string-mismatching translation pairs was annotated and evaluated.[29] The source for these pairs is a large coupling of ⟨1740, 1770⟩, uniformly initialized and optimized over sub-spaces of size $k = 20000$ with the default $\lambda = 5 \times 10^{-4}$.[30] $\Gamma^{20000}_{1740,1770}$ yields about 14000 mutual translation pairs, about

[29] Annotation schema and labeled pairs are openly available at https://github.com/SimonPreissner/get-shifty/tree/master/annotation.

[30] The annotation task was carried out *after* the 12 space pairs, selected in Chapter 4.4.1, had been optimized (cf. Chapter 5.1 for results). I chose this coupling because it appears to be in between the

Label	Name	Treated as	Proportion	Examples
O	orthographic/morphological	match	1.08%	toward–towards, ankles–ancles, fallen–falls
S	semantically similar	match	1.76%	producing–forming, article–essay, pond–pool
R	semantically related	noise	13.36%	artery–vein, confirmed–convinced, fruit-seed
A	antonymy	noise	1.72%	inferior–superior, floor–roof, wet–dry
N	noise	noise	81.28%	w–58, minus–rectangle, ancients–l'abbe, 21/–03
X	wildcard	noise	0.68%	laurin–rutty, quent–cerning, glewed–pluck

Table 6: Labels for the annotation of string-mismatching translation pairs, with relative frequency in the annotated data.

5800 or 41% of which are string-mismatches. Annotation was carried out on the 2500 highest-scoring of these mismatches. The decision to only annotate the highest-scoring pairs and to treat the other mismatching pairs as false positives is motivated by the time and resources available as well as by earlier observations that the coupling score of a pair is negatively correlated with the chances that it is noise (i.e., that there is no reasonable linguistic connection between the two words). The 2500 data points were randomly split into two sets of 1500 samples each, with an overlap of 500 used to estimate inter-annotator agreement via Cohen's κ (Cohen, 1960).

The annotation task was carried out by two annotators (fluent speakers of English), each one annotating one of the two sets. For each translation pair $\langle u, v \rangle$, the task was to assign one of six labels which most closely describes the relationship between u and v. Table 6 gives an overview of the classes and corresponding examples. The annotation scheme was designed to be as simple and straight-forward as possible while leading to relevant insights. It thus includes the wildcard label X, which can be assigned to cases in which the annotators are uncertain.

One critical point is the distinction between those pairs which are semantically *similar* (in the sense of near-synonymy) and those which are semantically *related* (having any semantic relationship, cf. Budanitsky and Hirst, 2001). While semantically similar words can be used interchangeably in a sentence without fundamentally changing its meaning, semantically related words do not necessarily refer to similar concepts (e.g., there is an important difference between "artery" and "vein" as shown by the sentence "the medicine was injected into the *artery/vein*").[31] For evaluation of the coupling's quality, the labels are interpreted either as denoting correct pairs (semantically similar pairs (label S) and orthographically or morphologically different pairs, label O) or incorrect pairs (all other classes of mismatches, including semantically related pairs).

Results. The annotations reach an inter-annotator agreement (IAA) of κ=0.45, which is usually interpreted as "moderate agreement" (Landis and Koch, 1977). Additionally to this notoriously conservative measure for IAA, the frequency of the wildcard label X sheds light onto the difficulty of the task. As the data contains many domain-specific

surprisingly good large couplings such as $\Gamma^{20000}_{1740,1750}$ and the very unreliable ones (e.g., $\Gamma^{20000}_{1710,1740}$) in terms of string-matching and mismatching translation pairs (cf. Figure 12 in Chapter 5.1).

[31] This description of semantic similarity stands in contrast with other definitions. For example, Resnik (1995) defines semantic similarity as the relation holding between co-hyponyms in a semantic hierarchy; this would include the pair *artery/vein*. Semantical relatedness).

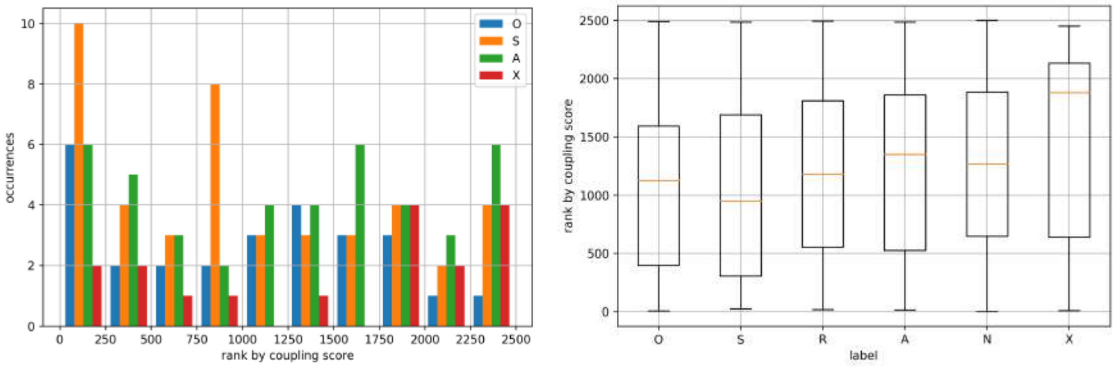


Figure 11: Distribution of labels. Bins were calculated over ranked lists of coupling scores instead of scores directly in order to factor out the number of pairs per score interval. *Left:* distribution of infrequent labels (R and N are evenly distributed). *Right:* box-and-whiskers plot of all labels (the horizontal symmetry of R and N indicates their even distribution).

terms and the annotators are not trained, X was included in the annotation scheme in order to prevent guesswork in difficult cases. X was used only for 0.7% of the pairs, which means that either the task was not very hard or that the annotators were shy to use this label and still guessed in difficult situations.

Table 6 reports on the relative frequencies of labels. The two most prevalent labels among the annotated data are, as expected, N and R. They make up about 95% of annotated pairs (81.3% for N and 13.4% for R). The two classes of mismatches deemed to be false negatives (O and S) are observed in only 2.8% of the annotated data.

What does this mean for the string-match estimator? Expressing that quality by the number of string-matching pairs relative to the coupling's vocabulary overlap, the estimated quality of $\Gamma^{20000}_{1740,1770}$ is 0.722. This improves only marginally to 0.728 when adding the 71 S- and O-annotated pairs to the matching ones. However, this conservative evaluation assumes that all mismatching pairs with a score lower than these 2500 annotated pairs are to be interpreted as noise. As Figure 11 shows, this assumption does not hold for $\Gamma^{20000}_{1740,1770}$: all labels are distributed more or less evenly across the whole set of annotated pairs. Only S- and X-pairs have a slight tendency to hank higher and lower, respectively. With the less conservative assumption that the labels are distributed evenly across all mismatching pairs, $\Gamma^{20000}_{1740,1770}$ achieves an estimated quality of 0.736, which is not very different from the most conservative score (0.722). These results show that in this quasi-monolingual scenario, the string match/mismatch analysis is a reliable estimator for the quality of a coupling.

The results also show that for this pair of spaces, lexical and/or orthographic changes are unlikely, which raises the question of whether it is useful to frame diachronic studies as bilingual. It is possible that other pairs of spaces which are farther apart in time or experience greater lexical and orthographic change still benefit from the bilingual approach. In Chapter 5, I will experiment from the monolingual as well as the bilingual point of view and compare the two variants. This will allow a final verdict on the question of whether the assumption of bilinguality is needed.

Further Use. The annotation task fulfills a second purpose: pairs which are labeled as either *O, S, R,* or *A* are potentially interesting for further investigations of semantic shifts, especially when they show a relatively high score in the coupling. Specifically, the pairs with the labels *O* and *S* form a set of 71 word pairs that can be used for further investigations on ⟨1740, 1770⟩. For example, it could be investigated whether *sorts* and *kinds* are close to each other when the 1740s space is projected onto the 1770s, and also to see where *kinds* 'comes from' and where *sorts* 'goes to'.

A final remark to the string-match estimator: it is clear that any set T of translation pairs obtained from GWOT will contain some noise. As mutual translation pairs are obtained directly from the scores of a coupling, there might be ways to transform these scores in a sensible way in order to increase the correlation between a pair's score and the probability that it is a true translation pair. I tested some methods, none of which yielded satisfactory results (cf. Figure 29 in the Appendix for more details); future work might include further investigations in this direction.

4.4 Conclusions and Considerations for Shift Detection

The following summarizes the insights gained from the previous sections and applies them to (a) determine a selection of space pairs on which to focus in the subsequent experiments and (b) determine the circumstances under which to optimize larger and smaller couplings. The answers to the main questions posed at the beginning of this chapter are as follows:

1. It is feasible to frame the detection of diachronic conceptual shifts as a bilingual task. However, annotation of the bilingual signal for one pair of spaces revealed that orthographic and lexical changes are relatively infrequent, so that the benefits of a bilingual approach can be deemed rather small.

2. Yes, diachronic embeddings can be aligned with GWOT *to a certain extent.* As an investigative method, it can uncover differences between embedding spaces. As a tool for alignment, its success is dependent on the particular spaces to be aligned (see Chapter 5.1).

3. Diachronic shifts are preserved by aligning the spaces only along the most stable concepts. Other, potentially shifted points are associated either via a dictionary of 'diachronic translations' or by assuming that a word shared by the vocabularies of two spaces corresponds to the same concept at two different points in time.

4. The quality of a coupling is most strongly influenced by its size and the availability of frequency information. Additionally, differences in the reliability of embeddings (correlated with corpus size) and the temporal distance of space pairs become apparent when performing GWOT.

4.4.1 Selection of Space Pairs

I will not perform full-fledged experimentation with all of the available spaces for three reasons. First, a pairwise comparison of all 27 word embedding spaces from the RSC would by far exceed the scope of this work. Second, not all space pairs can be aligned equally well with GWOT. It would be nonsensical to base experiments on unreliable large couplings. Third, the experiments are designed to go in-depth about conceptual shifts rather than describe the general diachronic tendencies present in the RSC. In order to focus on qualitative analysis of the results, I reduce the number of investigated space pairs to 2 in two steps.

In the first step, I select a variety of 12 pairs which differ in time span, epoch, and size of sub-corpus. The selection is presented in Table 7. In order to allow for experiments on a wide range of diachronic dynamics, the selection comprises space pairs of different chronic distances (1, 3, 10, and 20 decades apart) as well as a sequence of seven equally distant pairs, spanning from 1680 to 1890 in steps of 3 decades. This interval was chosen because the experiments on divergence in Section 4.3.3 suggest that space pairs at this interval already diverge enough to show shifts and are still similar enough to lead to reliable couplings. Another factor that might influence the feasibility of experiments is the amount of data available to each space. The pairs in the selection have various combinations of sub-corpus sizes, accounting for the earlier spaces with smaller as well as the later spaces with larger sub-corpora. Finally, the selection includes pairs such as $\langle 1700{,}1900\rangle$ and $\langle 1750{,}1760\rangle$ – spaces or space pairs which have shown to be difficult to align.

In the second step, described in Chapter 5.1, I optimize large couplings for each of these pairs and select two pairs which promise realistic results. Narrowing down the selection in two steps gives a more complete picture of how well large couplings of RSC spaces can be optimized. For the second step, it also allows to pick two space pairs which are both promising for the experiments and representative of the capabilities of GWOT.

4.4.2 Translation Pairs (Large Couplings)

As presented in Section 4.2, the set T of translation pairs is used to relate points across spaces to one another irrespective of whether the purposefully non-optimal projection $\mathbf{P}$ projects them to be close to one another or not. T is obtained directly from a relatively large coupling. The size of these couplings is limited by the algorithmic complexity of GWOT and with the available hardware, the largest couplings optimized in a reasonable time are of size 20000. The number of good translation pairs in T will be smaller than this, partly because the vocabularies of such sub-spaces do not overlap completely and partly because the coupling itself will lead to false translation pairs. At k=20000, however, T is expected to still contain a substantial number of translation pairs that show semantic shifts and other diachronic dynamics. For optimization, these couplings should be initialized with uniform probability distributions. The entropy regularization term is fixed to $\lambda = 5 \times 10^{-4}$, which is the smallest value that does not lead to numerical

Source, Target	Decades	Tokens $\times 10^6$		% Overlap*	D_{GW} $\times 10^{-2}$**
1740, 1750	1	1.01	1.18	64.88	0.36
1750, 1760	1	1.18	0.97	61.66	0.37
1680, 1710	3	0.57	0.49	73.73	0.48
1710, 1740	3	0.49	1.01	58.85	0.51
1740, 1770	3	1.01	1.50	56.79	0.50
1770, 1800	3	1.50	1.61	53.88	0.52
1800, 1830	3	1.61	2.61	58.63	0.47
1830, 1860	3	2.61	5.87	62.68	0.40
1860, 1890	3	5.87	10.20	62.11	0.32
1700, 1800	10	0.78	1.61	51.33	0.63
1800, 1900	10	1.61	10.33	47.98	0.53
1700, 1900	20	0.78	10.33	37.39	0.84

Table 7: Selection of space pairs for experiments on shift detection. *vocabulary overlap calculated from the 20K most frequent words from each space. **GW-distances of couplings obtained from 1K-sub-spaces (cf. Section 4.3.3).

errors.[32] The negative correlation between translation score and error rate (cf. Sections 4.3.1 and 4.3.7) backs the validity of higher-scoring translation pairs and indicates that lower-scoring ones should be treated with caution, irrespective of whether their strings match or not.

4.4.3 Projection Matrix (Small couplings)

The comparably tiny couplings that yield the bilingual signal for the computation of $\mathbf{P}$ via Procrustes behave differently to the large couplings. The question of size is a trade-off between expressiveness and reliability: on one hand, $\mathbf{P}$ should be as small and thus as non-optimal with respect to less stable words as possible in order for shifts to be observable more clearly. On the other hand, the set of anchor pairs needs to be large enough to be representative of the most stable concepts from each space, especially because even the most frequent words can be unstable at times (cf. Wendlandt et al., 2018).

In practice, the set of translation pairs for $\mathbf{P}$ suffers from the same limitations as T, namely an incomplete vocabulary overlap of the sub-spaces and a certain error rate of the coupling. Therefore it seems reasonable to optimize a coupling for slightly larger sub-spaces and then pick the highest-scoring translation pairs as anchors to compute $\mathbf{P}$. From the preceding experiments, a solid size for the bilingual signal is estimated to be $|T_P| = 100$. A coupling of sub-spaces of size $k = 300$ should yield enough translation pairs so that the highest-scoring 100 are correct. Note that the embeddings in the sub-spaces are selected on the basis of their tokens' frequencies in the sub-corpus. This can be seen as a first coarse-grained selection for alignment which is then followed by a more sensible selection by GWOT to the most stable points.

Small couplings will be initialized with (log-flattened) frequency information and, as with all couplings, with $\lambda = 5 \times 10^{-4}$ (cf. Section 4.3.6).

[32] This stands in contrast to Alvarez-Melis and Jaakkola (2018) who optimize with $\lambda = 10^{-4}$ and report numerical errors at the same coupling size only for smaller λ.

5 Detection of Conceptual Shifts

The previously gained insights about GWOT are now applied to a selection of embedding spaces to make them comparable to each other, optionally working under the assumption of bilinguality. This chapter treats the second major methodological part of the thesis: the detection of systematic conceptual shifts. To this end, I propose and experiment with a novel method which is based on the intuitions that concepts usually change together, and that only certain aspects of their semantics change at a time. The proposed method attempts to express these 'changes of components of meaning' in a natural way. The aim of this chapter is to answer the following questions from the introductory chapter:

1. Is it useful to frame the detection of diachronic conceptual change as a bilingual task?
2. Does the proposed method for unsupervised detection of shifts work; is it sensible and expressive?
3. Do the insights from the proposed method relate to previous findings?
4. Can the proposed method uncover new (general and/or specific) dynamics of language change within the RSC?

The chapter is structured as follows. Section 5.1 determines the pairs of embedding spaces on which to perform the experiments. The proposed method is presented in Section 5.2, where it is also embedded in the general experimental setup. These include two unsupervised experiments, presented in Section 5.3, and a supervised experiment in Section 5.4; both sections include quantitative and qualitative investigations. The chapter closes with answers to the questions above in Section 5.5.

5.1 Embeddings and Couplings for Subsequent Experiments

5.1.1 Optimization of Large Couplings

For each of the space pairs from the selection in Table 7, a large coupling is optimized with the parameters previously estimated to be optimal for this task.[33] As discussed previously, the main use of a large coupling is the compilation of a set T of translation pairs, which allows to carry out shift detection on the coupling's two embedding spaces while following the assumption of bilinguality. Table 8 shows the relevant statistics for all 12 large couplings; Figure 12 visualizes T for each of the couplings as a histogram of string-matching and mismatching pairs.

From Table 8 it becomes apparent that while certain couplings perform very well, GWOT does not successfully couple each space pair: only 5 out of the 12 couplings yield a T that can be deemed as at least moderately reliable in terms of the number of string-matching translation pairs; this is surprising. Given that any two embedding spaces from the RSC are probably more similar to each other than a pair of embedding spaces

[33] Parameters as set in Section 4.4.2: $k = 20000$; p/q uniform; $\lambda = 5e^{-4}$; distance matrices normalized with mean; maximum iterations: 300.

Space pair	Overlap	Pairs	Matches	Mismatches	Retrieved*
1740, 1750	12976	16407	11494	4913	88.58%
1750, 1760	12331	7510	24	7486	0.19%
1680, 1710	14746	12799	7466	5333	50.63%
1710, 1740	11769	6593	12	6581	0.10%
1740, 1770	11357	14031	8196	5835	72.17%
1770, 1800	10776	11009	4208	6801	39.05%
1800, 1830	11726	7107	265	6842	2.26%
1830, 1860	12535	5426	0	5426	0.00%
1860, 1890	12422	9179	3360	5819	27.05%
1700, 1800	10267	6828	39	6789	0.38%
1800, 1900	9596	5188	0	5188	0.00%
1700, 1900	7478	5075	0	5075	0.00%

Table 8: Statistics of the large couplings. *matches/overlap

in a true bilingual task[34] and given that Alvarez-Melis and Jaakkola (2018) achieved reliable couplings with this method in such a bilingual setup, the present results fall behind the expectations. The GW-distance between a pair of spaces is clearly not the decisive factor, as demonstrated by $\Gamma_{\langle 1750,1760\rangle}$. This is a coupling of *incrementally* trained embedding spaces covering two *adjacent* decades, but its T contains less than 0.2% of the possible matching translation pairs.

If it is not the data which causes these results, it might be the method of evaluation; in this case, the string-match estimator. However, the annotations of string-mismatching translation pairs in Chapter 4.3.7 demonstrate this estimator's validity. The numbers and the histograms in Figure 12 thus paint a relatively reliable picture. As almost all of the couplings' two vocabularies have a substantial overlap of 10000 words or more, 7 out of the optimized couplings do not manage to connect these common words to each other.

As neither the provided data nor the evaluation are likely causes for under-performing couplings, the issue lies withing GWOT itself. The couplings are optimized with the implementation from Alvarez-Melis and Jaakkola (2018) which was adapted only slightly to the twofold optimization approach. The most likely reason for the difference in results is the hyperparameter λ fr entropic regularization. Intuitively, larger values lead to a more 'fuzzy' coupling which tends to distribute probability mass more wirdespread than couplings with a smaller λ. The authors set λ between $5e^{-5}$ and $1e^{-4}$ and report that they "never had to go beyond these two values in all [their] experiments" (Alvarez-Melis and Jaakkola, 2018, §3.2). In the present case, however, I set $\lambda = 5e^{-4}$, because smaller values led to floating point errors. It is probable that some of the optimized couplings are too 'fuzzy' and distribute the translation probability mass over too many candidates, which in turn (1) makes it harder to find mutual best translations and (2) may cause more false translations. Table 9 displays these possible effects of an overly large λ in the form of highly probable but mismatching translation pairs.

[34] In fact, the logs from the optimization of 20K-couplings of individually trained spaces report GW-distances in the range of 0.0085 to 0.0095, which is about half of the smallest GW-distance of 0.017, reported in Alvarez-Melis and Jaakkola (2018) §4.4, achieved by a coupling between Spanish and French.

Space pair	Translation pairs
1740, 1750	(grin gram), (crystallizations crystallisations), (revives quenches), (wears stepping)
1750, 1760	(sea-breeze planman), (concamerations rain-gage), (austere undiluted)
1680, 1710	(cristalline rudus), (wrack calving), (ud ****), (portsmouth 1620), (eucl prax)
1710, 1740	(moivre historian), (-where funis), (**a plum), (-if austere), (***** sublime)
1740, 1770	(thunder-clouds reck), (elec tric), (half-complements 1000ths), (morbi lateribus)
1770, 1800	(810 1465), (threes tetrahedra), (back-horizon-glass ient), (rackwork carlings)
1800, 1830	(lepidolite opoponax), (acide cheltenham), (resin-like coarse-grained)
1830, 1860	(brome collin), (crassula eller), (1584 schonbein), (bipartite qg), (wrecked peligot)
1860, 1890	(peligot kellas), (friedel riedel), (grignon autun), (37’6 infrasternal), (sylvius lancisi)
1700, 1800	(bregmatis flucan), (sinistra lid), (javan tragacanth), (brachia iinch)
1800, 1900	(polarize litk), (respectability s.c.c.), (fiscal polymorph), (6i5 trachea), (=h jaggar)
1700, 1900	(jacobo hsemolysin), (pratense prolific), (mno correspondents), (re-action rectangles)

Table 9: Insights about the large couplings: highest-scoring translation pairs with string mismatch.

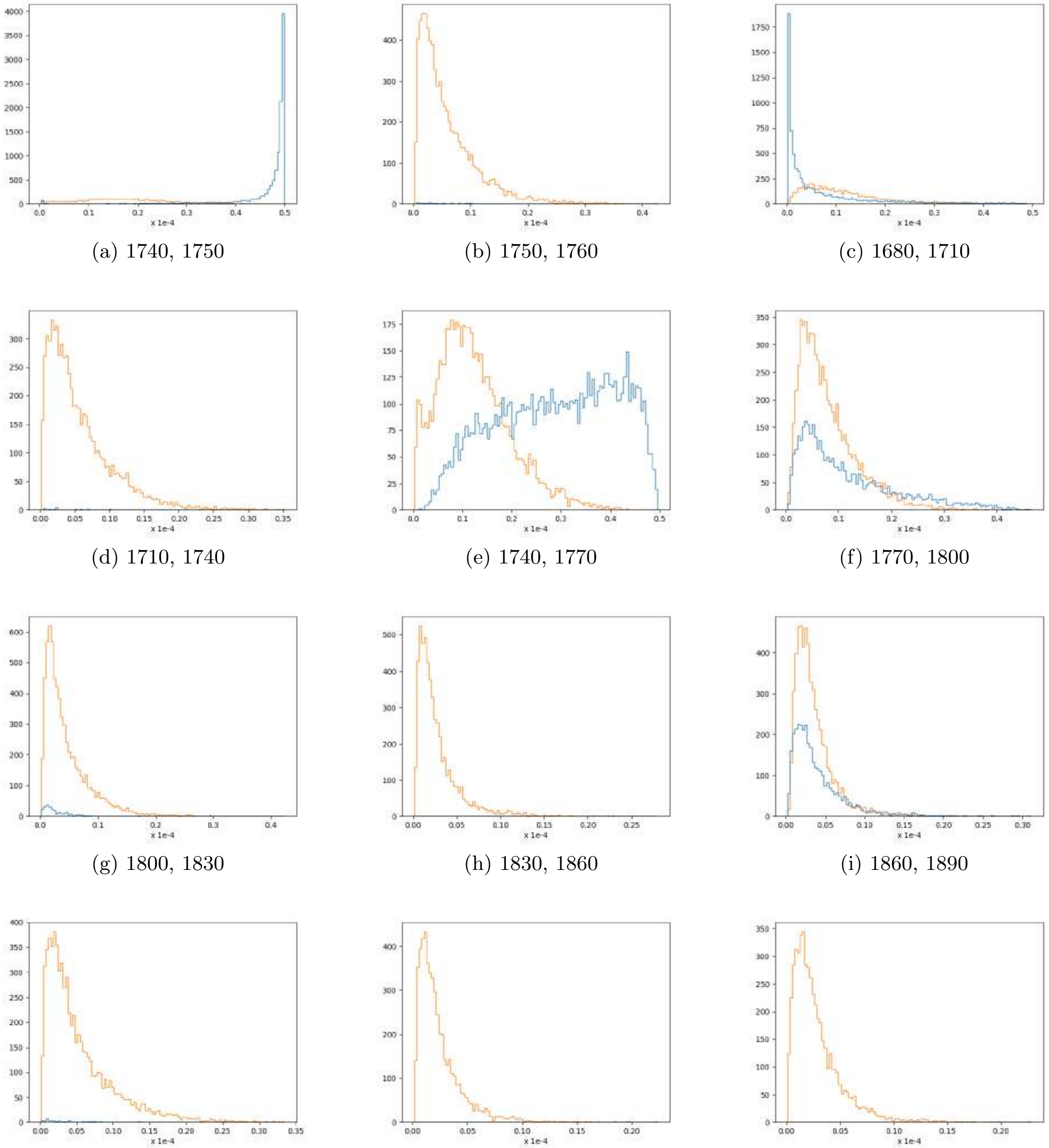
It is unclear why here the same values of λ in Alvarez-Melis and Jaakkola (2018) lead to floating point errors. It should be noted, however, that even with the relatively large $\lambda = 5e^{-4}$, several couplings manage to connect their two embedding spaces in a meaningful way. This is not only true for temporally adjacent spaces such as ⟨1740, 1750⟩, but also for pairs like ⟨1680, 1710⟩ or ⟨1740, 1770⟩, which are three decades apart. This leads to the conclusion that while λ does play an important role in optimization, it is not the sole reason for underperforming couplings. Further investigations on this issue are left for future work; for the next steps, i.e. the detection of systematic conceptual shifts, I focus on the couplings which do show reasonable and reliable performance.

5.1.2 Pick Two

In the second of the two steps to select pairings of embedding spaces (and their couplings), I choose ⟨1740,1770⟩ and ⟨1860,1890⟩ as the data for subsequent experiments. The primary reason to pick these pairs is that their couplings show moderate to good performance in terms of the makeup of their T, which is essential to carry out shift detection under the assumption of bilinguality (cf. Figure 12). However, this is not the only reason. As the spaces of both pairs are 3 decades apart, it is likely that there are more (and greater) conceptual shifts observable than in the temporally adjacent space pairs. Furthermore, these two pairs are 90 years apart from each other, promising to show different shifts and slightly different dynamics of language change.

The unsupervised bilingual scenario is not the only experiment to be conducted. Consequently, there are additional points individual to the two selected pairs which lead to the decision for ⟨1740,1770⟩ and ⟨1860,1890⟩ rather than for other space pairs like ⟨1680,1710⟩ with couplings which would be much better suited to construct a high-quality T.

As for ⟨1740,1770⟩, this space pair was already used for the validation of the string-match estimator (cf. Chapter 4.3.7). As a result, there is a set of 71 string-mismatching translation pairs annotated as either orthographically, morphologically, or semantically closely related. This labeled data can be used in the unsupervised bilingual case to



(a) 1740, 1750

(b) 1750, 1760

(c) 1680, 1710

(d) 1710, 1740

(e) 1740, 1770

(f) 1770, 1800

(g) 1800, 1830

(h) 1830, 1860

(i) 1860, 1890

(j) 1700, 1800

(k) 1800, 1900

(l) 1700, 1900

Figure 12: Distributions of mutual translation pairs by translation likelihood as obtained from the large couplings. Matches in blue, mismatches in orange; ranges of x- and y-axes set individually.

investigate the dynamics of orthographic and lexical changes.

The coupling of ⟨1860,1890⟩ is not as good as for other space pairs; however, the embeddings of these spaces are both trained on larger sub-corpora than spaces from earlier decades. As shown at the beginning of Chapter 4.3.3, these more recent spaces align better to each other than older spaces. In addition, the increased size of the sub-corpora allows to compare the shift detection across embedding methods.

5.1.3 Individually Trained Embeddings

All previous investigations have made use of the 'incremental' word embedding models, for which each decade's model is initialized with the model of the previous decade. The second, more common type of embedding method is to train a decade's embeddings solely on the corresponding sub-corpus. In order to compare these two methods, I optimized large couplings for both pairs of decades, using the *individual* instead of the incremental embedding models.[35] As the individually trained models do not use a shared vocabulary across decades, their individual vocabulariies may be smaller than 20K. This is the case for ⟨1740, 1770⟩, for which the largest possible coupling is of size 9323.

The evaluation of these models in Chapter 3.2.3 showed that the individually trained embeddings lag behind the incremental models up to the 1850s, where the ρ scores on the MEN similarity test set are almost on par. This leads to the expectations that the individually trained ⟨1740, 1770⟩ is difficult to align because of the embeddings' qualities, and that ⟨1860, 1890⟩ could yield a reliable coupling. The first of these expectations is met; the T obtained from $\Gamma^{9323}_{1740,1770}$ on individual spaces contains 6624 translation pairs, 5 out of which are string-matching (vocabulary overlap: 6168). The numbers for $\Gamma^{20000}_{1860,1890}$ are comparable to the large couplings on incremental spaces, but out of the 12000 translation pairs in its T, just 1916 are string-matching (vocabulary overlap: 12422).

These large couplings on individually trained spaces are thus not used in the unsupervised bilingual scenario. Still, the embedding spaces of ⟨1860, 1890⟩ are of a similar quality for both incrementally and individually trained models. This comparability is useful to inform about how the training method of the embedding spaces influences the detection of shifts in them. Therefore, the subsequent experiments will also be performed on the individually trained spaces of ⟨1860, 1890⟩.

5.2 Methods

5.2.1 Detection of Systematic Shifts

In the experiments I investigate a novel approach to finding and labeling systematic conceptual shifts. For this, I combine two well-known and frequently used techniques: nearest-neighbor search and unsupervised clustering with Affinity Propagation (Frey and Dueck, 2007). The novelty of the approach is to apply nearest-neighbor search to

[35] These models are also trained as structured skipgrams, with the same hyperparameters as were used for the incremental models.

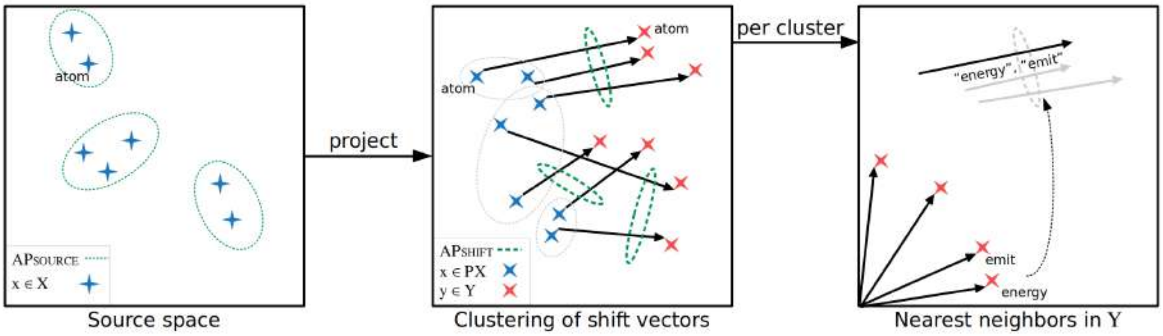


Figure 13: Schema of the proposed method of shift detection. *Left*: clustering variant APsource for comparisons with APshift. *Center*: Difference vectors are clustered with AP (inner distance of a cluster = average cosine between its members). *Atom* is chosen as a cluster exemplar. *Right*: 'labels' are determined by nearest-neighbor search to the exemplar, *atom*.

shift vectors rather than to original or projected ones. To the best of my knowledge, it has not yet been part of research for the purpose of shift detection.

Prior work on the detection of lexical semantic change in word embeddings has seen a wide variety of approaches, many of which are quantitative, measuring the amount of change of a lexical item. However, the aspect of interest here is also a qualitative one: additionally to measuring *how much* a concept or a group of concepts changes, the aim is to uncover in which *direction* this change happens. Given that two embedding spaces X and Y are aligned to each other, the direction of shift between two vectors x and y is trivially calculated as $y - x$. The result is the *shift vector* $s_{x,y}$ which expresses the component(s) of meaning of x that change in the process of shifting to y. For example, if the concept for *atom* shifts from a neighborhood of chemistry to that of physics, its core meaning may remain the same, but it might be understood from a slightly different viewpoint: *atom* might be treated less as a component of molecules, an elementary unit, or an agent in chemical reactions, and be more talked about from the aspects of storing and emitting energy, exercising certain forces, and so on. This slight conceptual change may be captured by $s_{x,y}$, but it cannot be interpreted by humans, because there is no label (i.e., vocabulary item) attached to $s_{x,y}$ to give an intuitive explanation for the expressed shift. The solution to this is similar to the analogy task in Mikolov et al. (2013a): as all shift vectors reside in Y, it should be possible to 'put a name' on any $s_{x,y}$ by finding those y in Y closest to it. Such nearest neighbors should approximately express the components of meaning which change between x and y. The intuition behind a nearest-neighbor search for $s_{x,y}$ in the example of *atom* is that this development towards a more 'physicized' conceptualization of *atom* can be described by those concepts whose embedding vectors point in a similar direction as the shift vector. In this example, such proxy labels might be *power* or *emit*, highlighting that these aspects of the meaning of *atom* become more prominent. Figure 13 visualizes this intuition.

With this approach to labeling conceptual shifts, it would be interesting to investigate how individual concepts change over time. In this case, each $s_{x,y}$ would be characterized by its k nearest neighbors, determined by similarity measures such as cosine distance or CSLS (cf. Conneau et al., 2018). The focus on the following experiments, however, is slightly more macroscopic, as they treat *clusters* of conceptual shifts rather than individual words. This has mainly two reasons. First, it is to be expected that the experiments operate in a noisy environment: every processing step (the construction of the corpus, the training of the embeddings, GWOT for alignment, and finally the orthogonal projection) introduces noise to the environment in which the shift vectors are created, and so it is to be expected that several of the shifts are spurious. Clustering the shift vectors allows to better assess their reliability. For example, when shift vectors are clustered by cosine distance and some $s_{x,y}$ belongs to only a small cluster with a high average pairwise cosine distance among its members (i.e., a 'loose' cluster), then it is likely that this $s_{x,y}$ is noisy. This is less likely for difference vectors belonging to larger and at the same time tighter clusters. The second reason to investigate clusters rather than individual difference vectors is that it offers a continuation of previous work on scientific English in the RSC (cf. Chapter 2.3). Here, the focus lies on the emergence and development of topics and language style rather than the lexical semantics of individual words. Experiments on systematic shifts rather than individual shifts allow to gain insights about these more general dynamics of scientific English from the point of view of distributional semantics.

5.2.2 Experimental Setup

All subsequent experiments on shift detection follow these two principles of clustering and nearest-neighbor search on difference vectors (Figure 14 shows a schematic). The starting point is a space pair $\langle X, Y\rangle$ with corresponding vocabularies U, V and corpus token counts f_X, f_Y. The main distinction between the experiments is the signal, i.e. the set T of translation pairs according to which pairs of embedding vectors from $\langle X, Y\rangle$ are selected to measure change. In the unsupervised bilingual experiment, UnsupBi, T is obtained from the large coupling of $\langle X, Y\rangle$ that has previously been optimized (this is the first of the two connections described in Chapter 4.2). In the monolingual variant, UnsupMono, T is simply the shared vocabulary; each word is paired with itself. The third, supervised experiment DisTech, operates separately on two sets of word pairs which are constructed from lists of discourse-related words and of technical terms (cf. Section 5.4 below). T is restricted to words with a minimum corpus count of 10 and 15 for $\langle 1740, 1770\rangle$ and $\langle 1860, 1890\rangle$, respectively.[36] This filters out the less reliable embeddings of very infrequent words and should improve clustering. For better nearest-neighbors search, Y as well is reduced to embeddings with a minimum token count.

The embeddings of the source words in T are projected with $\mathbf{P}$ (obtained from solving a Procrustes problem with the 100 highest-scoring translation pairs of the space pair's small coupling; the second of two connections between X and Y, cf. Section 4.2). Now

[36] $\langle 1860, 1890\rangle$ has a larger sub-corpus than $\langle 1740, 1770\rangle$; this allows to increase the minimum count while retaining a similar number of words as for $\langle 1740, 1770\rangle$. This reduction in size was also necessary because of memory restrictions during nearest-neighbor search for individual $x \in X$ and $y \in Y$.

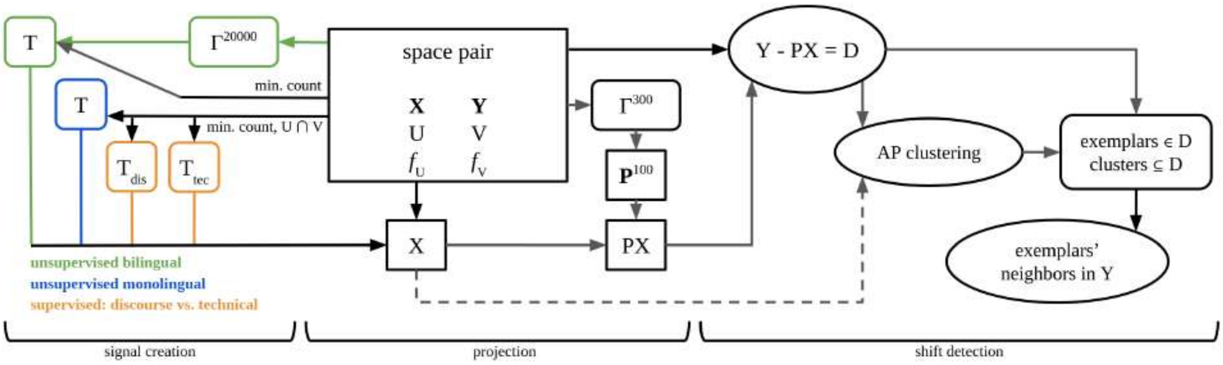


Figure 14: Basic structure of the experiments on shift detection. Clustering alternative APsource vectors of the source space (dotted line) instead of the difference vectors D.

the set D of difference vectors (i.e.: shift vectors) s_{x_u,y_v} is computed by subtracting the projected embedding vector of any u in T from the embedding vector of its associated v. As differences in corpus frequency between the two end points of a shift vector are likely to influence its length, each shift vector is scaled with the logarithm of the absolute token count difference:

$$D := \left\{ \frac{y_v - \mathbf{P}x_u}{log(|f_X(u) - f_Y(v)|)} \mid \langle u, v \rangle \in T \right\}. \tag{22}$$

This scaling method is inspired by Cafagna et al. (2019), who use a similar measure to mitigate frequency effects in shift vectors across two distinct words.

The vectors in D are then grouped into clusters in one of two ways; either by clustering D directly (APshift) or by first clustering the selected x of the source space and subsequently grouping the shift vectors by the cluster of their source vector (APsource). Both variants matter: the default APshift is used to group words with similar conceptual changes; APsource allows to compare the conceptual changes among semantically similar words. Figure 13 visualizes the difference between the two variants. The clustering algorithm used here is Affinity Propagation (AP, Frey and Dueck, 2007), which creates clusters with exemplars as centers. Differently to K-Means or KNN clustering, AP does not require a fixed number of clusters; in this sense, it works 'unsupervised'. The similarity values on which to perform AP are the pairwise cosines of the vectors to be clustered. The intuition behind AP is that data points iteratively exchange messages of two kinds: how well they deem other data points as the exemplar of their cluster; and how well other data points deem them as the exemplar of a cluster. The exchange of messages converges after c iterations without changes in the number of clusters. In the subsequent experiments I set the maximum number of iterations to 100 (empirically) and start out with SciPy's default criterion of $c = 15$, decrementing by one whenever AP does not converge.

The last step collects the potential 'labels' of each cluster by determining the nearest neighbors in Y of each cluster's exemplar. The measure used for this is cosine distance, which is computed from the dot product of two vectors x and y, scaled by their euclidean norm (already mentioned in equation 17 in Chapter 4.1.3):

$$d_{cos}(x, y) = 1 - \frac{x \cdot y}{\|x\|_2 \; \|y\|_2}. \tag{23}$$

For quantitative measurements, the clusters are measured as units in two ways: by the average length of their members (maximum, mean, median, standard deviation) and by the 'inner distance', i.e. the average cosine distance between members of a cluster. Inner distance is inspired by Bizzoni et al. (2020), who devise a similarity measure between two clusters C_1, C_2 by summing all possible pairwise cosine similarities between the members of C_1, C_2 and normalizing with $|C_1| \cdot |C_2|$. In that context, inner distance is calculated in the case that $C_1 = C_2$. As the only cluster distances of interest in this context are inner distances, I do not take into account all possible pairwise measurements. Rather, given a square matrix of cosine distances for C, the inner distance measure only considers the upper triangle of that matrix. Formally, the inner distance of a cluster $C := \{c_1, ..., c_n\}$ is defined as

$$d_{inner}(C) := \frac{\sum_{i=0}^{n} \sum_{j>i}^{n} d_{cos}(c_i, c_j)}{\frac{n^2 - n}{2}}. \tag{24}$$

Inner distance expresses the spread of a systematic shift. A small inner distance within a shift cluster implies that its members experience a similar degree of change in similar components of meaning (a 'harmonic' shift). A high inner distance indicates that the cluster's members change their relative distance to each other, either by growing apart or by converging. For example, the *atom*-cluster in Figure 13 (center) has a relatively small inner distance, because only one of its shifts grows apart from the rest. One of the other two clusters shows a slightly larger inner distance, because its members are less parallel than for the *atom*-cluster; here, however, the concepts shift towards each other.

This is the proposed pipeline for unsupervised detection of systematic conceptual shifts. In addition to these measurements on shift vectors, I also measure the cosine distance $d_{cos}(\mathbf{P}x_u, y_v)$ for each word pair $(u, v) \in T$, and compute each of the words' nearest neighbors, using CSLS by Conneau et al. (2018) to mitigate hubness effects as much as possible.

5.2.3 Evaluation

Post-processing and visualization comprises three parts: the generation of a baseline to the performance of AP clustering; the selection of shift clusters with unusual features; and the visualization of shift clusters.

After the run of an experiment has finished, I construct a clustering baseline in order to put into perspective the effect of APshift with respect to APsource. The baseline for an experiment is constructed by sampling the same number of clusters as created by APshift from a normal distribution, ensuring that (a) the average size of a cluster is similar to APshift and (b) every shift vector in D is a member of a cluster. These random clusters are expected to have the largest inner distances and approximately normally distributed length measurements.

The manual qualitative evaluations treat ‘interesting’ clusters in order to determine whether certain types of words experience different conceptual shifts than others, and if so, how these shifts differ. At first glance, clusters are ‘interesting’ if they have unusual measurements. As measurements I take cluster size (small/large), inner distance (tight/loose), mean length (short/long), and standard deviation of length (even/uneven). To determine interesting clusters in terms of these measurements, I select clusters with significantly small or large values. Formally, I select the set $\mathcal{C}_m \subset \mathcal{C}$ of interesting clusters by means of measurement m from the set $\mathcal{C}$ of all available clusters as follows:

$$\mathcal{C}_m = \left\{ C \in \mathcal{C} \mid \frac{m(C) - \mu(m)}{\sigma(m)} < -1.65 \right\} \cup \left\{ C \in \mathcal{C} \mid \frac{m(C) - \mu(m)}{\sigma(m)} > 1.65 \right\}, \quad (25)$$

where μ and σ are the mean and standard deviation, respectively. This approximately corresponds to the smallest 5% and largest 5% of m-values under the assumption that m follows a normal distribution (Figure 30 in the Appendix shows the KDE-approximated distributions of the measurements).

Shift clusters are visualized as follows: given a cluster $C := \{s_0, ..., s_n\}$ of shift vectors, I retrieve the $x_i \in \mathbf{P}X$ and the $y_i \in Y$ which constitute the $s_i \in C$ as well as the nearest neighbors (i.e., the labels) of the exemplar of C. All of these vectors are simultaneously processed with t-SNE (van der Maaten and Hinton, 2008), which computes 2-dimensional vectors as approximations to the 100-dimensional embedding vectors. Shifts are expressed as lines going from a Px (darker color) to a y (lighter color). The exemplar of a cluster is marked as a red line. The directions of the exemplar’s nearest neighbors (i.e., the cluster’s labels) are marked as colored lines going out from the origin.

By definition, the dimensionality reduction from 100 to 2 implies that a lot of information of the original vectors is lost. The resulting 2D images thus only show a snapshot of the actual spatial arrangement of a cluster’s vectors. As for the lengths of shifts, this means that the lengths of t-SNE vectors do not correlate one-to-one with the lengths of the original shift vectors. Furthermore, as t-SNE is non-deterministic, there are many possible different visualizations (or many different ‘angles’ from which to look at a shift cluster). A fixed point, however, is that the cluster’s labels are defined as the nearest neighbors to the cluster’s exemplar. In order to present more expressive visualizations, I create multiple visualizations and select one, judging the success of t-SNE reductions by how similar (in direction) the labels are to the exemplar.

Note also that t-SNE is sensitive to the number of points visualized. The reason for this is the objective of t-SNE to differentiate between the data points in a given set or arbitrary size (cf. van der Maaten and Hinton (2008)). Consider two cases of visualization: once of a single small shift cluster and once of multiple large shift clusters. In both cases, the algorithm utilizes the same 2D space to maximally differentiate all the given points. In the first (small) case, however, this is often much easier, as there are simply not as many points, each with its nuances, as in the second case. T-SNE visualizations of small point sets are thus often more nuanced; for example, the angle between two shifts will be greater than if these same two shifts were to be visualized as

⟨1740,1770⟩ |X|=5809, |Y|=7408

		APshift			APsource		
	pairs	conv-it	clusters	avg. size	conv-it	clusters	avg. size
UnsupBi	3720	14	324	11.5	15	336	11.1
UnsupMono	4544	5	387	11.7	15	388	11.7
Dis	626	15	68	9.2	15	71	8.8
Tech	119	15	19	6.3	15	18	6.6

⟨1860,1890⟩ |X|=17374, |Y|=25165

		APshift			APsource		
	pairs	conv-it	clusters	avg. size	conv-it	clusters	avg. size
UnsupBi	7970	15	559	12.3	15	559	14.3
UnsupMono	10233*	3	774	13.2	5	748	13.7
Dis	1237	15	126	9.8	15	117	10.6
Tech	252	15	35	7.2	15	33	7.6

⟨1860,1890⟩ (individual) |X|=17374, |Y|=25165

		APshift			APsource		
	pairs	conv-it	clusters	avg. size	conv-it	clusters	avg. size
UnsupMono	10233*	15	545	18.8	15	565	18.1
Dis	1237	11	94	13.2	15	100	12.4
Tech	252	15	29	8.7	15	37	6.8

Table 10: Clustering statistics per experiment and clustering variant.. *conv-it*: criterion for AP clustering convergence (= number of iterations without changes in centroid candidates). *minimum corpus frequency: 15 (instead of 10).

part of a larger point set.

5.3 Unsupervised Experiments

The following sections report on the first two experiments on shift detection, UnsupBi and UnsupMono, which both aim at detecting new systematic shifts without being provided human intuition. The main difference between the two experiments is the signal, i.e., the set T of word pairs which is used to relate the shifted embeddings. In UnsupBi, T is constructed from the large coupling corresponding to the investigated space pair. UnsupMono however does not need to take the step of optimizing a large coupling: the two spaces are assumed to belong to the same language, so T is simply the shared vocabulary of the investigated space pair.

The discussion of the results is divided into several parts, each one addressing and comparing a particular aspect of the two experiments from a quantitative point of view. This is followed by a discussion of specific shift clusters of interest. These sections evaluate both the findings of diachronic language change and the methods used for detection and visualization. For a general reference, Table 10 reports the sizes of the X, Y, and T as well as clustering statistics of all experiments and clustering variants. The default variant of clustering is APshift; the default method of embedding construction is ‘incremental’ (as opposed to the one pair of individually trained embedding models).

5.3.1 Bilingual vs. Monolingual

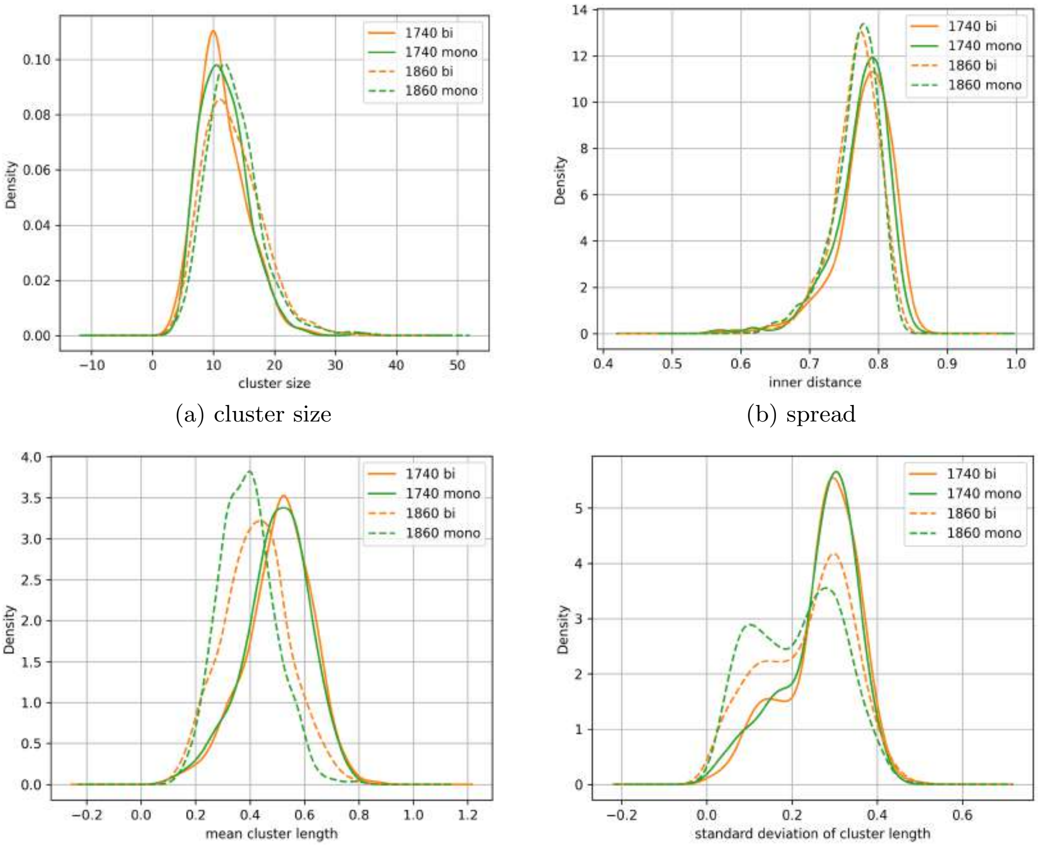
The first aspect of interest is concerned with the effect of the signal on clustering. Put concretely, the questions to be answered are whether the bilingual experiment yields results different to the monolingual experiment and if so, whether this difference is clear enough to justify the efforts of optimizing a large coupling, taken to satisfy the assumption of bilinguality.

Figure 15 shows the approximated distributions of cluster sizes, inner distances (i.e., the 'tightness' or 'spread' of a cluster), mean cluster lengths (i.e., the size of the shift), and standard deviations of shift lengths, for both space pairs in both experiments. The distributions are approximated with SciPy's implementation of Gaussian Kernel Density Estimation (KDE) using Scott's factor (cf. Scott, 2015). It is clearly visible that APShift creates shift clusters of very similar sizes across all relevant conditions, and these sizes are distributed very similarly. The average sizes of a cluster for UnsupBi/UnsupMono are 11.5/11.7 on ⟨1740, 1770⟩ and 12.3/13.2 on ⟨1860, 1890⟩ (cf. Table 10). The clusters' inner distances as well show practically no difference between the two experiments, and neither between the two space pairs (0.780/0.774 and 0.767/0.761), and the distributions are practically identical (Figure 15b).

The averages of mean lengths of the ⟨1740, 1770⟩ experiments are very similar (0.506/0.497); for ⟨1860, 1890⟩, they differ slightly (0.422/0.390). As Figure 15c indicates, the distributions of mean lengths are very similar across the experiments. Lastly, the standard deviation of cluster lengths shown in Figure 15d indicates virtually no difference between experiments for ⟨1740, 1770⟩, but slight differences for ⟨1860, 1890⟩: UnsupBi has fewer evenly long clusters than UnsupMono.

In order to put the observations on UnsupBi and UnsupMono into perspective, it is important to note that filtering infrequent words with a minimum count reduces the number of mismatching translation pairs in the T of the two UnsupBi experiments. For ⟨1740, 1770⟩, this filtering increases the percentage of string-matching pairs in T from 58.4% (in T as obtained from the large coupling) to 90.2% (in the T used in UnsupBi). For ⟨1860, 1890⟩, this percentage increases from 36.6% to 48.7% (cf. Table 8). It is possible that the differences observed with the ⟨1860, 1890⟩ UnsupBi experiment stem from the fact that it is being provided a signal with about 50% of mismatching pairs which, as determined in Chapter 4.3.7, are mostly noise.

These observations on the clusters' statistics suggest that UnsupBi and UnsupMono do not behave differently and that their differences are mainly due to noise in the bilingual signal. This affirms the speculation from Chapter 4.3.7 in which the annotation of translation pairs revealed that only a small number of string-mismatching pairs is valuable (i.e, meaningful at the same time). The advantage of using these valuable mismatching pairs in a bilingual setup can be deemed marginally small. Thus there is no reason to approach the detection of diachronic conceptual shifts as a bilingual task; at least not with respect to the RSC. Consequently, the remaining comparisons of experimental conditions will mostly treat UnsupMono.



(a) cluster size

(b) spread

(c) length

(d) unevenness

Figure 15: Cluster sizes and lengths for both space pairs and both unsupervised experiments (clustering variant: APSHIFT). Distributions approximated with KDE.

5.3.2 ⟨1740, 1770⟩ vs. ⟨1860, 1890⟩

With respect to the two different space pairs, Figure 15c shows that later clusters are on average 21.5% shorter than earlier ones (0.497 and 0.390 for ⟨1740, 1770⟩ and ⟨1860, 1890⟩, respectively).[37] This indicates that language in the RSC changes more between 1740 and 1770 than between 1860 and 1890. In addition, these greater changes vary more in their size: for most of the ⟨1740, 1770⟩ clusters, the standard deviation of their members' length is 0.29; ⟨1860, 1890⟩ instead shows two modes of standard deviation, at 0.28 and at 0.10 (cf. Figure 15d). Although the graphs in Figure 15 are only approximations, it can be said that there are certain conceptual changes in ⟨1860, 1890⟩ which go in a similar direction at relatively similar rates.[38] The cluster size and spread/tightness of the clusters is not considerably different between the two space pairs (cf. igures 15a and b).

The observation of greater changes happening in ⟨1740, 1770⟩ than in ⟨1860, 1890⟩

[37] the default: UNSUPMONO

[38] This is not necessarily a showcase of the law of parallel change, because in the APSHIFT variant here, the 'similar vectors' are not vectors of concepts, but vectors of *shifts* of concepts.

is in line with the findings of Bizzoni et al. (2020). There, lexical innovation (new words entering the vocabulary) of ⟨1860, 1890⟩ is lower and only reaches the levels of ⟨1740, 1770⟩ in the first couple of years. In terms of lexical thinning, i.e., the development of words being used less frequently or not anymore, ⟨1860, 1890⟩ again shows less development. Intuitively, new terminology as a result of lexical innovation has a tendency to change more than established terminology, because it is yet to be firmly positioned in language. The greater lexical innovation in ⟨1740, 1770⟩ might explain the higher mean length of shift clusters.

5.3.3 APshift vs. APsource

All measurements in the experiments are taken on shift vectors or groups of shift vectors. Consequently, it is to be expected that APSHIFT, operating directly on the set of shift vectors, will yield the tightest clusters. The shift clusters created with APSOURCE are based on the starting points of the shifts; as it is not guaranteed that similar concepts shift in similar directions, it is expected that the inner distances of clusters from APSOURCE is higher. The baseline of randomly created shift clusters is expected to have the loosest clusters (i.e., the least harmonic 'systematic' shifts).

Figure 16 shows the KDE-approximated distributions of inner distances for both space pairs and the three clustering variants. The results are as expected and the differences are clear (means for APSHIFT/APSOURCE/BASELINE: 0.774/0.828/0.924 and 0.766/0.821/0.950 for ⟨1740, 1770⟩ and ⟨1860, 1890⟩, respectively). The improvement of inner distance of APSOURCE over the baseline is about twice as large as the improvement APSHIFT over APSOURCE. It is thus possible to group shift vectors into more or less coherent clusters solely based on their source vectors' similarities. This suggests that there is a correlation between the cosine similarity of a given word vector and its shift vector; that additionally to the law of parallel change ('similar words change at similar rates'), similar words tend to change *in similar ways.*

The differences in clustering behavior do not have a big effect on the clusters' length measurements. Figure 16 depicts the maximum, mean, and median lengths as well as the clusters' standard deviations from their mean lengths. The distributions are practically the same for APSHIFT and APSOURCE, but they differ a lot from the baseline which, despite being constructed from the same shift vectors, shows smaller length values for all measurements. For example, the longest shift of most AP-created clusters from ⟨1740, 1770⟩ has a length of about 1.18. For the baseline, almost no cluster contains a shift vector of this length, even though the AP variants clearly indicate the existence of shifts with this length. The cause for this mismatch is unclear; further investigations might shed light onto it.

5.3.4 Incremental vs. Individual

As the earlier decades of the RSC each comprise relatively little data, it would be unfeasible to train them individually, i.e. to initialize them from zero. Shift detection for earlier portions of the RSC (e.g., ⟨1740, 1770⟩) is therefore bound to be carried out on spaces which are constructed with some form of initialization. However, embedding

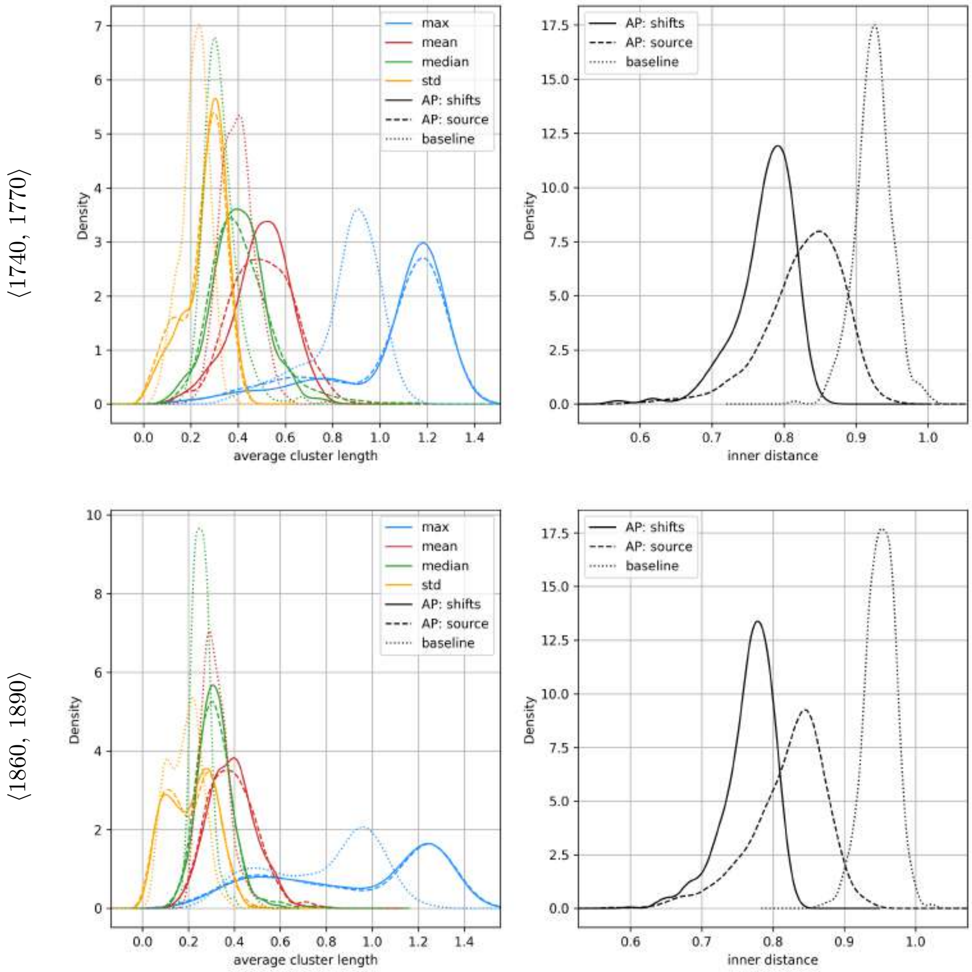
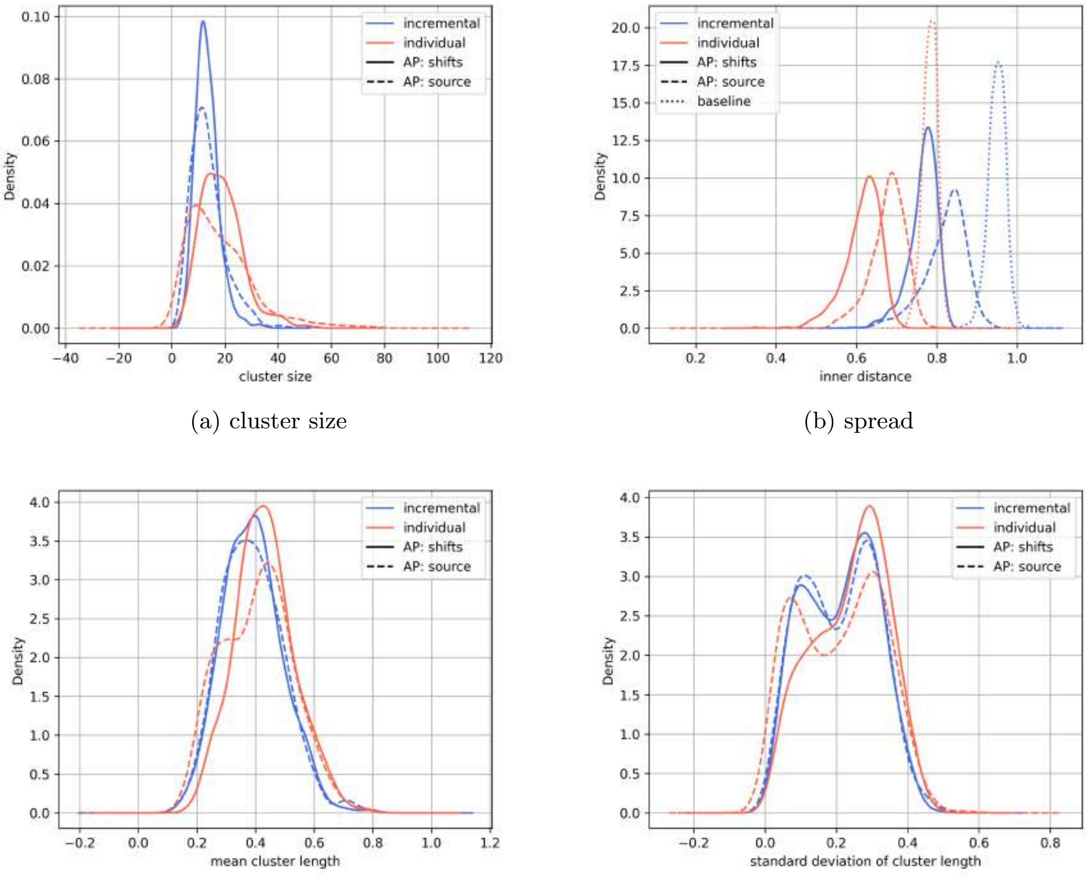


Figure 16: Clusters' average shift length measures (left) and inner distances (right) for the UnsupMono experiment (APshift: solid; APsource: dotted; baseline: finely dotted). Distributions approximated with KDE.

spaces are normally trained individually. Therefore it is of interest to shed light on possible differences between these two training methods and see whether one of them might be better suited for the task of diachronic shift detection. For this, I will discuss the results from UnsupMono on the 1860, 1890⟩ space pairs (incremental/individual), concentrating on the default clustering variant APshift.

The first insight comes from Table 10, which reports that clustering of the incremental space converged at a criterion of 3 (for APsource: 5), whereas AP on the individual space converged with a criterion of 15. In other words: while AP on the incremental space achieved at most only 3 consecutive iterations without changes of possible cluster exemplars, AP on the individual space achieved up to 15 unchanged iterations. This indicates that the clusters found on individual spaces are more stable.[39]

[39] Note that contrary results can be observed for clustering discourse words (convergence criteria for incremental/individual: 15/11). I deem the difference between these convergence criteria not large



(a) cluster size

(b) spread

(c) length

(d) unevenness

Figure 17: KDE-approximated distributions of cluster sizes and lengths for both training methods of the ⟨1860, 1890⟩ spaces (experiment: UnsupMono).

Regarding the measurements of cluster size, spread, and length, further differences are observable. Figure 17 shows the corresponding measurements (the same as for Figure 15, but with the focus on the training method). At the same size of T, the shift clusters determined in the individually trained space pair comprise 4 to 5 more vectors than the clusters on the incremental space pair (cf. Table 10); Figure 17a furthermore shows that the sizes have a larger variance. The mean lengths of these clusters are relatively similar (0.390/0.431), and so is their distribution (cf. Figure 17c). The same is true for the standard deviations of these mean lengths. However, the APShift clustering leads to fewer clusters with even shift sizes.

The greatest difference between the training methods is observable for the inner distances. As Figure 17b shows, the shift clusters of incrementally trained embeddings are considerably looser than those of individually trained ones, indicated by a larger inner distance. This holds for both clustering conditions (0.766/0.616 for APShift,

enough to express a clear difference in ease of clustering.

0.821/0.678 for APsource) as well as for the baseline (0.950/0.784). This means that even when clustered randomly, the vectors from the individually trained space pair change more 'harmonically' in general. This is peculiar, given that the shift vectors are computed from the same T. There are two possible explanations, both based on the fact that the incremental spaces are trained on more and more diverse data. After all, with the incremental method, training is initialized with embeddings which themselves incorporate information about the previous 195 to 225 years of the RSC, which amounts to 23.0M - 48.1M tokens.

The first explanation follows the intuition that having observed more text, the incremental spaces are more nuanced. This increases their ability to express nuances not only of concepts, but also of conceptual shifts – put differently, the spaces are more sensitive to the orthogonal projection. The inner distance measure captures these nuances of shifts. The individually trained spaces, while relatively expressive on their own (cf. Chapter 3.2.3), do not capture as many slight differences when being projected and thus the average cosine distance in their shift clusters is smaller.

The second possible explanation focuses on the fact that while the incrementally trained spaces have more data available, these data are not homogeneous. Spanning a couple of centuries, the data that is supposed to give the incremental spaces a head start captures not just the bare conceptual content, but all of the conceptual change that happened over the years. Moreover, as pointed out by Kim et al. (2014), an embedding from one decade only gets re-positioned well in the next incremental step if its corresponding word occurs sufficiently often in the new portion of text. If that is not the case, the embedding in question remains in its outdated position. It is possible that the 'legacy' of language change as well as the existence of 'outdated' embeddings lead to more slightly spurious embeddings. These slight errors may not be detected in a simple similarity task, but possibly do carry some weight in the shift detection, where they are captured by more difficult AP clustering and a greater average inner distance.

The observations from this comparison indicate that the method of training embedding models matters to the proposed method for shift detection. Individually trained embedding spaces appear to express systematic shifts in larger and at the same time tighter clusters. This has implications for the applicability of the proposed shift detection method to pairs of incremental spaces: in Section 5.3.1, there is practically no difference in inner distance between UnsupBi and UnsupMono on 1860, 1890⟩. UnsupBi, however, uses a T with about 50% noisy translation pairs. If this large amount of noise does not affect inner distance, then it could be that the proposed method for shift detection is not feasible for incremental spaces.

5.3.5 Qualitative Results

Despite the assumption that the detection of systematic shifts on the individually trained spaces leads to better results, I will continue investigations on the results from experiments on incrementally trained spaces. This decision is motivated by the aim to connect the present insights and methods to previous research, which uses incrementally trained embedding spaces (cf. Bizzoni et al., 2019a, 2020). Potential insights on these em-

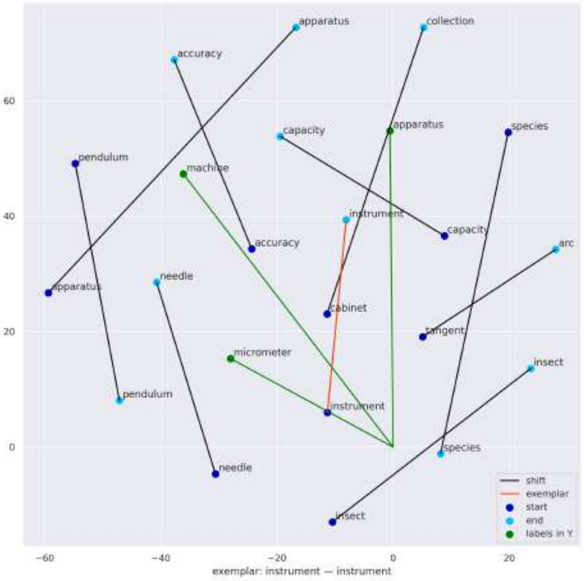


Figure 18: The (*instrument,instrument*)-cluster from UnsupBi on ⟨1740, 1770⟩.

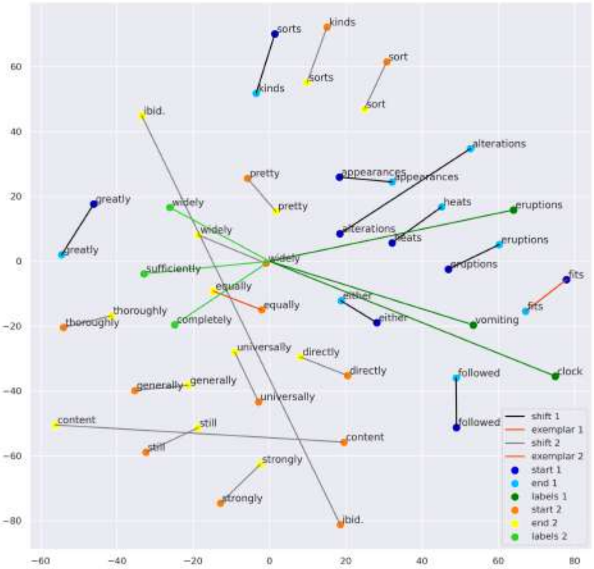


Figure 19: Clusters of (*fits, fits*) and (*equally, equally*), with members (*sorts, kinds*) and (*kinds, sorts*) in the upper part.

bedding spaces are better comparable than findings on individually trained embedding spaces. Furthermore, it is likely that future work on language change within the RSC will make use of incremental training rather than individual training in order to be able to also research the earlier periods of scientific English.[40]

Unsupervised Bilingual, ⟨1740, 1770⟩. The word pairs in T in the bilingual experiments on ⟨1740, 1770⟩ are mostly string-matching; as mentioned earlier, the requirement of a minimum count reduces the proportion of string-mismatching word pairs to about 10%. Shift vectors of mismatching words are thus observed rather infrequently. As most of mismatching pairs are noise, they might be expected to lead to noisy shifts and thus to be included in topically completely different clusters. Within the unusual clusters, there are members of both kinds, matching and string matching, which are topically noisy, i.e. do not fit into the cluster. However, there is no observable difference between matching and mismatching noisy members. For example, the (*instrument, instrument*) cluster, visualized in Figure 18, contains two mismatching shift vectors, (*tangent,arc*) and (*cabinet,collection*). While the earlier is depicted as divergent from the exemplar, the latter mostly follows the exemplar's direction. (*cabinet,collection*) probably is a noisy bilingual signal, but the conceptual shift from *cabinet* in the 1740s to *collection* in the 1770s seems to be similar enough to the shift of certain components of *instrument* to be included in the cluster.

[40] That being said, the promising quantiative results on individually trained word embeddings invite to apply the proposed method of shift detection to the later decades of the RSC, or to other, larger corpora.

Exemplar	Labels	Members
enquiries	labours, botanical, invention, members, purposes	severity, honoured, labours, friends, purposes, examination, petrification–preparation,* notions, descent, removal, qualities, invention, enquiries

Table 11: One of the unusually long clusters from UnsupBi on ⟨1740, 1770⟩. *string-mismatching pair.

Pair	Type	Overlap*	Pair	Type	Overlap*
assise – assize	orth.	45	assign – determine	sem.	14
pour – poured	morph.	27	adjoining – adjacent	sem.	27
aequilibrium – equilibrium	orth.	18	cloud – clouds	morph.	31
toward – towards	morph.	17	adjacent – neighbouring	sem.	13
perpendicularly – vertically	sem.	19	terms – expressions	sem.	22
longitude – longitudes	morph.	24	suppose – imagine	sem.	14
commotion – shock	sem.	19	sorts – kinds	sem.	21
kinds – sorts	sem.	26			

Table 12: String-mismatching word pairs with type of mismatch (morphological/orthographic/semantic) used as part of T in UnsupBi on ⟨1740, 1770⟩. *mutual nearest neighbors of a pair's individual words in their respective space, out of 100.

Among the unusually long clusters, there are several clusters which contain shifts of words related to discourse, that is, words which are typically used across scientific fields to communicate findings. While the clusters are not clear-cut into discourse terms and other words, this is a hint towards the findings of Bizzoni et al. (2020) that there are developments to the style of scientific writing between the 1740s and the 1770s.

One of the unusually long clusters, the *enquiries*-cluster reported in Table 11, concerns the topical changes in the RSC. 3 out of 5 of the exemplar's nearest neighbors in Y are also members of the cluster. This can be interpreted as a form of 'standardization' which I call *conceptual establishment*: the conceptual change occurs most strongly in those components which already characterize the concept. From this point of view, the concepts in the *enquiries*-cluster become stronger in their dominant meaning.

In the case of the *enquiries*-cluster, however, the second-best label to express the conceptual shift (at least of the exemplar) is *botanical*, which clearly belongs to a specific field of science. On the one hand, this is in line with the diachronic topic model of Bizzoni et al. (2020) which categorizes botany and biology as part of a topic 'LifeScience' become slightly more prominent. On the other hand, botany and biology are also summarized under the topic 'Reporting', which experiences a drastic decline throughout the 18th century (cf. Figure 20). This is a point at which further research could start. With the unusually long *enquiries*-cluster, the proposed method for shift detection indicates that these rather domain-independent concepts related to research are amplified in their main components of meaning and at the same time experience a shift towards a specific scientific topic; they receive a slight connotation of '*botanical* enquiries'.

Human-Annotated Translation Pairs. Out of the 2500 string-mismatching translation pairs from the large coupling of ⟨1740, 1770⟩ that have been annotated (cf. Chap-

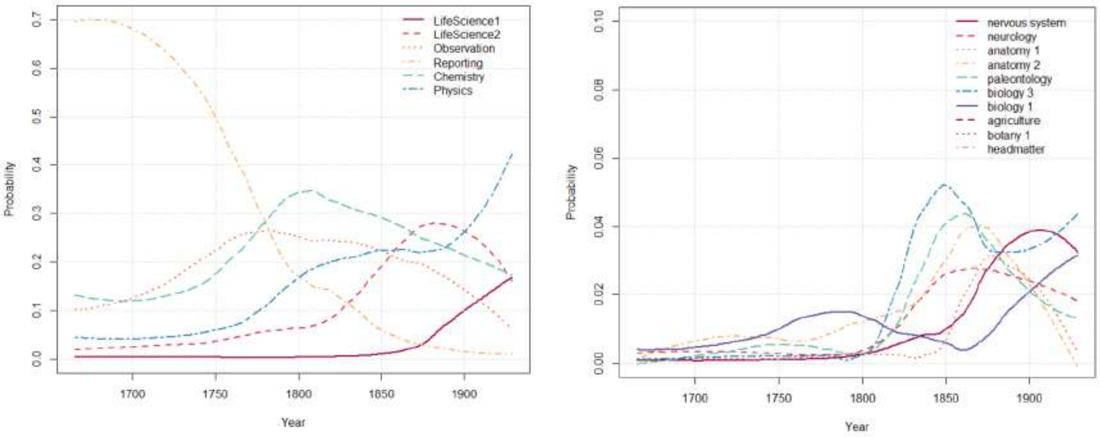


Figure 20: Topic models on the RSC, taken from Bizzoni et al. (2020). *Left*: Distribution of coarse topics over time. *Right*: Distribution of the 'LifeScience2' sub-topics over time.

ter 4.3.7), 71 are labeled as 'good translations', showing only slight differences in spelling, morphological form, or semantic content. 15 of these 71 were considered in UnsupBi (the rest was filtered out by the minimum count requirement). The word pairs are listed in Table 12, which also reports for each pair (u, v) the overlap of the $k = 100$ nearest neighbors of u in X and v in Y. Neighbors were computed with CSLS (Conneau et al., 2018). This approach to measuring individual words' changes via the intersection of their nearest-neighbors sets is similar to Gonen et al. (2020), albeit on a much smaller scale.

The average overlap of orthographically different word pairs is 31.5 (2 pairs); the ones of morphologically different and semantically similar pairs are 24.8 (4 pairs) and 19.4 (9 pairs), respectively. While the number of samples is too small to make a statistically valid statement, the differences should be noted. As the structured skipgram models capture syntactic and semantic information, it is intuitive that the pairs with morphological and semantic differences change more over time than words which differ in spelling.

A somewhat peculiar case are the two pairs (*kinds, sorts*) and (*sorts, kinds*), which are members of distinct shift clusters. The joint t-SNE reduction for an expressive visualization turned out to be difficult; however, the shifts corresponding to the two pairs were consistently placed parallel to each other (cf. Figure 19). These two word pairs in the bilingual signal were constructed by GWOT based on the embeddings' spatial roles. If GWOT erroneously mixed up these two words, this is likely because *kinds* in the 1740s 'behaves' like *sorts* in the 1770, and vice-versa. This suggests that the two words experience a conceptual switch between the 1740s and the 1770s.

Observations on Both Space Pairs. For insights about UnsupMono, I will change space pair and investigate the results of ⟨1860, 1980⟩. Before that, three observations shared for both periods are to be discussed.

The first observation is that shift vectors of words (for UnsupBi: word pairs) with similar grammatical forms tend to be clustered together. At first glance, this is not

Feature	Exemplar	Labels	Members
large	providing	ensuring, save, sustaining, secure, forcing	close, carrying, supposing, allowing, preparing, constituting, absorbing, holding, constitutes, indefinite, regulating, preventing, amply, subjacent, demands, completes, confocal, providing, ensuring, implying, remedied
large	nerves	nerve-roots, roots, nerve, cervical, ganglia	nerve, roots, nerves, skin, cervical, ganglion, ganglia, holes, employment, resulted, thickening, oblongata, nerve-cells, knew, 13th, nerve-roots, turbinal, branchial, labyrinth, hundredths, trigeminus, bellows, pulsations, cava
large	vienna	hamburg, copenhagen, munich, haarlem, amsterdam	proximal, constituent, continue, lately, engineers, excretion, student, screens, miller, vienna, petersburg, aim, popular, munich, addressed, duplicate, thank, bread, twenty-three, quadrilateral, hie, seated, sunset, wh, ejected, clothed, wants, dm, tour, exactitude, constrained, replied
short	percentages	proportions, weights, percentage, proportion, volumes	volumes, indices, freezing, lighter, stature, percentages, speeds
short	must	cannot, may, should, might, can	which, may, will, must, could, should, shall, cannot

Table 13: Clusters from UnsupMono on ⟨1860, 1890⟩ with unusual features.

surprising, given that the structured SGNS embeddings are designed to capture syntactic information; it is expected that within a word field of verbs, for example, the gerund forms and the participle forms tend to form separate sub-fields. However, the clusters are based on the changes of concepts and not on the concepts themselves. The tendency of shift clusters to form along the lines of syntactic differences is an indicator for the great non-semantic diachronic changes in the RSC (e.g., the development a scientific style of writing, cf. Bizzoni et al., 2020).

The second observation is that many clusters have some of their members as labels. As discussed with the example of the *enquiries*-cluster from ⟨1740, 1770⟩, this can be interpreted as the concepts in the cluster becoming more established (mostly) in their main meaning(s). The same is true of the *nerves*-cluster from ⟨1860, 1890⟩ (cf. Table 13), which has all of its labels as members. This leads to the assumption that the overlap of a cluster's labels with its members indicates to a certain extent whether a group of words is enriched with a new meaning (like the *enquiries*-cluster with *botanical*) or further establishes its conceptual core.

The third observation is that as expected, most of the shortest shift clusters mostly consist of function words (e.g., articles, conjunctions), numerals, units of measurement, and modal verbs. However, among these rather stable concepts, there are also content-bearing words, for example the *percentages*-cluster from ⟨1860, 1890⟩ (cf. Table 13). The fact that these usually non-functional words experience significantly little change suggests that they are being used like functional words. Put differently, they are used in scientific writing, but not to express specific concepts — rather, these stable content words contribute to the lexical aspects of the emerging scientific language in the RSC.

Unsupervised Monolingual, ⟨1860, 1890⟩. Many of the shifts in this period which cluster into large groups are numerals and particles, but some of them largely comprise

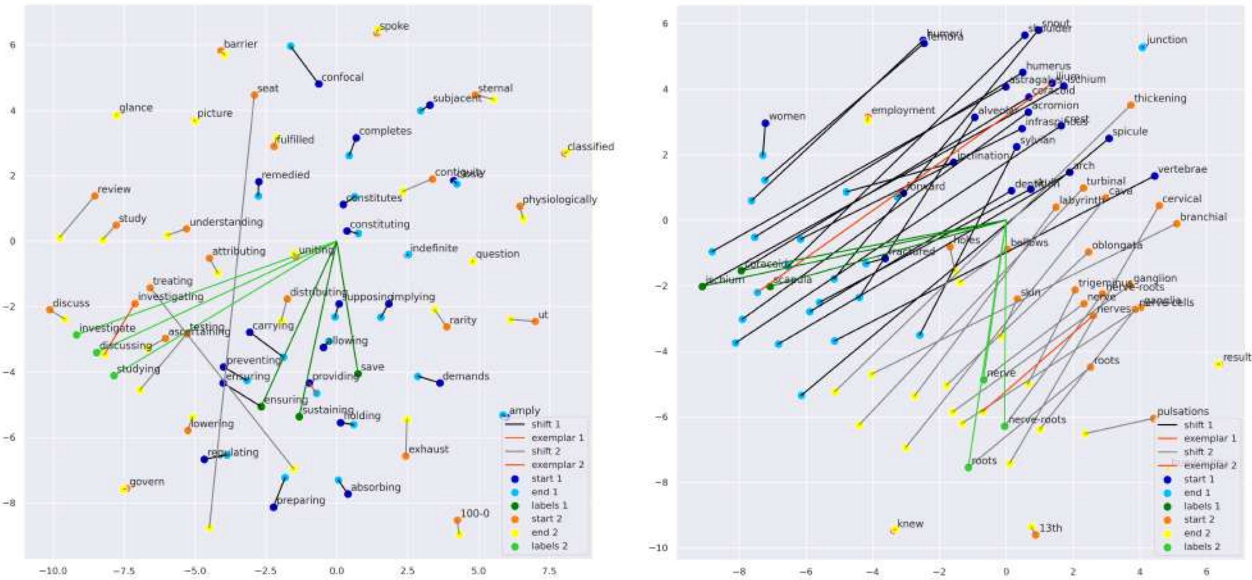


Figure 21: Two of the largest shift clusters from UnsupMono on ⟨1860, 1980⟩.
Left: clusters of *providing* and *investigating* (blue/orange). *Right*: clusters of *ilium* and *nerves* (blue/orange).

either gerunds and participles or words related to anatomy. Two clusters of each kind are visualized in Figure 21. The two pairs of clusters differ fundamentally. The *providing*- and the *investigating*-cluster are difficult to visualize; the depicted shifts do not point into similar directions and are overall shown as short. In contrary, the *ilium*- and the *nerves*-clusters show clear systematic shifts into very similar directions, both within and across clusters. All four clusters are of a similar size (21, 28, 24, and 24), so it is unlikely that these differences in the visualizations can be attributed to t-SNE's sensitivity to point set size. These observations clearly demonstrate that the concepts related to anatomy are under the influence of dynamics of conceptual change. I deem this dynamic to be conceptual establishment because of the large overlap of cluster labels with their members.

Indeed, the topic model of Bizzoni et al. (2020) (cf. Figure 20) shows that between 1825 and 1880, the 'LifeScience2' topic greatly increases in popularity. Furthermore, the sub-topics 'anatomy1' and 'nervous system' experience a lot of growth. As for the *providing*- and the *investigating*-cluster, it is plausible that the gerund and participle forms only change marginally given that the formation and consolidation of English scientific writing happened before this time, between the 1750s and the 1850s (cf. Degaetano-Ortlieb and Teich, 2016).

From a methodological point of view, the cluster with *vienna* as exemplar marks a shortcoming of the proposed approach to finding appropriate labels systematic conceptual shifts. This cluster appears to be made up of concepts which are often used in the editorial parts of publications (cf. Table 13). These sections are mainly concerned with the personal lives of renowned scientists and members of the Royal Society. As the exemplar is the shift of *vienna*, which on its own does not shift significantly, it is

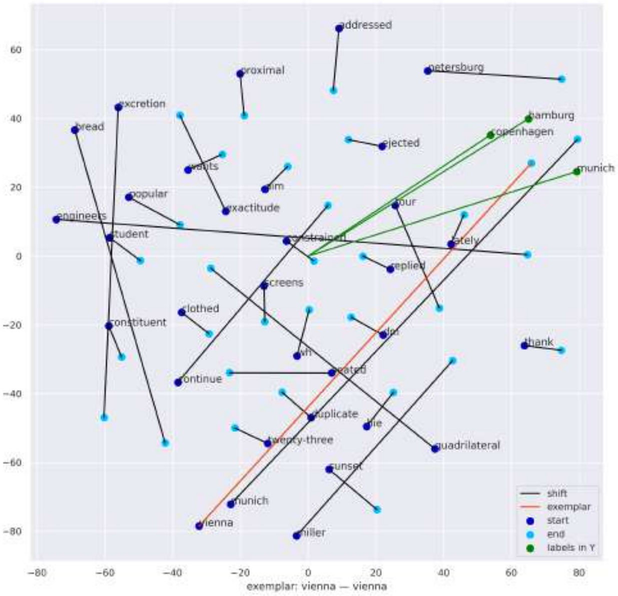


Figure 22: The *vienna*-cluster from UnsupMono on ⟨1860, 1890⟩.

understandable that the cluster's labels (i.e., the exemplars nearest neighbors) are all names of cities. The issue here is that these labels are not representative of the whole shift cluster, because the cluster itself contains the shifts of many other words which themselves are likely experience slight shifts into many other directions. The t-SNE visualization thus positions especially the shifts of city names in parallel to the labels, but neglects almost all other members of the cluster (cf. Figure 22). This shows that the proposed method for labeling the systematic shifts needs to be improved.

5.4 Supervised Experiment: Discourse Words vs. Technical Terms

The third experiment builds on sets of words that have previously been found to behave in specific ways in the RSC. These words of interest can be categorized as 'discourse terms' and 'technical terms'; Table 14 gives an overview and shows examples. The set of discourse terms consists largely of adjectives occurring in constructions like *it is possible*. Other types of discourse terms are words ending with *-ing* (often poly-functional words, which can be used as gerunds, participles, or present continuous forms). Most of the *-ing* verbs in the set occur before *that* to form phrases such as *assuming that...*. The rest are two out of three groups of *-ing* verbs which have been characterized by cf. Bizzoni et al., 2019a as 'change of state' and 'motion' verbs. The set of technical terms contains the third group of *-ing* verbs, which are characterized as 'academic' words. The main portion of this set, however, is a combination of two large topical clusters from Bizzoni et al. (2019b), one consisting of chemistry-related words, the other one with words related to galactic astronomy.

For the investigations with these words of interest, the main experiment is performed

Source	Name	Size at 0/10/15*			Use	Examples
G&M	-ing verbs1	30	10	26	DIS	determining, establishing, studying, ascertaining
	-ing verbs2	30	19	28	DIS	passing, extending, running, reaching, bending
	-ing verbs3	30	13	21	TECH	purifying, warming, pouring, removing, plunging
	-ing-that verbs	1613	200	234	DIS	producing, pouring, denoting, breaking, preventing
PHD	it-adjectives	2796	434	684	DIS	adequate, larger, wrong, suspected, special
DWNET	galaxy	234	30	54	TECH	magnitudes, nebula, methods, faint, polar
	chemistry	681	78	163	TECH	zinc, carbonaceous, gramme, incandescence, potash

Table 14: Words of interest for the DISTECH experiment. G&M: Bizzoni et al. (2019a). PHD: Degaetano-Ortlieb (2015). DWNET: Bizzoni et al. (2019b). *Original size and size with minimum 10/15 occurrences in $\langle 1740, 1770\rangle/\langle 1860, 1890\rangle$.

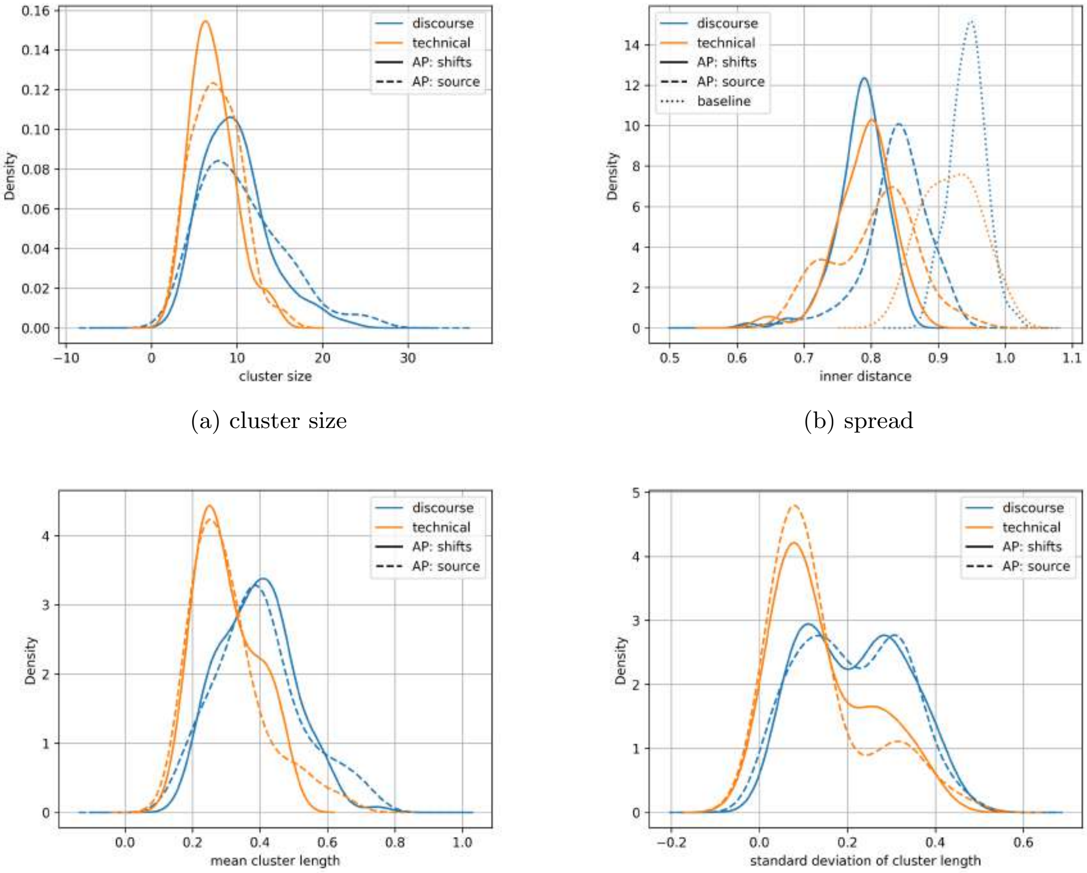
twice, once with the discourse terms and once with the technical terms serving as T. Both of these sets of words contain a small number of words which at first sight belong to the other type. That is, there are *-ing* verbs in the set of technical terms and science-related adjectives in the set of discourse words. This is because the distinction between 'discourse words' and 'technical terms' is not clear-cut. For example, one of the three smaller groups of *-ing* verbs is considered to belong to technical terms, because it contains change-of-state verbs like *purifying, discharging*, or *bending.* As DISTECH is not performed on the two T simultaneously, these cross-over concepts can potentially be informative for the qualitative analysis of the results.

The following sections will focus on the differences between these two groups of words, first from a statistical (quantitative) and then from a qualitative perspective. As an in-depth discussion of DISTECH on all three space pairs would exceed the scope of this thesis, most of the reported results stem from DISTECH on the incrementally trained $\langle 1860, 1890\rangle$ only.

5.4.1 Quantitative Results

Table 10 in Section 5.3 shows the overall results on clustering. The imposed minimum corpus occurrence of 10 leads to different numbers of words considered; for ⟨1860, 1890⟩, the two sets contain $|T_{dis}| = 1237$ and $|T_{tech}| = 252$ words, which are grouped into 126 and 35 clusters, respectively. The difference in set size between discourse word and technical terms could influence the results from AP clustering; this should be kept in mind.

Considering differences in clustering, AP converged successfully at a criterion of 15 iterations for most runs of DISTECH, irrespective of the T. The average size of the clusters (discourse/technical) is 9.8/7.2 for APSHIFT and 10.6/7.6 for APSOURCE; the approximated distributions are shown in Figure 23a. It shows a tendency for discourse terms to form larger shift clusters in general and more very large clusters in specific. The differences between clustering variants is negligible. These two facts in combination indicate that technical terms are overall more specialized: they form smaller groups in the source space already and when diachronic changes into a certain direction occur, they apply to smaller groups of concepts when these are technical. However, the observed differences could partly be due to the differing sizes of T_{dis} and T_{tech}, so this inference



(a) cluster size

(b) spread

(c) length

(d) unevenness

Figure 23: Distributions of cluster sizes and lengths for DisTech on ⟨1860, 1890⟩.

should be taken with caution.

The approximated distributions of inner cluster distance depicted in Figure 23b show that when clustered by similarity of diachronic shift, discourse terms do not behave differently from technical terms. The two sets of words do however differ in the size of their shifts; Figure 23c shows that the mean length of a shift cluster is greater for discourse words than for technical terms (0.391/0.302); in other words: discourse words, on average, change more than the technical terms selected for this experiment. Most of these technical terms belong to one of two topics; galactic astronomy and chemistry. The topic models in Figure 20 report that the presence of the general topics ‘Observation’ (for galactic astronomy) and ‘Chemistry’ are both on a slight decline, but still among the more frequently treated topics. The relatively little change of technical shift clusters thus does not lie in these words not being used. Rather, it is plausible that these concepts established a stable meaning over the course of the the preceding decades so that between the 1860s and the 1890s, they are used in a limited set of contexts; they have undergone *standardization*.

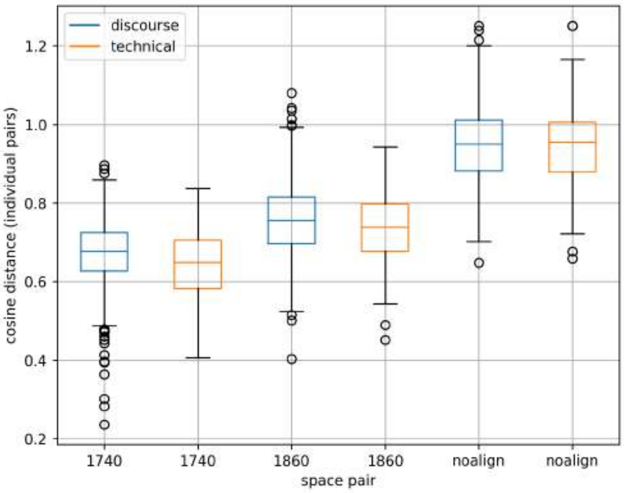


Figure 24: Average cosine distance of individual words ('noalign' = individually trained).

This difference between discourse words and technical terms in terms of the amount of conceptual change is not clearly observable from individual diachronic shifts alone. Figure 24 shows the average diachronic differences of single words $w \in T$, measured by the cosine distance $d_{cos}(w) = d_cos(\mathbf{P}x_w, y_w)$, for the DisTech experiments on all 3 space pairs. Across the space pairs, the average distances do differ. However, within one and the same run of DisTech, e.g. on $\langle 1860, 1890\rangle$ in the middle, the distances of discourse words differ too little from those of technical words in order to conclude a difference in shift behavior. This shows that the type of measure for conceptual shift matters: the frequently used cosine distance is often employed successfully, but in some cases like the present, other approaches such as measuring the length of difference vectors might be more sensitive to changes.[41]

Lastly, Figure 23d shows differences in standard deviation of the clusters' shifts' lengths. While shift clusters of discourse terms vary widely in their 'evenness' (mean: 0.271), most shift clusters of technical terms show a relatively small standard deviation (mean: 0.146); their shifts tend to have very similar lengths. The greater variety of evenly and unevenly long shift clusters for discourse terms may be an effect of the larger T_{dis}, which possibly comprises a wider range of concepts (compared to the mostly galaxy- and chemistry-related technical terms). Accommodating all these concepts' shifts into clusters, it becomes more likely that the evenness of lengths of a cluster decreases.

To conclude the quantitative results of the DisTech experiments, there are certain observable differences between the diachronic behavior of discourse words and technical terms. The latter tend to cluster in smaller, shorter, and more evenly-sized groups. In terms of spread of these shift clusters, there is no observable difference.

[41] Of course, the same is true in the opposite direction. Using both measures in a complementary way at the same time would be the ideal solution, if this is feasible.

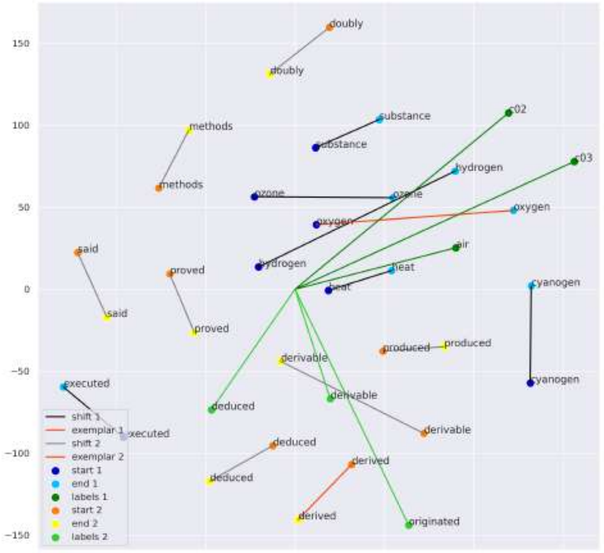


Figure 25: The tightest and loosest clusters of technical terms: *oxygen* (blue) and *derived* (orange).

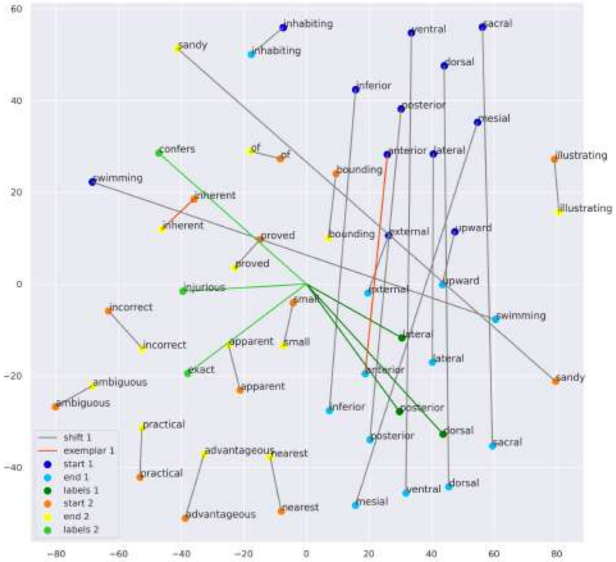


Figure 26: Discourse term clusters of *anterior* (blue) and *inherent* (orange).

5.4.2 Qualitative Results

Among the unusual clusters of the technical terms shifts, one of the tightest clusters is the *oxygen*-cluster. Figure 25 visualizes this in contrast to the *derived*-cluster, which is the loosest of technical clusters. Table 15 reports these two and the other clusters mentioned in the following. Although many of the members of the *derived*-cluster can be seen as the shifts of technical terms, their meaning is more general, because these words express methodology (e.g. *deduced, proved*). Furthermore, they are mostly verbs in past tense and could pass as discourse words. This is the first indicator that the two groups devised for the DisTech experiment behave differently on the qualitative level.

Further insights of this kind are gained from the words potentially to belonging to both groups, or even clearly belonging to the other group. A case of the latter is the *anterior*-cluster formed by DisTech on discourse words. It is relatively (albeit not significantly) tight and comprises mostly expressions of direction which are specific to anatomy. Similarly to the *ilium*- and the *nerves*-cluster from UnsupMono, all of the labels of the *anterior*-cluster are also present as members; it experiences conceptual establishment. This cluster is another example for the development in the highly specialized field of (neuro-)anatomy, which apparently progresses without the influence of other sciences. Figure 26 shows this cluster in comparison to the *inherent*-cluster of discourse words which is equally large and relatively (but not significantly) loose. It visualizes the movement of the anatomy-related concepts from the 1860s to the 1890s as a collective shift from one small region in Y, namely the region onto where the x were projected (upper-right part in the 2D image), to another small region in Y. In contrast to this, the *inherent*-cluster is depicted as a collection of diachronic shifts for which the

	Feature	Exemplar	Labels	Members
Discourse	long	accepting	assuming, finding, assigning, classifying, granting	accepted, grouping, discovering, believing, adopting, old, overcoming, accepting, asserting, many, rejecting, mistaken
	long	spreading	curving, shoots, slanting, turning, shooting	paying, investing, curving, spreading, supplying, inconspicuous, floating, hard, swelling, ing
	short	seen	found, observed, noticed, disclosed, evident	seen, occurring, so, observable, obvious, less, noticed, evident, noted, observed
	short	special	especial, particular, external, undue, valvular	new, particular, special, immediate, former, same
	est.	anterior	posterior, lateral, dorsal, ventral, mesial	upward, lateral, sacral, anterior, external, dorsal, ventral, mesial, inferior, swimming, inhabiting, posterior
	loose	inherent	exact, confers, injurious, ciose, discriminate	incorrect, practical, sandy, illustrating, of, proved, nearest, advantageous, inherent, bounding, small, ambiguous, apparent
Technical	tight	oxygen	c02, c03, air, nitrogen, carbon	substance, ozone, oxygen, executed, heat, cyanogen, hydrogen
	loose	derived	derivable, originated, deduced, arisen, eliminated	said, doubly, proved, derived, produced, methods, derivable, deduced
	large	purifying	refilling, sealing, regulating, tap, aspirating	sealing, closing, palladium, aspirator, purifying, discharging, bending, baryta, transmitting, exploding

Table 15: Clusters from DisTech on ⟨1860, 1890⟩ with unusual features (*est.*: large overlap of labels with members).

starting points (i.e., the $\mathbf{P}x$) are not as close to each other and subsequently shift into various directions; their diachronic development is less synchronized.

Among more ‘clear cases’ of discourse words, there is an observable difference between *-ing* verbs and adjectives. Many of the shift clusters with the longest average shifts comprise mainly *-ing* verbs, such as the clusters of *accepting* and *spreading.* The shortest shift clusters of discourse words, in contrast, are often formed by adjectives, such as the *seen-* and the *special*-cluster. This indicates that gerund and participle forms take a prominent role in the developments of the style of scientific writing in the second half of the 19th century. Further inquiries comparing these verb forms to other types of discourse-related words such as the adjectives in the aforementioned shift clusters will likely confirm this finding.

Besides showing greater changes in relation to other discourse-related words, *-ing* verbs tend to form large shift clusters in general. Even among technical terms, one of the largest shift clusters is *purifying*-cluster, containing many change-of-state verbs which are likely used to report experimental setups and describe processes within experiments. The fact that large groups of *-ing* verbs change in similar ways is explained by their wide-spread usage across scientific disciplines which prevents them from forming smaller, distinct groups of concepts which would likely continue to change independently from each other.

5.5 Conclusions

The experiments and their evaluation paint a positive picture with respect to the questions at the beginning of the chapter:

1. Is it useful to frame the detection of diachronic conceptual change as a bilingual task? The assumption of bilinguality, originally taken to be able to account for the most subtle shifts in a diachronic setup, is not necessary. The outcomes of the bilingual and the monolingual variants of the experiments are similar to such an extent that the efforts of optimizing a large coupling with GWOT can easily be avoided.

2. Does the proposed method for unsupervised detection of shifts work; is it sensible and expressive? As it is tested on only two cases of diachronic comparison, it cannot be said with certainty whether the proposed method can detect conceptual changes reliably. Nonetheless, its results are promising. For example, the proposed method correctly indicates previously identified diachronic developments like topical changes such as the rise of botany in the second half of the 19th century. The nearest-neighbor approach to expressing the direction of conceptual change in an interpretable way has shown its strengths and weaknesses. With this approach, the proposed method confirms the intuition that groups of highly specialized concepts retain their meaning; however, it is not sensible enough to express the nuances of conceptual change happening in larger and more diverse shift clusters.

3. Do the insights from the proposed method relate to previous findings? There are multiple findings in accordance to previous research. Firstly, fact that scientific language experiences specialization (cf. Bizzoni et al., 2019b) is expressed by concepts from neurology and anatomy collectively experiencing 'conceptual establishment', i.e. shifts in the direction of their own main conceptual components. Secondly, it is shown via the size of collective conceptual shifts that language change in the RSC happens at a greater rate between the 1740s and the 1770s than between the 1860s and the 1890s, which is in line with the results from Bizzoni et al. (2020). Thirdly, specialized, technical terms change differently from words which are mainly used to communicate about research, supporting the findings of Bizzoni et al. (2019a). The latter form larger clusters, even when they have a specialized meaning. Contrary to Bizzoni et al. (2019a), the experiments also show that between the 1860s and the 1890s, discourse words tend to show greater shifts than specialized technical terms. This does not show in measures of individual cosine distances; it is the clustering of shift vectors which makes the tendency observable.

4. Can the proposed method uncover new (general and/or specific) dynamics of language change within the RSC? The main insight about the relationship between groups of concepts and their diachronic behavior relative to each other is that similar concepts tend to shift in similar directions. This is closely connected to the dynamic I call 'conceptual establishment' — concepts are confirmed in their core components of meaning rather than changing connotations. There are many cases of conceptual establishment observable in the experiment results. Indeed, conceptual establishment is likely the default mode of change to be observed in diachronic investigations, as naturally, most of language is rather stable across time. In these cases, the contexts in which a word is used do not change considerably; observing more text data means observing more of the context patterns already observed. When two similar concepts (i.e. concepts with overlapping core components of meaning) both experience this conceptual establishment, they likely experience amplification in the shared components of meaning, which makes them shift into similar directions. Further investigations, possibly with more sophisticated methods for describing a cluster's main components of change, may discover dynamics of language change still unknown.

6 Discussion

Characterizing conceptual shifts in word embeddings is no straightforward task. That is, detecting and quantifying change can be as simple as comparing a concept's nearest-neighbors at different points in time (cf. Gonen et al., 2020). Rather than the question of *how much*, it is the question of *how* a concept changes, and of how it changes relative to other changing concepts, that makes this a challenging task. This is because the answer to this question cannot be expressed in a single number (e.g., cosine distance, number of shared neighbors, number of nearest neighbors); it is as multi-faceted as the changing concept itself.

In the previous chapters, I have taken on the task of characterizing conceptual change on the level of small groups of concepts. Each of the various techniques which I combined to arrive at the results in Chapter 5 has certain limitations and thus introduces possible weaknesses to my approach. In this chapter, I reflect upon these points as well as upon potential methodological shortcomings, presenting ideas for improvement in future research. The structure of the discussion follows the main components of the approach, beginning in Section 6.1 with the discussion about the embedding models employed for the task. It continues with the alignment via GWOT in Section 6.2 and the shift detection in Section 6.3. This is followed by a discussion of the less intuitive findings in Section 6.4. The chapter closes with directions for future research in Section 6.5.

6.1 Word Embedding Methods

The central requirement for the construction of reliable word embedding models is the amount and quality of text data available. For diachronic and especially for domain-specific studies of language, it is difficult to collect homogeneous and representative data in quantities large enough to construct reliable word embeddings. As shown in Chapter 3, the first 200 years worth of text in the RSC do not provide enough information to create reliable decade-wise embedding models by simply training each decade separately. Training embedding spaces incrementally, i.e. initializing the training of one time span with the embedding space from the previous time span (cf. Fankhauser and Kupietz, 2017; used in Bizzoni et al., 2019a and re-used in the present work) alleviates the sparsity problems, but it leads to embedding spaces which are slightly skewed towards the past.

To illustrate this, consider the metaphor[42] of a reader of the RSC consuming text chronologically, and in contrast to this, two observers of two small and in themselves

[42] This metaphor was brought up by Yuri Bizzoni in a conversation.

chronologically coherent sub-corpora of the RSC. Starting with the earliest texts, the 'reader in time' models their conceptual space according to these first contributions to the Royal Society. As the reader progresses in time, the same words occur with new meanings. In order to account for these changes, the reader in time has to modify the old concept and re-position it in their conceptual space. The new position of the concept is likely a trade-off between the positions of the old meaning (i.e. what the reader already knows) and the new meaning (i.e., where the concept would fit best for the current time). In contrast, each of the two independent observers is agnostic to anything outside of their small sub-corpus, which includes any of the meanings that a word might have had previous to the time span of their sub-corpus. Upon encountering the word, the observer is free to place it in their conceptual space wherever it is best suited to model the observed sub-corpus, without being required to take into account the word's position(s) in the conceptual space at earlier points in time.

This difference between the incremental (i.e., reader in time) and the individual (observers of periods) training of embedding models shows in Chapter 5.3.4. Here, the temporal interdependence of incremental embedding models results in smaller, less harmonic shifts when compared to shift detection on individually trained models. It is possible that incremental models are not as well-suited for the task. Despite this, the main portion of experiments is carried out on such incremental models, for two reasons. First, the results are better comparable to previous research (Bizzoni et al., 2019a, 2020) that was carried out on the exact same embeddings. Second, the proposed method is tested under the conditions at hand, i.e. the diachronic setting of the RSC, in order to investigate possible future directions of research on this corpus. The experiments on $\langle 1740, 1770 \rangle$ yield informative results and would not have been possible with individually trained embeddings, given that the sub-corpora of the 1740s and the 1770s each contain less than 2M tokens.

That being said, the more recent parts of the RSC, starting with the 1850s, comprise sub-corpora large enough to train reliable models individually, as shown in Chapter 3.2.3. Research on these periods should not be restricted to using incremental embeddings and may profit from applications of the proposed method for shift detection to individually trained models.

Moreover, it would be interesting to apply the proposed method to embedding models of larger diachronic corpora such as CCOHA (Alatrash et al., 2020) — which, however, does not focus on the scientific domain.

6.2 Alignment with Gromov-Wasserstein Optimal Transport

The approach to alignment of embedding spaces for diachronic studies in this thesis is subject to the assumption of bilinguality: in order to make sure that no subtleties of language change are overlooked, I assume that embedding spaces built on corpora from two different points in time should be treated as embedding spaces from two different languages (or, less extreme, as distinct language varieties). Furthermore, I do not directly inspect the incremental (and therefore in principle comparable) embedding spaces; instead, they are first aligned. This step promises to minimize the mostly spurious diachronic differences of the most stable concepts while at the same time amplifying the developments present in concepts which actually experience diachronic change.

Under the assumption of bilinguality, GWOT (Alvarez-Melis and Jaakkola, 2018) is appealing as a method, because it does not require any information about the two embedding spaces to be aligned; there is no initial 'bilingual signal' involved. Furthermore, the alignment with GWOT is convenient for the diachronic scenario, because it is carried out in two steps: the construction of a bilingual signal first, the actual alignment via Procrustes (Schönemann, 1966) second. The bilingual signal from the first step allows to relate pairs of vectors in the aligned spaces irrespective of how far they are apart. Lastly, GWOT has previously shown reliable performance in real bilingual tasks (cf. Alvarez-Melis and Jaakkola, 2018).

However, the insights gained from the application of GWOT to the incremental embeddings of the RSC dampen the view on this method for alignment. The outcomes vary a lot in quality and despite thorough investigations on the present embedding spaces (cf. Chapter 4.3), the reasons for these mixed results — as well as for the numerical errors which require to set the regularization parameter considerably higher than usual — remain unclear. Further investigations on a greater variety of space pairs could paint a clearer picture.

In any case, the difficulties encountered with GWOT do not hinder the alignment. The evaluation of the bilingual signal obtained from a successful coupling with GWOT suggests that the assumption of bilinguality is not necessary. This allows to avoid the largest part of computation with GWOT, which is required under the assumption of bilinguality to create the bilingual signal. GWOT as an unsupervised method remains useful for the task of diachronic alignment as it can identify those concepts which change the least over time and thus help to base the alignment of two spaces on their most stable concepts.

Most criticism concerning unsupervised approaches to alignment (cf. Søgaard et al., 2018; Vulić et al., 2019) is targeted at the limitations of these approaches when they are confronted with alignment tasks involving embedding spaces of typologically distant languages, different embedding methods, or different domains. Furthermore, the use of orthogonal projections in many of these approaches has been shown to be inadequate (cf. Patra et al., 2019; Glavaš and Vulić, 2020), because embedding spaces, especially those of more distant languages, are not isometric to one another. This criticism is justified and supported by evidence. However, the embedding spaces of the RSC are clearly not from typologically distant languages, all alignments are carried out between embedding models constructed with the same method, and the RSC is designed as a domain-specific corpus. In other words: none of the identified and criticised weaknesses applies to the alignment task in this thesis. Still, this does not mean that GWOT is the only or best solution to alignment in this context. Other methods such as the ones sketched out in Chapter 2.2 could be applied.

Lastly, I did not investigate shift detection on the incrementally trained embedding spaces without aligning them. Determining the effect size of such alignments (i.e., minimization of shifts of stable words and amplification of other shifts) is left for further research.

6.3 Detection of Systematic Shifts

The proposed method for shift detection is relatively modular: it combines several techniques to relate word embeddings and their changes to one another, but the combined techniques operate mostly independently from each other. Therefore, it is possible to alter single components of the shift detection pipeline without the need to make compromises at other points. Considering the results in chapters 5.3 and 5.4, there are points for improvement as well as alternatives for some of these components.

First, the difference vectors, which are at the center of the shift detection, may be scaled more sensibly. They are not normalized, because differently to measurements via cosine distance (i.e., angle between vectors), it is the shift vectors' length which expresses the magnitude of the conceptual shift. Nonetheless, the shift vectors need to be scaled in order to take into account differences in corpus frequency. The method of scaling used here (cf. Chapter 5.2), adopted from Cafagna et al. (2019), is a good start. However, it is likely that there are more sensible and mathematically or algorithmically more sound ways to mitigate frequency effects, and these alternatives, once identified, can simply replace the present method.

Second, the number of concepts investigated for shifts is limited by the experimental setup in the form of a minimum count requirement (i.e., only shifts of embeddings with a minimum corpus count are considered). In the unsupervised experiments, this limitation is partly due to hardware restrictions during nearest-neighbor search for individual embeddings. Improvements of the memory management during these experiments will allow to lift the requirement and investigate every concept for diachronic shifts.

The restriction by minimum count is also used to reduce the number of noisy shifts (i.e., conceptual shifts into individual directions, not belonging to any tight cluster) in further processing steps. The inspection of shift clusters (cf. chapters 5.3.5, 5.4.2) reveals that there is still some noise present. A possible solution to both lifting the minimum count restriction and mitigating noise is to introduce a preliminary clustering step employing DBSCAN (Ester et al., 1996; Schubert et al., 2017). This clustering algorithm can recognize and exclude outliers from the clusters. Applied to the identification of noisy shift vectors, it could be used to indicate which vectors to consider for further processing.

The third point for improvement is the method used to find 'labels' that best describe the main components of a cluster's collective shift. Before discussing this point, the difficulty of the task should be put into perspective. The key property exploited to obtain a shift vector is the compositionality of word embeddings (cf. Chapter 2.1.2): by subtracting one concept from the other, the resulting vector expresses those components of meaning which are different between two concepts (or: a concept at two points in time). This shift vector is the result of a combination of two embedding models, one of which additionally is rotated and scaled by a projection operation. As a comparison, the word analogy task (e.g., find the x in '*Rome is to Italy as Berlin is to* x') tests the quality as well as the degree of compositionality of an embedding space. Mikolov et al.

(2013c), for example, report almost 40% correct answers with their best model in such an analogy task – but this is in a syntactic variant of the task (*cat:cats :: dog:x*), it involves only one single model, and this model is trained on 320M tokens. Compared to the circumstances under which these 'almost 40%' were achieved with the present situation, it becomes clear that finding the correct label for a conceptual shift (i.e., for certain *components* of a concept) is not at all an easy task.

Coming back to the third point for improvement, the qualitative investigations show that the labels of a shift cluster, being defined as the nearest neighbors (in the target space) of the cluster's exemplar, sometimes fail to express the general tendency of shift. This is especially likely when the cluster's members take on multiple slightly different directions. The reason is that the point of reference for nearest-neighbor search is the exemplar shift, determined by the Affinity Propagation (AP) clustering algorithm (cf. Chapter 5.2). As Dubossarsky et al. (2015) show, there can be an important difference between distances computed with respect to a cluster's exemplar and distances computed to its mathematical center.[43] In the proposed method for shift detection, AP is the clustering algorithm of choice, because differently to DBSCAN or K-Means clustering (Sculley, 2010), it can be applied without making assumptions about the shape or layout of the shift vectors. The drawback in this context is that there is no notion of (mathematical) center; AP is exemplar-based. A possible solution is demonstrated by Bizzoni et al. (2019a) who first perform AP to inspect the vector space and then use the gained information to set the parameters for K-Means. The centers of such K-Means clusters might be better suited for the nearest-neighbor search for labels.

The last step of shift detection, after clustering and nearest-neighbor search, is the selection of 'interesting' clusters. For this step, I rely on statistical significance and inspect those clusters which show unusual measurements, e.g. especially long, tight, or large clusters. In the qualitative analyses, I explore these outliers and identify that certain types of words tend to shift in certain ways (e.g., *-ing* verbs tend to shift in large clusters; numbers and numerals tend to shift in tight clusters). This method of selection also retains shift clusters which express specific developments of language in the RSC. Still, there are two aspects which could be improved. First, the selection could combine and prioritize factors. For example, clusters of long shifts which at the same time point into very similar directions might be of special interest and probably more interesting than clusters which merely shift a lot. Second, the largest part of conceptual shifts, namely those clusters which are not statistically significant in terms of length, spread, size etc., is ignored, even though these measurements are not validated to indicate whether a cluster of conceptual shifts is 'interesting' or not. Here, different measures or a different criterion for the selection of shift clusters might be better suited than the current method.

Indeed, there is no simple definition of what makes a (systematic) conceptual shift 'interesting'; statistical significance is merely to be taken as a proxy. In general, the proposed method for unsupervised shift detection is not designed to yield finalized results: while no human intuition is involved in the alignment or the detection of systematic

[43] Dubossarsky et al. (2015) are interested in the correlation between a concept's distance to its cluster's center and the concept's rate of change.

shifts, the final step, namely the interpretation of the shift clusters, is left to human judgement.

6.4 Findings and Evaluation

I discuss most of the findings of the experiments directly in Chapters 5.3 and 5.4. To recapitulate: similar words tend to change in similar ways; numbers and function words, but also measurement-related content words, change very little; more change is observable for $\langle 1740, 1770\rangle$ than for $\langle 1860, 1890\rangle$; in $\langle 1740, 1770\rangle$, botanic studies are on the rise; in $\langle 1860, 1890\rangle$, the specialized nature of anatomy shows in the form of conceptual establishment; *-ing*-verbs, whith respect to diachronic shift, tend to form large clusters; discourse-related words change differently from technical terms. The following points address the less straightforward and the controversial findings as well as a remark on the visualization via t-SNE.

The first point is concerned with statistics of the UnsupMono experiment; more precisely, the measurements of shift cluster length (maximum, mean, median length, and the standard deviation). In Figure 16, both iterations, once each on $\langle 1740, 1770\rangle$ and $\langle 1860, 1890\rangle$, show unusual differences of statistics of the baseline to those of the actual performance. The baseline was constructed on the exact same set of shift vectors as the actual experiments. Thus, it is surprising that, for example, almost no cluster of the $\langle 1740, 1770\rangle$ baseline comprises a shift vector with length between 1.1 and 1.3, while the experiments find that *the majority* of clusters contain a vector of this length.

A second controversy in the statistics relates to the differences between training methods of embeddings (incremental vs. individual). Here I compare the amount of shift measured in UnsupMono to that from DisTech (both for $\langle 1860, 1890\rangle$). Figure 17c reports almost no differences in mean length of clusters, but Figure 24 shows a clear difference when measured by pair-wise cosine distance (i.e., $d_{cos}(\mathbf{P}x_u, y_v)$ for $(u, v) \in T$).

One reason for this is that clearly, the two measurements are not directly comparable to each other; one measures averages of vector length, the other measures individual angles. Recall also that the shift vectors need to be scaled in order to reduce frequency effects which arise when a word occurs differently often in the two sub-corpora of a space pair. It is likely that the two measurements react differently to the projection, and it is possible that the vector scaling makes the size estimation via shift vector length less reliable than cosine distance. Here, a more faceted approach to measuring vectors, involving both measures of length and of angles, could help to make better use of the full potential of the proposed method.

Another explanation lies in the different contents of $T_{UnsupMono}$ and $T_{DisTech}$. One experiment is performed for a large, unspecific set of words ($|T_{UnsupMono}| = 10233$), while the other is comprises a smaller set of specific words ($|T_{DisTech}| = 1489$). It is possible that the differences which show for the discourse words and technical terms in DisTech also occur in UnsupMono, but are overshadowed by the large number of additional clusters, which themselves might behave rather similarly across training methods. If this is true, then a differentiation of groups of concepts like the one in DisTech could help to make more precise judgements about which kinds of concepts experience more change than others. Such distinctions may be syntactic (as in DisTech) or topical (e.g., *botany/anatomy/astronomy/chemistry*).

Lastly, the visualization via t-SNE used throughout Chapters 5.3.5 and 5.4.2 shows to be very much dependent on the number of points to be visualized. This is useful when single clusters of shifts are to be investigated: t-SNE utilizes as much of the 2D space as possible and brings out the differences in the direction of shift between the members of the cluster (cf. Figure 18, §5.3.5). The visualization is also useful when comparing two clusters of similar size, e.g. the *anterior*- and the *inherent*-clusters (Figure 26, §5.4.2).

However, with an increasing number of points to visualize, t-SNE naturally cannot depict as many of the nuances as for single clusters or pairs thereof. Figure 31 in the Appendix illustrates this issue with four relatively tight clusters: when visualized jointly, the systematic changes are clearly visible, but this collective movement shows much less when the clusters are visualized individually.

The visualization of diachronic shifts via t-SNE is not new; Hamilton et al. (2016b) and Bizzoni et al. (2019a), for example, visualize conceptual change by projecting both the concepts of interest together with their nearest neighbors onto the 2D space. The reason for applying t-SNE in a different manner here is complexity: while Hamilton et al. (2016b) and Bizzoni et al. (2019a) are interested in the shift of single concepts, the visualization in the context of systematic changes involves groups of concepts. As the expressiveness of t-SNE projections reduces with an increasing number of points, the joint visualization of a shift cluster and the nearest neighbors of all points involved (all start points, all end points) would likely lead to cluttered images with little expressiveness. This increased complexity, however, should not be seen as a limitation. It is rather to be seen as a challenge to improve the usability and reliability of t-SNE for the visualization of shift clusters.

These points underline that despite the various promising findings, it is a challenging task to not only quantify, but also to characterize conceptual change. It is important to have a realistic picture of the potential and the limitations of methods like the one proposed in this thesis. A major challenge with respect to this is to distinguish artefacts of vector space processing (projection, subtraction, reduction etc.) from actual conceptual shifts.

6.5 Future Work

An important future direction is validation. It is clear that the experiments carried out in chapters 4 and 5 are only some of the first steps towards embedding-centered research on language change in the RSC. The results, while partly promising, do not have a strong statistical foundation; they are at times carried out on only a small number of samples. In order to validate the findings as well as the methods, the experiments should be replicated. This is especially necessary for the experiments on shift detection (Chapter 5), which should be replicated both on the $\langle 1740, 1770\rangle$ and $\langle 1860, 1860\rangle$ space pairs as well as other combinations of embedding spaces. Also, the proposed method of shift detection is only evaluated manually. A more statistically sound method for evaluation, e.g. the use of synthetic data (cf. Rosenfeld and Erk, 2018), may help to untangle the effects of embedding models, alignment, and of the processing of shift vectors on the results.

In addition to this, I provided several suggestions for improvement of the proposed method of shift detection and its evaluation in the previous sections (e.g., the use of clustering for noise filtering; different clustering of shift vectors; better label approximation via centroids; more sensitive quantification of shifts; more elaborate filtering of interesting shifts; more reliable visualization). Future applications may refine some of its components and make it more reliable and expressive.

Apart from embedding-based research on the development of scientific English, the proposed method could be applied to comparative studies of conceptual spaces of different language resources and/or along different axes. For example, Cafagna et al. (2019) show conceptual differences of words between the language varieties used in two newspapers; the axis of distinction is the political dimension (left vs. right). On another note, the proposed method might be applied to compare how the world is conceptualized by humans across different cultures, and whether differences in conceptualization are captured or even amplified by embedding models. Systematic shifts, potentially expressing human or machine bias, might be uncovered in this way.

7 Conclusion

In this thesis, I investigated diachronic conceptual change with methods from Distributional Semantics. The investigations were carried out on word embeddings of the Royal Society Corpus, a collection of English texts from the scientific domain dating from 1665 to 1929 (Fischer et al., 2020). Following the intuition that diachronic conceptual changes are often very subtle, I approached the topic assuming bilinguality; that is, not taking for granted that the vocabulary is mostly persistent over time and instead treating any two points in time as distinct varieties of language. With respect to the detection of conceptual shifts, I did not make assumptions about which changes should or should not be observed.

The starting point for comparative research with word embeddings is the knowledge about how reliable and expressive these vector representations are. To this end, I evaluated the embedding models (taken from Bizzoni and Teich, 2019) in a semantic similarity test (§3.2.3). This confirmed their adequacy for the task at hand, showing at the same time that not all training methods are well-suited for the diachronic scenario which often cannot provide large amounts of text data.

In order to make the word embeddings comparable to each other, I used Optimal Transport with Gromov-Wasserstein Distance (GWOT, Alvarez-Melis and Jaakkola, 2018), a method that relates points across embedding spaces purely by how similar their spatial roles within their respective spaces are (§4.1). GWOT had been shown to be effective in a true bilingual scenario. I thoroughly inspected GWOT and showed that it can be adapted to the diachronic scenario (§4.3), but that it does not work reliably for the embedding spaces at hand. This is problematic if working under the assumption of bilinguality; at the same time, the investigations also show that it is not necessary to make this strong assumption of bilingualiy in the diachronic context (§4.3.7, §5.3.1). GWOT still proved to be useful in the (conventional monolingual) diachronic scenario to align embedding spaces while preserving conceptual shifts. In addition to this, GWOT can generally inform about the divergence of embedding spaces, which is useful to survey the overall development of language.

For the detection of systematic conceptual shifts, I proposed a novel method which is based on clustering and nearest-neighbor search (§5.2). Crucially, it operates directly on the shifts, which, being high-dimensional vectors themselves, allows to observe directions of shifts in addition to magnitudes. The proposed method then approximates the meaning of these directions with nearest-neighbor search in order to make them human-interpretable. In the conducted experiments (§5.3, §5.4), the proposed method showed to be an effective tool for exploratory studies of conceptual change. Most results are in line with previous research. In addition to this, the proposed method yielded new insights about language change in the RSC. For example, from the 1860s to the 1890s, discourse-related words tend to shift more than technical terms (§5.5). The proposed method also uncovered general dynamics such as conceptual establishment – the amplification of a concept's core components of meaning. Many groups of concepts with systematic shifts experience conceptual establishment, irrespective of how much they change. It is likely the default dynamic of change, present in concepts which become conceptually stronger in their core semantics rather than taking on new or different meanings. The limitations of the proposed method which showed during the inspection of the results can be attributed to weaknesses in specific processing steps. These issues can be addressed with modifications and extensions, which are straightforward to implement, given that the proposed method is modular and follows simple principles of Distributional Semantics.

In the introduction, I stated five main questions to be answered by the thesis. I found and gave answers to all of them in detail in chapters 4 and 5. In short, the answers to these questions are as follows:

1. **Is it useful to frame the detection of diachronic conceptual change as bilingual task?** There is no consistent benefit in assuming that one language at two points in time should be treated as two distinct varieties. This was shown for a temporal distance of 30 years. It is possible that the assumption of bilinguality is more useful to larger intervals of time.

2. **How well does the method for unsupervised alignment work with respect to the RSC; is it reliable and feasible?** The results are mixed; there is no clear tendency to why and when the alignment via GWOT works and when it does not. Thorough investigations point to algorithmic issues and idiosyncrasies of the underlying data. That being said, GWOT does yield good results reliably in certain cases.

3. **Does the proposed method for the unsupervised detection of shifts work; is it sensible and expressive?** Yes, the proposed method works; it

detects shifts and allows to inspect them from the qualitative aspect in an intuitive way. In terms of sensibility and expressiveness, the results are promising. Modifications to the method are straightforward to implement because of its simplicity.

4. **Do the insights from the proposed method relate to previous findings?** Yes, most of the findings confirm previous research on the development of the language in the RSC.

5. **Can the proposed method uncover new (general and/or specific) dynamics of language change within the RSC?** Yes, it expresses the tendency of similar words to change in similar ways. The proposed method also confirms the intuition that most concepts continue to become established in their main meaning(s) rather than becoming more faceted. This tendency to *conceptual establishment* is present in the scientific language of the RSC, which developed under the need for precise and conceptually stable expressions for a clear form of communication. It remains to be shown whether other varieties of language show conceptual establishment to a similar degree.

The thesis aims to contribute to research in two ways; in addition to gaining new insights on the RSC and confirming previous findings, it aims to provide a basis of knowledge about the investigation of conceptual change via word embeddings to researchers from varying fields. To this end, I gave descriptions of the technical background and a variety of related research, outlining the possibilities and applications of the field (§2. The goal of providing entry points for further research is also reflected in the discussion of my findings (§6): I discussed controversial findings, identified points for improvement, and suggested alternatives and solutions to these points, drawing from the previously established background of applications.

Future work includes, but is not limited to: validation and improvement of the method; further application of the method for research on the RSC; and application of the method to different fields of comparative research. With the work presented in this thesis, I hope to have provided a solid basis for these future directions and to have inspired research on conceptual change via distributional models of semantics, irrespective of language or resource.

References

R. Alatrash, D. Schlechtweg, J. Kuhn, and S. Schulte im Walde. CCOHA: Clean Corpus of Historical American English. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6958–6966, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.859.

D. Alvarez-Melis and T. Jaakkola. Gromov-Wasserstein Alignment of Word Embedding Spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1214. URL https://www.aclweb.org/anthology/D18-1214.

D. Alvarez-Melis, S. Jegelka, and T. S. Jaakkola. Towards Optimal Transport with Global Invariances. *Proceedings of Machine Learning Research*, 89:1870–1879, Apr. 2019. URL http://proceedings.mlr.press/v89/alvarez-melis19a.html.

M. Artetxe, G. Labaka, and E. Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1042. URL https://www.aclweb.org/anthology/P17-1042.

M. Artetxe, G. Labaka, and E. Agirre. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

D. Atkinson. The "Philosophical Transactions of the Royal Society of London," 1675-1975: A Sociohistorical Discourse Analysis. *Language in Society*, 25(3):333–371, 1996. ISSN 0047-4045. URL https://www.jstor.org/stable/4168717. Publisher: Cambridge University Press.

M. Baroni and A. Lenci. Distributional Memory: A General Framework for Corpus-Based Semantics. *Computational Linguistics*, 36(4):673–721, Oct. 2010. ISSN 0891-2017. doi: 10.1162/coli_a_00016. URL https://doi.org/10.1162/coli_a_00016. Publisher: MIT Press.

M. Baroni, G. Dinu, and G. Kruszewski. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland, 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1023. URL http://aclweb.org/anthology/P14-1023.

Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, (3):1137–1155, Feb. 2003.

Y. Bizzoni and E. Teich. Analyzing variation in translation through neural semantic spaces. In *Proceedings of the 12th Workshop on Building and Using Comparable Corpora at RANLP*, Varna, Bulgaria, 2019.

Y. Bizzoni, S. Degaetano-Ortlieb, K. Menzel, P. Krielke, and E. Teich. Grammar and Meaning: Analysing the Topology of Diachronic Word Embeddings. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 175–185, Florence, Italy, Aug. 2019a. Association for Computational Linguistics. doi: 10.18653/v1/W19-4722. URL https://www.aclweb.org/anthology/W19-4722.

Y. Bizzoni, M. Mosbach, D. Klakow, and S. Degaetano-Ortlieb. Some steps towards the generation of diachronic WordNets. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 55–64, Turku, Finland, Sept. 2019b. Linköping University Electronic Press. URL https://www.aclweb.org/anthology/W19-6106.

Y. Bizzoni, S. Degaetano-Ortlieb, P. Fankhauser, and E. Teich. Linguistic Variation and Change in 250 years of English Scientific Writing: A Data-driven Approach. *Frontiers in Artificial Intelligence*, (3):73, Apr. 2020.

T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*, July 2020. URL http://arxiv.org/abs/2005.14165. arXiv: 2005.14165.

E. Bruni, N.-K. Tran, and M. Baroni. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47, 2014.

A. Budanitsky and G. Hirst. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. *Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics*, pages 29–34, June 2001.

J. A. Bullinaria and J. P. Levy. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3): 510–526, Aug. 2007. ISSN 1554-351X, 1554-3528. doi: 10.3758/BF03193020. URL http://link.springer.com/10.3758/BF03193020.

M. Cafagna, L. D. Mattei, and M. Nissim. Embeddings Shifts as Proxies for Different Word Use in Italian Newspapers. In *Proceedings of the Sixth Italian Conference on Computational Linguistics*, Bari, Italy, Nov. 2019. URL http://ceur-ws.org/Vol-2481/paper12.pdf.

J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, Apr. 1960. ISSN 0013-1644. doi: 10.1177/001316446002000104. URL https://doi.org/10.1177/001316446002000104. Publisher: SAGE Publications Inc.

A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou. Word Translation Without Parallel Data. *Proceedings of ICLR*, Jan. 2018. URL http://arxiv.org/abs/1710.04087. arXiv: 1710.04087.

M. Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2292–2300. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/4927-sinkhorn-distances-lightspeed-computation-of-optimal-transport.pdf.

M. Davies. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora*, 7(2):121–157, Nov. 2012. ISSN 1749-5032. doi: 10.3366/cor.2012.0024. URL https://www.euppublishing.com/doi/abs/10.3366/cor.2012.0024. Publisher: Edinburgh University Press.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990. ISSN 1097-4571. doi: 10.1002/(SICI)1097-4571(199009)41:6⟨391::AID-ASI1⟩3.0.CO;2-9. URL https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28199009%2941%3A6%3C391%3A%3AAID-ASI1%3E3.0.CO%3B2-9.

S. Degaetano-Ortlieb. Evaluative meaning in scientific writing : macro- and micro-analytic perspectives using data mining. 2015. doi: 10.22028/D291-23641. URL https://publikationen.sulb.uni-saarland.de/handle/20.500.11880/23697. Universität des Saarlandes.

S. Degaetano-Ortlieb and E. Teich. Information-based Modeling of Diachronic Linguistic Change: from Typicality to Productivity. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 165–173, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2121. URL https://www.aclweb.org/anthology/W16-2121.

S. Degaetano-Ortlieb and E. Teich. Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the Second Joint SIGHUM*

Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pages 22–33, Santa Fe, New Mexico, Aug. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-4503.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

H. Dubossarsky, Y. Tsvetkov, C. Dyer, and E. Grossman. A bottom up approach to category mapping and meaning change. In *Proceedings of the NetWordS Final Conference*, Pisa, 2015. URL http://ceur-ws.org/Vol-1347/paper14.pdf.

H. Dubossarsky, D. Weinshall, and E. Grossman. 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, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1118. URL https://www.aclweb.org/anthology/D17-1118.

S. Eger and A. Mehler. On the Linearity of Semantic Change: Investigating Meaning Variation via Dynamic Graph Models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 52–58, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2009. URL https://www.aclweb.org/anthology/P16-2009.

M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pages 226–231, Portland, OR, 1996. AIII Press. URL https://aaai.org/Library/KDD/1996/kdd96-037.php.

P. Fankhauser and M. Kupietz. Visualizing Language Change in a Corpus of Contemporary German. In *Proceedings of the Corpus Linguistics International Conference*, Birmingham, UK, 2017.

M. Faruqui and C. Dyer. Improving Vector Space Word Representations Using Multilingual Correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden, Apr. 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-1049. URL https://www.aclweb.org/anthology/E14-1049.

J. R. Firth. *Papers in Linguistics 1934-1951: Repr.* Oxford University Press, 1961.

S. Fischer, J. Knappen, K. Menzel, and E. Teich. The Royal Society Corpus 6.0: Providing 300+ Years of Scientific Writing for Humanistic Study. In *Proceedings of*

The 12th Language Resources and Evaluation Conference, pages 794–802, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.99.

R. Flamary and N. Courty. POT Python Optimal Transport library, 2017. URL https://github.com/rflamary/POT.

B. J. Frey and D. Dueck. Clustering by Passing Messages Between Data Points. *Science*, 315(5814):972–976, Feb. 2007. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1136800. URL https://science.sciencemag.org/content/315/5814/972.

G. Glavaš and I. Vulić. Non-Linear Instance-Based Cross-Lingual Mapping for Non-Isomorphic Embedding Spaces. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7548–7555, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.675.

Y. Goldberg and O. Levy. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *CoRR*, abs/1402.3722, 2014. URL http://arxiv.org/abs/1402.3722. _eprint: 1402.3722.

H. Gonen, G. Jawahar, D. Seddah, and Y. Goldberg. Simple, Interpretable and Stable Method for Detecting Words with Usage Change across Corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.51.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

E. Grave, A. Joulin, and Q. Berthet. Unsupervised Alignment of Embeddings with Wasserstein Procrustes. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1880 – 1890, Okinawa, Japan, 2019. PMLR. URL http://proceedings.mlr.press/v89/grave19a.html.

S. Haker, L. Zhu, and A. Tannenbaum. Optimal Mass Transport for Registration and Warping. *International Journal of Computer Vision*, 60(3):225–240, 2004. doi: https://doi.org/10.1023/B:VISI.0000036836.66311.97. Kluwer Academic Publishers.

W. L. Hamilton, J. Leskovec, and D. Jurafsky. Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas, Nov. 2016a. Association for Computational Linguistics. doi: 10.18653/v1/D16-1229. URL https://www.aclweb.org/anthology/D16-1229.

W. L. Hamilton, J. Leskovec, and D. Jurafsky. 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, Berlin, Germany, Aug. 2016b. Association for Computational Linguistics. doi: 10.18653/v1/P16-1141. URL https://www.aclweb.org/anthology/P16-1141.

D. Helbing. Derivation of Non-local Macroscopic Traffic Equations and Consistent Traffic Pressures from Microscopic Car-Following Models. In *Modelling and Optimisation of Flows on Networks: Cetraro, Italy 2009, Editors: Benedetto Piccoli, Michel Rascle*, Lecture Notes in Mathematics, pages 247–269. Springer, Berlin, Heidelberg, 2013. ISBN 978-3-642-32160-3. URL https://doi.org/10.1007/978-3-642-32160-3_3.

P. Jawanpuria, M. Meghwanshi, and B. Mishra. Geometry-aware domain adaptation for unsupervised alignment of word embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3052–3058, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.276.

T. Kekeç, L. van der Maaten, and D. M. J. Tax. PAWE: Polysemy Aware Word Embeddings. In *Proceedings of the 2nd International Conference on Information System and Data Mining*, ICISDM '18, pages 7–13, New York, NY, USA, Apr. 2018. Association for Computing Machinery. ISBN 978-1-4503-6354-9. doi: 10.1145/3206098.3206101. URL https://doi.org/10.1145/3206098.3206101.

H. Kermes, S. Degaetano-Ortlieb, A. Khamis, J. Knappen, and E. Teich. The Royal Society Corpus: From Uncharted Data to Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1928–1931, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L16-1305.

Y. Kim, Y.-I. Chiu, K. Hanaki, D. Hegde, and S. Petrov. Temporal Analysis of Language through Neural Language Models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-2517. URL https://www.aclweb.org/anthology/W14-2517.

V. Klema and A. Laub. The singular value decomposition: Its computation and some applications. *IEEE Transactions on Automatic Control*, 25(2):164–176, Apr. 1980. doi: 10.1109/TAC.1980.1102314.

J. Knappen, S. Fischer, H. Kermes, E. Teich, and P. Fankhauser. The Making of the Royal Society Corpus. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 7–11, Gothenburg, May 2017. Linköping University Electronic Press. URL https://www.aclweb.org/anthology/W17-0503.

T. C. Koopmans. Optimum Utilization of the Transportation System. *Econometrica*, 17:136–146, 1949. ISSN 0012-9682. doi: 10.2307/1907301. URL https://www.jstor.org/stable/1907301. Wiley, Econometric Society.

V. Kulkarni, R. Al-Rfou, B. Perozzi, and S. Skiena. Statistically Significant Detection of Linguistic Change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 625–635, Florence, Italy, May 2015. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-3469-3. doi: 10.1145/2736277.2741627. URL https://doi.org/10.1145/2736277.2741627.

A. Kutuzov, E. Velldal, and L. Øvrelid. Tracing armed conflicts with diachronic word embedding models. In *Proceedings of the Events and Stories in the News Workshop*, pages 31–36, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2705. URL https://www.aclweb.org/anthology/W17-2705.

A. Kutuzov, L. Øvrelid, T. Szymanski, and E. Velldal. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1117.

T. K. Landauer and S. T. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997. ISSN 1939-1471(Electronic),0033-295X(Print). doi: 10.1037/0033-295X.104.2.211. American Psychological Association.

J. R. Landis and G. G. Koch. An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers. *Biometrics*, 33(2):363–374, 1977. ISSN 0006-341X. doi: 10.2307/2529786. URL https://www.jstor.org/stable/2529786. Wiley, International Biometric Society.

A. Lazaridou, E. Bruni, and M. Baroni. Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1414, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1132. URL https://www.aclweb.org/anthology/P14-1132.

A. Lazaridou, G. Dinu, and M. Baroni. Hubness and Pollution: Delving into Cross-Space Mapping for Zero-Shot Learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 270–280, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1027. URL https://www.aclweb.org/anthology/P15-1027.

O. Levy and Y. Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems 27*, pages 2177–2185. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5477-neural-word-embedding-as-implicit-matrix-factorization.pdf.

O. Levy, Y. Goldberg, and I. Dagan. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational*

Linguistics, 3:211–225, 2015. doi: 10.1162/tacl_a_00134. URL https://www.aclweb.org/anthology/Q15-1016.

W. Ling, C. Dyer, A. W. Black, and I. Trancoso. Two/Too Simple Adaptations of Word2Vec for Syntax Problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1299–1304, Denver, Colorado, May 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1142. URL https://www.aclweb.org/anthology/N15-1142.

Y. Liu, Z. Liu, T.-S. Chua, and M. Sun. Topical word embeddings. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2418–2424, Austin, Texas, Jan. 2015. AAAI Press. ISBN 978-0-262-51129-2.

A. Lu, W. Wang, M. Bansal, K. Gimpel, and K. Livescu. Deep Multilingual Correlation for Improved Word Embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 250–256, Denver, Colorado, May 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1028. URL https://www.aclweb.org/anthology/N15-1028.

M. Mancini, J. Camacho-Collados, I. Iacobacci, and R. Navigli. Embedding Words and Senses Together via Joint Knowledge-Enhanced Training. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 100–111, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics. doi: 10.18653/v1/K17-1012. URL https://www.aclweb.org/anthology/K17-1012.

M. Martinc, P. Kralj Novak, and S. Pollak. Leveraging Contextual Embeddings for Detecting Diachronic Semantic Shift. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.592.

A. V. Miceli Barone. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 121–126, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-1614. URL https://www.aclweb.org/anthology/W16-1614.

J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, T. G. B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014):176–182, Jan. 2011. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1199644. URL https://science.sciencemag.org/content/331/6014/176. American Association for the Advancement of Science.

T. Mikolov, Q. V. Le, and I. Sutskever. Exploiting Similarities among Languages for Machine Translation. *arXiv:1309.4168 [cs]*, Sept. 2013a. URL http://arxiv.org/abs/1309.4168. arXiv: 1309.4168.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013b. URL http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June 2013c. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N13-1090.

L. Moessner. The influence of the Royal Society on 17th-century scientific writing. *ICAME journal*, 33:65–87, 2009. URL http://icame.uib.no/ij33/ij33-65-88.pdf.

F. Mémoli. Gromov–Wasserstein Distances and the Metric Approach to Object Matching. *Foundations of Computational Mathematics*, 11(4):417–487, Aug. 2011. ISSN 1615-3375, 1615-3383. doi: 10.1007/s10208-011-9093-5. URL http://link.springer.com/10.1007/s10208-011-9093-5.

N. Nakashole. NORMA: Neighborhood Sensitive Maps for Multilingual Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 512–522, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1047. URL https://www.aclweb.org/anthology/D18-1047.

K. A. Nguyen, S. Schulte im Walde, and N. T. Vu. Integrating Distributional Lexical Contrast into Word Embeddings for Antonym-Synonym Distinction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 454–459, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2074. URL https://www.aclweb.org/anthology/P16-2074.

M. Ono, M. Miwa, and Y. Sasaki. Word Embedding-based Antonym Detection using Thesauri and Distributional Information. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 984–989, Denver, Colorado, May 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1100. URL https://www.aclweb.org/anthology/N15-1100.

B. Patra, J. R. A. Moniz, S. Garg, M. R. Gormley, and G. Neubig. Bilingual Lexicon Induction with Semi-supervision in Non-Isometric Embedding Spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1018. URL https://www.aclweb.org/anthology/P19-1018.

J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://www.aclweb.org/anthology/N18-1202.

G. Peyré and M. Cuturi. Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000073. URL https://www.nowpublishers.com/article/Details/MAL-073.

G. Peyré, M. Cuturi, and J. Solomon. Gromov-Wasserstein Averaging of Kernel and Distance Matrices. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 2664–2672, New York, NY, USA, 2016. JMLR: W&CP.

M. Radovanović, A. Nanopoulos, and M. Ivanović. Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data. *The Journal of Machine Learning Research*, 11:2487–2531, Dec. 2010. ISSN 1532-4435.

P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA, Aug. 1995. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-363-9.

D. L. T. Rohde, L. M. Gonnerman, and D. C. Plaut. An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence. *Communications of the ACM*, 8((627-633)):116, 2006.

A. Rosenfeld and K. Erk. Deep Neural Models of Semantic Shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 474–484, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1044. URL https://www.aclweb.org/anthology/N18-1044.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *EMC^2 5th Edition Co-ocated with*

NeurIPS, Feb. 2020. URL https://www.emc2-ai.org/assets/docs/neurips-19/emc2-neurips19-paper-33.pdf.

D. Schlechtweg, A. Hätty, M. Del Tredici, and S. Schulte im Walde. A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1072. URL https://www.aclweb.org/anthology/P19-1072.

E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Transactions on Database Systems*, 42(3):1–21, Aug. 2017. ISSN 0362-5915, 1557-4644. doi: 10.1145/3068335. URL https://dl.acm.org/doi/10.1145/3068335.

P. H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, Mar. 1966. ISSN 1860-0980. doi: 10.1007/BF02289451. URL https://doi.org/10.1007/BF02289451.

D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley and Sons, Hoboken, NJ, 2 edition, 2015. ISBN 978-0-471-69755-8. URL http://adsabs.harvard.edu/abs/2015mdet.book.....S.

D. Sculley. Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 1177–1178, New York, NY, USA, Apr. 2010. Association for Computing Machinery. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772862. URL https://doi.org/10.1145/1772690.1772862.

Y. Shigeto, I. Suzuki, K. Hara, M. Shimbo, and Y. Matsumoto. Ridge Regression, Hubness, and Zero-Shot Learning. *Proceedings of the 2015th European Conference on Machine Learning and Knowledge Discovery in Databases*, 1:135–151, Sept. 2015. URL https://doi.org/10.1007/978-3-319-23528-8_9.

S. L. Smith, D. H. P. Turban, S. Hamblin, and N. Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. Toulon, France, Feb. 2017. URL https://openreview.net/pdf?id=r1Aab85gg.

N. G. Stern. *Swift, swiftly and their synonyms. A contribution to semantic analysis and theory.* PhD thesis, University of Göteborg, 1921. URL https://gupea.ub.gu.se/handle/2077/14287.

A. Søgaard, S. Ruder, and I. Vulić. On the Limitations of Unsupervised Bilingual Dictionary Induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1072. URL https://www.aclweb.org/anthology/P18-1072.

A. Søgaard, I. Vulić, S. Ruder, and M. Faruqui. Cross-Lingual Word Embeddings. *Synthesis Lectures on Human Language Technologies*, 12(2):1–132, June 2019. ISSN 1947-4040, 1947-4059. doi: 10.2200/S00920ED2V01Y201904HLT042. URL https://www.morganclaypool.com/doi/10.2200/S00920ED2V01Y201904HLT042.

N. Tahmasebi, L. Borin, and A. Jatowt. Survey of Computational Approaches to Lexical Semantic Change. *arXiv:1811.06278 [cs]*, Nov. 2018. URL http://arxiv.org/abs/1811.06278. arXiv: 1811.06278 version: 1.

E. C. Traugott. Semantic Change. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press, Mar. 2017. ISBN 978-0-19-938465-5. doi: 10.1093/acrefore/9780199384655.013.323. URL http://linguistics.oxfordre.com/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-323.

P. D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010. URL https://www.aaai.org/Papers/JAIR/Vol37/JAIR-3705.pdf.

S. Upadhyay, M. Faruqui, C. Dyer, and D. Roth. Cross-lingual Models of Word Embeddings: An Empirical Comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1661–1670, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1157. URL https://www.aclweb.org/anthology/P16-1157.

L. van der Maaten and G. Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, Nov. 2008. URL https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?source=post_page---------------------------.

I. Vulić and A. Korhonen. On the Role of Seed Lexicons in Learning Bilingual Word Embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 247–257, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1024. URL http://aclweb.org/anthology/P16-1024.

I. Vulić, G. Glavaš, R. Reichart, and A. Korhonen. Do We Really Need Fully Unsupervised Cross-Lingual Embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4407–4418, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1449. URL https://www.aclweb.org/anthology/D19-1449.

L. Wendlandt, J. K. Kummerfeld, and R. Mihalcea. Factors Influencing the Surprising Instability of Word Embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2092–2102, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1190. URL https://www.aclweb.org/anthology/N18-1190.

Y. Xu and C. Kemp. A Computational Evaluation of Two Laws of Semantic Change. page 6, 2015. URL https://www.cs.toronto.edu/~yangxu/xu_kemp_2015_parallelchange.pdf.

Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in neural information processing systems*, page 11, 2019.

Z. Yao, Y. Sun, W. Ding, N. Rao, and H. Xiong. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the eleventh acm international conference on web search and data mining*, pages 673–681, 2018.

M. Zhang, Y. Liu, H. Luan, and M. Sun. Adversarial Training for Unsupervised Bilingual Lexicon Induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1179. URL https://www.aclweb.org/anthology/P17-1179.

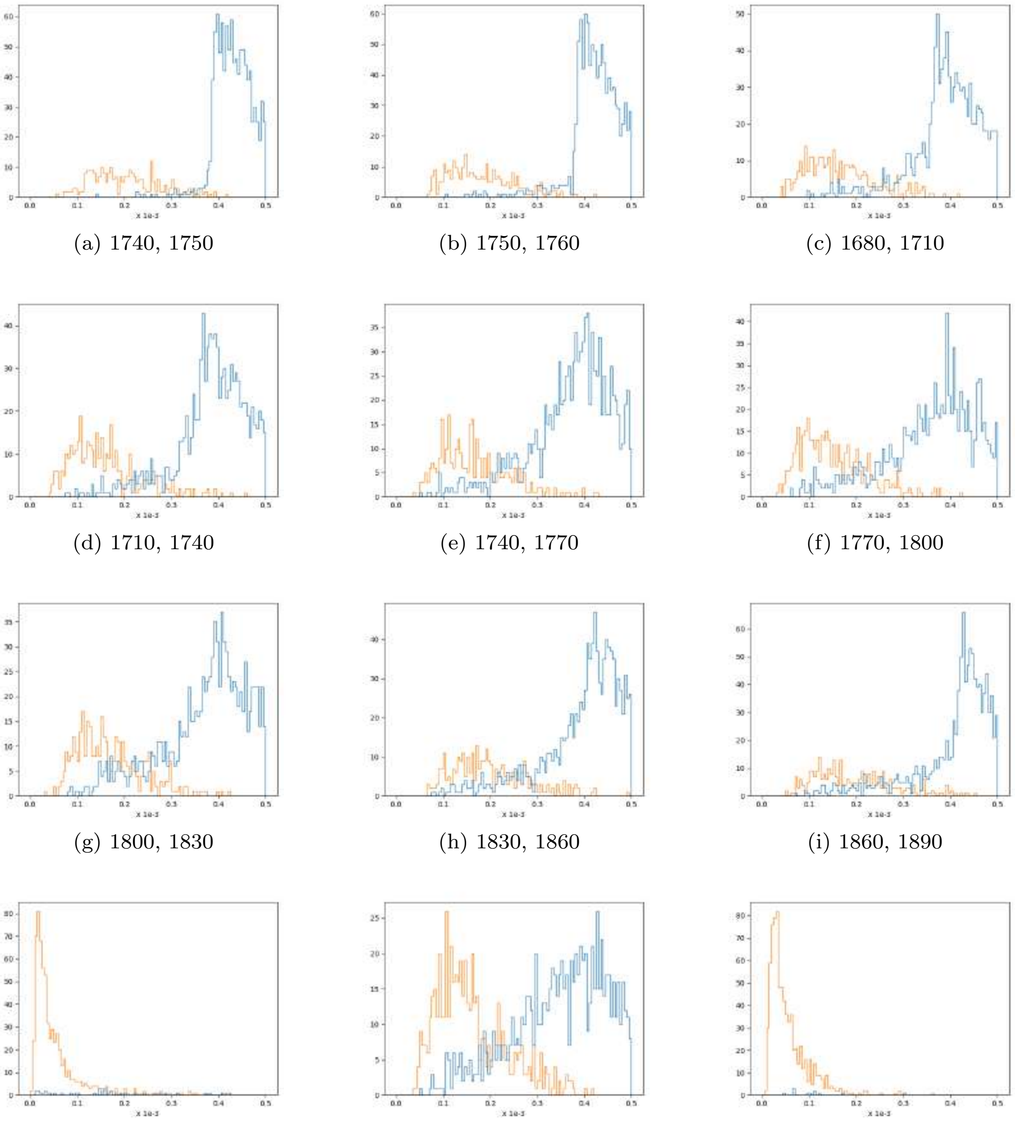
Y. Zhang, D. M. Gaddy, R. Barzilay, and T. S. Jaakkola. Ten pairs to tag - Multilingual POS tagging via coarse mapping between embeddings. *MIT Web Domain*, June 2016. URL https://dspace.mit.edu/handle/1721.1/110739. Accepted: 2017-07-17T18:13:57Z ISBN: 9781941643914 Publisher: Association for Computational Linguistics.

G. K. Zipf. *Selected studies of the principle of relative frequency in language*. Harvard university press, 1932.

Appendix

size	PC	per 1K	cluster	per 1K
1K	1:30	1:30	0:50	0:50
2K	4:50	2:25	3:20	1:40
3K	14:45	4:55	12:36	4:12
5K	40:45	8:09	33:40	6:44
10K	—	—	196:28	19:39
20K	—	—	1278:20	63:54

Table 16: Times taken for GWOT on different sizes of sub-spaces, carried out on a PC and a CPU cluster (cf. Chapter 3.3 for details). Maximum number of iterations for optimization: 300. Time in minutes.



(a) 1740, 1750

(b) 1750, 1760

(c) 1680, 1710

(d) 1710, 1740

(e) 1740, 1770

(f) 1770, 1800

(g) 1800, 1830

(h) 1830, 1860

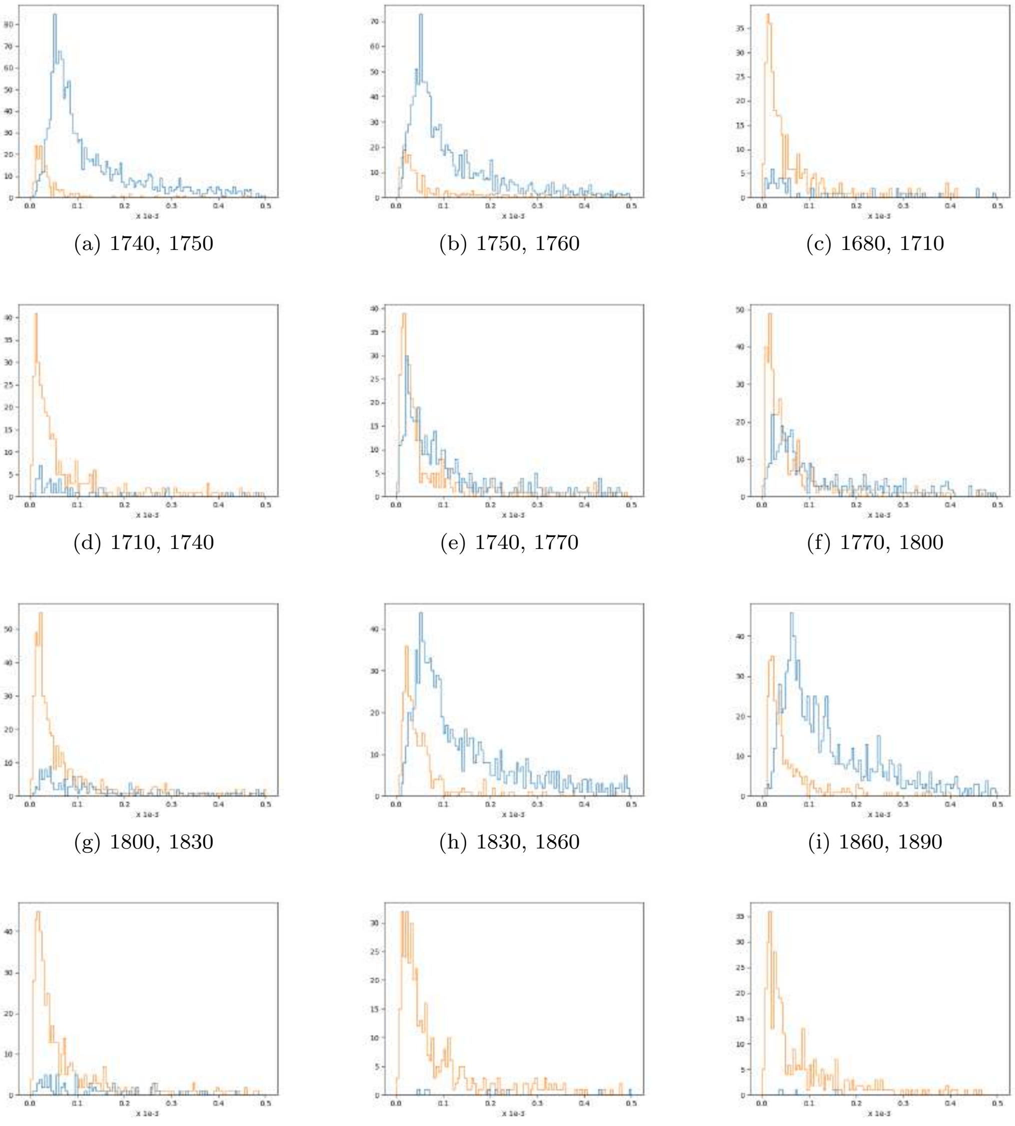
(i) 1860, 1890

(j) 1700, 1800

(k) 1800, 1900

(l) 1700, 1900

Figure 27: Distributions of string-matching mutual translation pairs by translation likelihood (range: $[0, 0.5e^{-3}]$). Optimal coupling computed for the 2K most frequent words in a pair's spaces (not necessarily mutual) with *log-flattened* relative frequencies provided for optimization. Matches in blue, mismatches in orange. Note that the range was clipped at $0.5e^{-3}$ for better comparability with other figures while scores reach values of up to $1.0e^{-3}$.



(a) 1740, 1750

(b) 1750, 1760

(c) 1680, 1710

(d) 1710, 1740

(e) 1740, 1770

(f) 1770, 1800

(g) 1800, 1830

(h) 1830, 1860

(i) 1860, 1890

(j) 1700, 1800

(k) 1800, 1900

(l) 1700, 1900

Figure 28: Distributions of string-matching mutual translation pairs by translation likelihood (range: $[0, 0.5e^{-3}]$). Optimal coupling computed for the 2K most frequent words in a pair's spaces (not necessarily mutual) with *raw* relative frequencies provided for optimization. Matches in blue, mismatches in orange.

p, q Top10-entropy Top10-mean Original

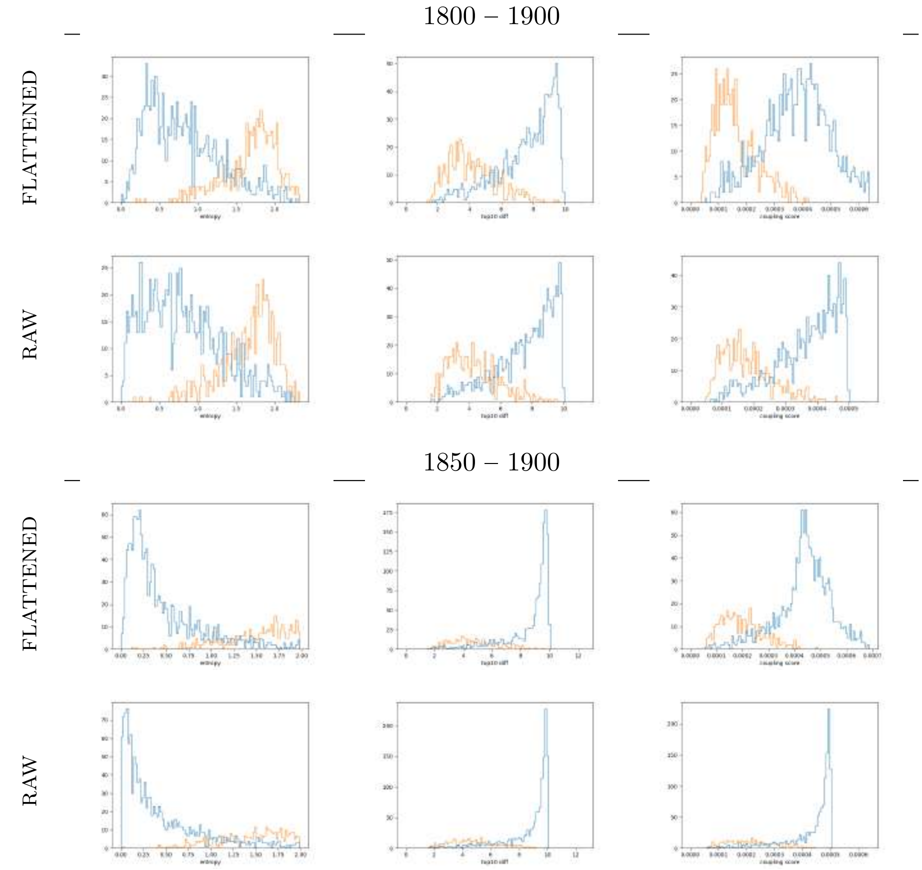
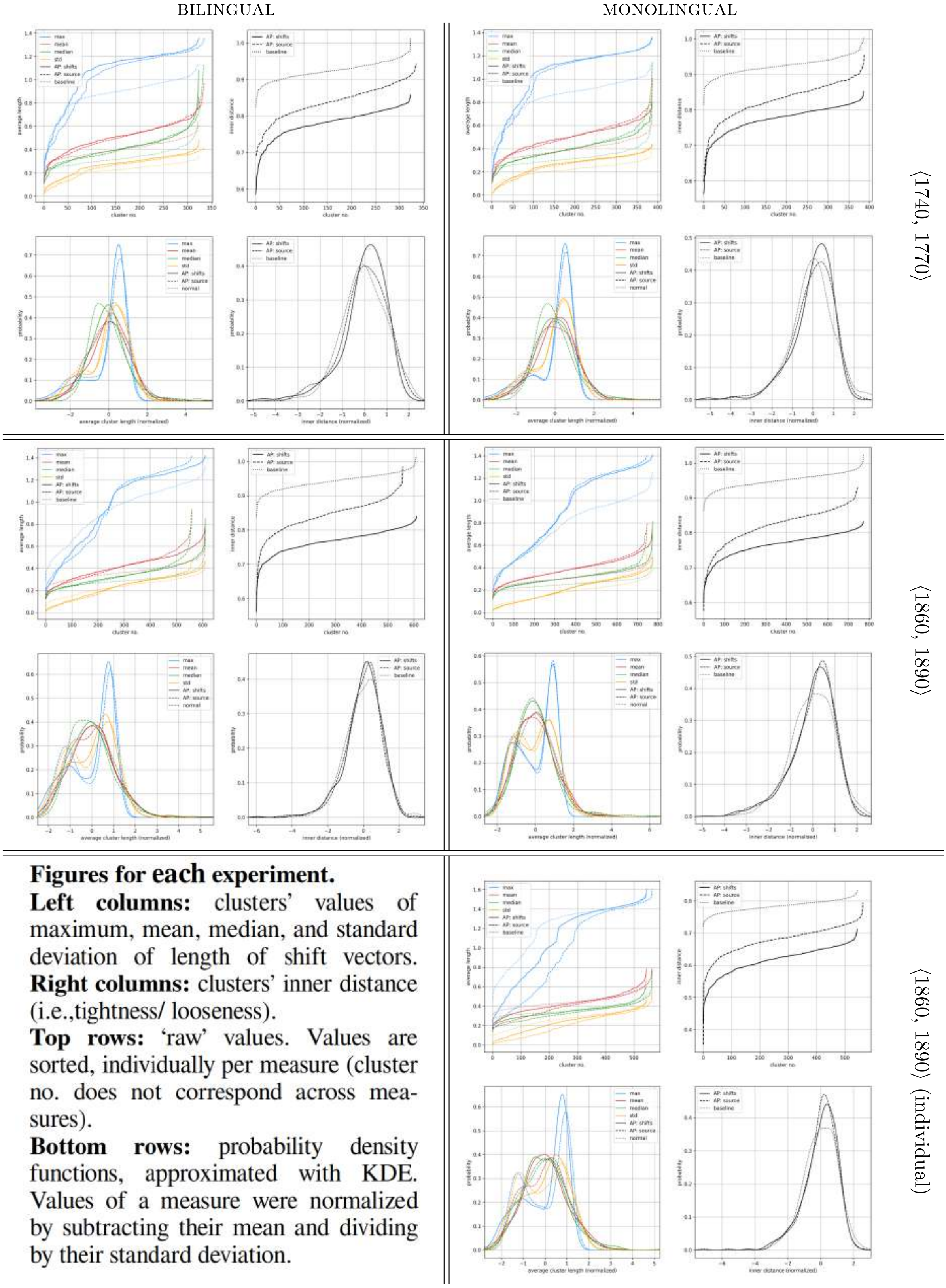


Figure 29: Distribution of translation pairs by different scoring techniques. Couplings optimized for k=2K sub-spaces, once initialized with uniform p/q and once with frequency information. Ranges of x-axes calculated from the scores: $[0, \mu + 2\sigma]$ per set of scores. Matches in blue, mismatches in orange. *Top-10-entropy* $H_{10}(\Gamma_{ij})$: arithmetic mean of the entropy values of the 10 highest scores in row i and column j. *Top-10-mean* $M_{10}(\Gamma_{ij})$: analogous to H_{10}; arithmetic mean instead of entropy.

Mutual translation pairs are defined as those word pairs which are the most likely translation of each other, i.e. those $\langle u_i, v_j \rangle$ for which Γ_{ij} is the maximum value in row i and column j. The intuition behind re-scoring translation pairs is that taking the maximum does not express entirely how confident a translation is: on one side, equally likely secondary translations are not taken into account; on the other side, a translation pair with a clearly dominant coupling score are not treated as “more confident” translations than uncertain ones. I investigated several re-scoring techniques that aim to enable a clear partition of good translation pairs from incorrect ones. The techniques were applied to couplings optimized on two k=2000 sub-spaces of ⟨1800,1900⟩ and ⟨1850,1900⟩, once with and once without frequency information for initialization. I evaluated with the same string-match heuristic as before, assuming that false positives (i.e, translation errors) are mostly found among the mismatching pairs.

Out of several proposed methods, the above presents *Entropy* and *Distance to Mean.* Both measures operate only on the 10 highest scores; lower scores are negligible. From the above, there is no clear positive or negative effect observable with respect to the partitioning of string-matching and mismatching translation pairs. Further and more thorough investigations might be needed to establish a reliable re-scoring method. Furthermore, the match-mismatch approach to evaluation and visualization of this experiment might not be expressive enough.



Figures for each experiment.
Left columns: clusters' values of maximum, mean, median, and standard deviation of length of shift vectors.
Right columns: clusters' inner distance (i.e.,tightness/ looseness).
Top rows: 'raw' values. Values are sorted, individually per measure (cluster no. does not correspond across measures).
Bottom rows: probability density functions, approximated with KDE. Values of a measure were normalized by subtracting their mean and dividing by their standard deviation.

Figure 30: Length measures and inner distances of unsupervised experiments.

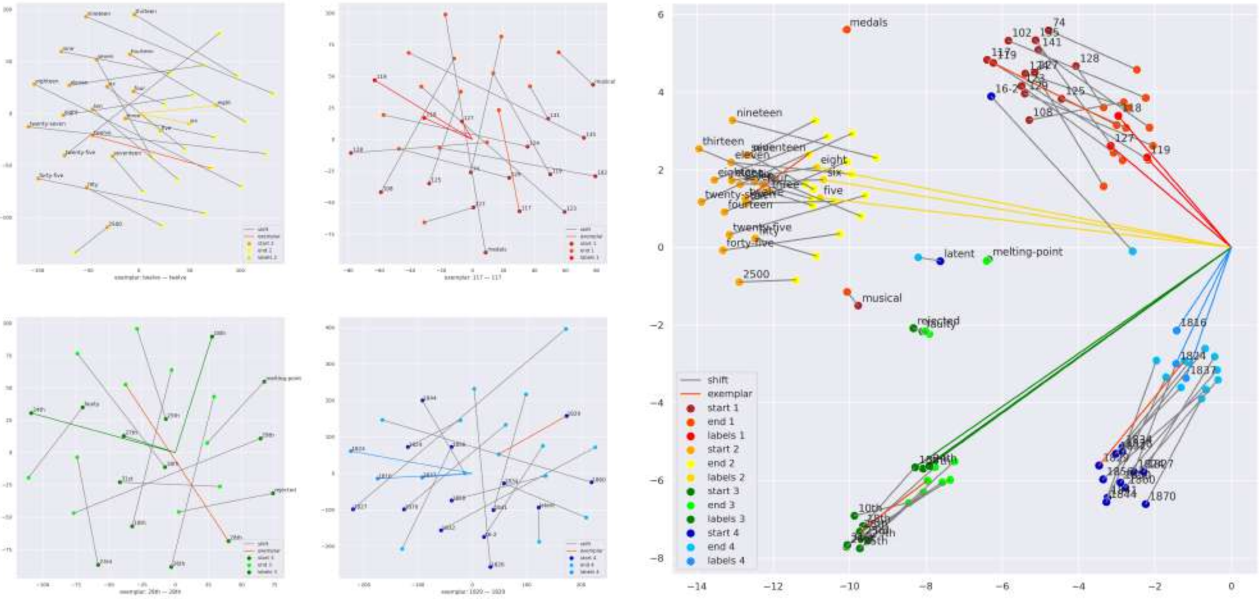


Figure 31: Four similar clusters from UnsupMono on ⟨1860, 1890⟩ (top right, counter-clockwise): *117* (red), *tweve* (yellow), *28th* (green), *1829* (blue).
Left: t-SNE for each individual cluster. *Right*: t-SNE for all clusters combined.

Note that the shifts on the right appear as directed towards the origin, rather than away from it (as would be expected, given that space expands over time, cf. Bizzoni et al., 2020, p.20). A likely explanation is the following trade-off of similarities: the label vectors $z \in Y$ (lines from the origin) are relatively similar to each other. Coming from the same model, they are likely more similar to each other than a $\mathbf{P}x$ of one cluster to a z of another cluster. As a trade-off, t-SNE decides to position all $\mathbf{P}x$ farther away from the origin than the y or the z. They are literally treated as *outsiders*, coming from a different space, and the large number of points forces t-SNE to depict these shifts in a partly counter-intuitive manner.