

**Universität Leipzig**  
**Fakultät für Mathematik und Informatik**  
**Institut für Informatik**

## **Bachelorarbeit**

Analyse des semantischen Wandels des Wortes „Nachhaltig“  
durch Methoden des Text Minings

Leipzig, 30. Januar 2022

eingereicht von  
Gabriel Huppenbauer

Betreuer der Abschlussarbeit:

Dr. Christian Kahmann  
Fakultät für Mathematik und Informatik  
Lehrstuhl für Automatische Sprachverarbeitung

---

### *Abstract - Deutsch*

Diese Arbeit beschäftigt sich mit der Untersuchung semantischen Wandels bestimmter Worte durch Methoden des Text Minings anhand des Wortes „nachhaltig“. Die Untersuchung des Wandels von Kontext mit den Methoden Word Embeddings, zeitliche Kookkurrenzen und Topic Modelle wird miteinander verglichen und die Frage gestellt, ob sich dadurch ein semantischer Wandel zwischen den Jahren 1995 und 2021 feststellen lässt. Die Methoden werden beschrieben und durchgeführt und die Ergebnisse miteinander verglichen, um semantischen Wandel nachzuvollziehen und sowohl das Potenzial, als auch die Limitierungen der Methoden zu finden. Dabei konnten sowohl bei Word Embeddings, als auch bei Topic Modellen und eingeschränkt bei Kookkurrenzen ein größerer Wandel des Kontexts in den 90er- und früheren 2000er-Jahren festgestellt werden. Durch qualitative Interpretation der Ergebnisse konnte aber nicht eindeutig festgelegt werden, ob dieser Kontext-Wandel auch einem semantischen Wandel entspricht.

### *Abstract - English*

This work examines semantic change of certain words through methods of text mining via the German word „nachhaltig“. The examination of context change through the methods word embeddings, temporal cooccurrences and topic models is being compared and the question asked, if it's possible to conclude semantic change between the years 1995 and 2021. The methods get explained, carried out and the results are getting compared to find semantic change and to find potentials and limits of the used methods. Word embeddings, topic models and in parts cooccurrences show a larger change in context in the 90s and early 2000s. However, through qualitative interpretation of the results it couldn't be said for sure if this context change is the same as semantic change.

# Inhaltsverzeichnis

<b>1. Einleitung</b>	<b>1</b>
1.1. Motivation . . . . .	1
1.2. Forschungsfrage . . . . .	2
1.3. Linguistische Einordnung . . . . .	2
1.4. Relevanz/Forschungslücke . . . . .	4
<b>2. Themenbezogene Arbeiten</b>	<b>6</b>
<b>3. Daten</b>	<b>9</b>
3.1. Word Embeddings - Uni Leipzig Korpora . . . . .	9
3.2. Zeitungen - TAZ . . . . .	9
3.3. Bundestag . . . . .	10
<b>4. Methoden</b>	<b>11</b>
4.1. Word Embeddings . . . . .	11
4.1.1. Preprocessing . . . . .	11
4.1.2. Model . . . . .	13
4.2. Kurze Einführung in den iLCM . . . . .	14
4.3. Zeitliche Kookkurrenzen . . . . .	15
4.4. Topic Modelle . . . . .	15
<b>5. Ergebnisse</b>	<b>17</b>
5.1. Word Embeddings . . . . .	17
5.2. Zeitliche Kookkurrenzen . . . . .	21
5.3. Topic Modelle . . . . .	25
5.4. Vergleich der Ergebnisse . . . . .	27
<b>6. Diskussion</b>	<b>28</b>
<b>7. Fazit</b>	<b>30</b>
7.0.1. Limitierungen . . . . .	31
7.0.2. Ausblick . . . . .	32
<b>A. Anhang</b>	<b>33</b>
A.1. Links zum Quellcode . . . . .	33

## *Inhaltsverzeichnis*

---

A.2. Zusätzliche Grafiken . . . . .	33
<b>Erklärung</b>	<b>41</b>

# 1. Einleitung

## 1.1. Motivation

Sprache verändert sich ständig. Alte Wörter verlieren an Bedeutung oder werden in neuen Kontexten genutzt, neue Wörter werden unserem Wortschatz hinzugefügt [Elizabeth Closs Traugott, 2017]. Oft benutzen unterschiedliche Menschen Wörter auf unterschiedliche Weisen und einigen sich nicht auf eine bestimmte Definition. Im digitalen Zeitalter werden englische Worte in den deutschen Wortschatz aufgenommen, manchmal fast unabhängig ihrer ursprünglichen Bedeutung im Englischen, wie „Handy“ (wörtliche Übersetzung „praktisch“) oder „Public Viewing“ („Öffentliche Besichtigung“, „Live Open Air Screening“ würde im Englischen genutzt werden). Viele Wörter werden in kommerziellen Kontexten genutzt, um das Image einer Firma zu verbessern oder werden speziell zu Großereignissen wie der Olympiade oder Klimakonferenzen vermehrt gebraucht.

Bei all dem stellt sich die Frage, wie und ob sich Bedeutungen von bestimmten Wörtern verändern. Ist es möglich, diesen semantischen Wandel zu messen? Kann ein Kontext zu Ereignissen hergestellt oder kann sogar die Geschwindigkeit des Bedeutungswandels gemessen werden?

„Nachhaltig“ und seine Substantiv-Form „Nachhaltigkeit“ sind Wörter, deren Untersuchung sich im Kontext der Klimakrise und der seit 1995 stattfindenden UN-Klimakonferenzen anbietet. Die Bedeutung der Worte ist nicht eindeutig. Ursprünglich wurde das Wort im 18. Jahrhundert in Deutschland im Kontext der Forstwirtschaft verwendet [Grober, 2013]. Heute wird das Wort in vielen unterschiedlichen Kontexten verwendet - nicht nur in Wirtschaft oder Politik, doch auch viel in sozialen Kontexten. So wird „Nachhaltigkeit“ in drei Bedeutungen unterschieden:

1. „eine längere Zeit anhaltende Wirkung“
2. „forstwissenschaftliches Prinzip, nach dem nicht mehr Holz gefällt werden darf, als jeweils nachwachsen kann“

3. „Prinzip, nach dem nicht mehr verbraucht werden darf, als jeweils nachwachsen oder sich regenerieren und künftig wieder bereitgestellt werden kann.“ [Rödel, 2013]

Diese Arbeit soll unterschiedliche Möglichkeiten betrachten, den Bedeutungswandel von „nachhaltig“ quantitativ zu untersuchen. Die vorgestellten Methoden könnten unter Umständen auch für die Untersuchung des semantischen Wandels anderer Wörter genutzt werden.

## 1.2. Forschungsfrage

Die Fragen, die diese Arbeit beantworten soll, lassen sich insofern formulieren:

1. Lässt sich der Wandel des Kontexts des Wortes „nachhaltig“ mit den computerlinguistischen Methoden von Word Embeddings, zeitlichen Kookkurrenzen und Topic Modellen quantitativ untersuchen?
2. Lässt sich anhand dieser ein semantischer Wandel, eine Veränderung der Wortbedeutung von „nachhaltig“ nachvollziehen?
3. Hat sich die Bedeutung zwischen den Jahren 1995 und 2021 substantiell verändert?
4. Wie unterscheiden sich die drei Methoden in ihren Resultaten und welche Methode lässt die größten Rückschlüsse auf semantischen Wandel zu?
5. Welche Einschränkungen finden sich sowohl bei einzelnen Methoden und welche Limitierungen finden sich bei allen drei Methoden?

## 1.3. Linguistische Einordnung

Es gibt unterschiedliche Formen des semantischen Wandels. Seit dem frühen zwanzigsten Jahrhundert wurde der Bedeutungswandel hauptsächlich mit Änderungen im „Sinn“ eines Wortes assoziiert. Dieser semantische Wandel kann von Veränderung in der Entwicklung eines Vokabulars unterschieden werden, wobei es Überlappungen gibt. Ein Bedeutungswandel beeinflusst sowohl den Sinn, als auch die lexikale Domäne, in welchem ein Wort genutzt wird.

Es gibt zwei Hauptperspektiven auf die Untersuchung von Bedeutungswandel. Die erste ist „semasiologisch“, in der sich die Bedeutung eines Wortes verändert, während die Form zum größten Teil gleich bleibt (Ausnahmen sind dabei pho-

nologischer und morphosyntaktischer Wandel). Dabei wird untersucht, welche Bedeutungen mit einem Wort assoziiert sind, wie diese Bedeutungen miteinander zusammenhängen und wie sie zeitlich entstanden sind [[Elizabeth Closs Traugott, 2017, S. 1](#)]. Die andere Perspektive schaut auf den „onomasiologischen“ Wandel. Darin wird besonders darauf Acht gegeben, wie Wörter mit ihrem Sinn verbunden sind und welche Worte für bestimmte Konzepte genutzt werden. Onomasiologie schneidet sich oft mit Untersuchungen über Veränderungen im Wortschatz [[Elizabeth Closs Traugott, 2017, S. 2](#)]. Oft erweitert sich die Bedeutung eines Wortes über die Zeit, wie neue Bedeutungen für „nachhaltig“.

Wie lässt sich ein Bedeutungswandel identifizieren? Sprecher und Schreiber benutzen existierende Sprache in kreativen neuen Formen und Zuhörer bzw. Leser interpretieren das Gesagte unterschiedlich. Viele dieser Vorgänge passieren nebenbei und nicht absichtlich, viele werden nicht weiter verfolgt. Damit sich ein wirklicher Bedeutungswandel vollzogen haben muss, muss eine „Konventionalisierung“ eingesetzt haben. Als Beweis dafür sollte die neue Bedeutung in mehreren Texten oder Tonspuren benutzt werden.

Arten von Bedeutungswandel können klassifiziert werden. Unter die Bedeutendsten fallen:

1. Metaphorisierung: Eine Sache mit der Beschreibung einer anderen Sache konzeptualisieren.
2. Metonymisierung: Eine Sache durch die Assoziation mit einer anderen Sache zu beschreiben, zum Beispiel ein „runder Tisch“.
3. Pejoration: Die Assoziation eines Begriffs mit einer negativen Bedeutung.
4. Verbesserung: Die Assoziation eines Begriffs mit einer positiven Bedeutung.
5. Eingrenzung: Die Einschränkung der Bedeutung eines Begriffs, zum Beispiel die Assoziation von BLM mit dem Black-Lives-Matter Movement.
6. Generalisierung: Die Erweiterung der Bedeutung eines Begriffs.

[[Elizabeth Closs Traugott, 2017, S. 3](#)]

Über einen längeren Zeitraum können mehrere dieser Effekte auf einen einzelnen Begriff zutreffen. Auch Relationen zwischen Begriffen können sich ändern. So kann ein Wort in der Vergangenheit synonym mit einem anderen gewesen sein, doch heute etwas anderes bedeuten.

Grundsätzlich wird in der linguistischen Theorie davon ausgegangen, dass Worte keine festen Bedeutungen besitzen. Sie rufen bestimmte Bedeutungen hervor oder geben Hinweise auf Bedeutungen. In ihrem Kontext funktionieren

Worte als Anleitung für bestimmte Bedeutungen. Selbst dabei besitzen Worte prototypische Eigenschaften, mit der Möglichkeit, den Kern oder periphere Bedeutungen zu lesen [Elizabeth Closs Traugott, 2017]. Daraus folgt, dass der Kontext eines Wortes die wichtigste Eigenschaft für die Bedeutung eines Wortes ist. Das kommt dieser Arbeit entgegen - durch die Untersuchung des Wortes „nachhaltig“ in verschiedenen Kontexten über einen längeren Zeitraum kann ein möglicher Bedeutungswandel erfasst werden.

Eine Distinktion linguistischer Studien besteht in den Fragen warum Bedeutungswandel entsteht oder wie er entsteht. Aufgrund der technischen/informatischen Natur dieser Arbeit wird die erste Frage nicht betrachtet werden. In dieser Arbeit werden die technischen Methoden erklärt, ob Bedeutungswandel am Beispiel von „nachhaltig“ nachgewiesen werden kann.

„Nachhaltig“ ist im aktuellen allgemeinen Sprachgebrauch stark mit Umweltschutz, insbesondere im Hinblick auf dem Klimawandel, assoziiert. Hypothetisch kann es sich beim semantischen Wandel um eine Eingrenzung handeln. Andererseits ist es auch möglich, dass der Begriff in vielen unterschiedlichen Themengebieten eingesetzt wird und es sich um eine Generalisierung handelt.

## 1.4. Relevanz/Forschungslücke

Die Untersuchung von Bedeutungswandel mit unterschiedlichen computerwissenschaftlichen Methoden ist ein recht neues Feld, welches erst in den letzten zehn Jahren in größerem Stil weiterentwickelt wurde. Die meisten Arbeiten konzentrieren sich dabei auf die technischen Methoden und überlassen die Analyse Linguisten. Allerdings, wie [Tahmasebi et al., 2021, S. 72] es ausdrücken: „When it comes to evaluating methods and systems, there is a general lack of standardized evaluation practices. Different papers use different datasets and testset words, making it difficult or impossible to compare the proposed solutions. Proper evaluation metrics for semantic change detection and temporal analog detection have not been yet established.“

Oft werden in den Arbeiten die Wörter mit den größten Veränderungen herausgesucht, jedoch nicht auf die Entwicklung eines bestimmten Wortes geachtet. Diese Arbeit dient dagegen dem Zweck, den semantischen Wandel des Wortes „nachhaltig“ zu verfolgen. Dabei sollen unterschiedliche computerlinguistische Methoden angewandt und miteinander verglichen werden. Ein Fokus wird dabei auf Word Embeddings gelegt. Verglichen wird dieser Ansatz mit zwei weiteren möglichen Methoden zur Quantifizierung von semantischem Wandel:



Diachrone Kookkurrenzanalyse und Topic Modelle.

## 2. Themenbezogene Arbeiten

Die hier vorgestellten Methoden, Bedeutungswandel zu untersuchen, sind nicht die einzigen Möglichkeiten, auch nicht die einzigen computer-unterstützten statistischen Methoden. Im späten zwanzigsten Jahrhundert veränderte sich das linguistische Feld insofern, dass von Einzeluntersuchungen und Interpretationen auf große Korpora zugegriffen werden konnte. So wird in [Hilpert, 2008] die Untersuchung von Mustern mit statistischen Kookkurrenzen durchgeführt. Wörter wie „will“ und „be going“ werden zusammen mit den Verben untersucht, mit denen sie vorkommen. Falls ein Wort zu unterschiedlichen Zeitpunkten mit einem anderen Hauptverb auftritt, wird das als Indikator für Bedeutungswandel gewertet.

Weitere Untersuchungen basieren auf Web-Crawling [Kerremans et al., 2012] oder auf „Semantic Maps“. Dabei werden Bereiche in konzeptuellen/linguistischen Räumen visuell dargestellt und sowohl grammatische, als auch zeitliche Entwicklungen betrachtet [Croft, 2001].

[Tahmasebi et al., 2021] bieten einen Überblick über eine große Anzahl computergestützter Analysen von semantischem Wandel. Dabei wird zwischen unterschiedlichen Ansätzen unterschieden: Kookkurrenz-basierte Methoden, statische Neural Embeddings und dynamische Word Embeddings. Für jeden dieser Ansätze können unterschiedliche Methoden genutzt werden. Dabei wird sich grundsätzlich auf diese Aussage gestützt: „Regardless of whether pure co-occurrence computing, or contextualized embedding methods are used, a word’s meaning or senses rely on the context in which they appear in a written corpus“ [Tahmasebi et al., 2021, S. 17].

Für die Analyse in dieser Arbeit sind vor allem Kookkurrenzen, Topic Modelle und Word Embeddings von Interesse. Eine Annahme bei der Untersuchung von Word Embeddings und Topic Modellen besteht darin, dass die meisten Wörter ihre Bedeutung über die Zeit nicht ändern. Dadurch bleibt die Repräsentation der Kontextwörter über einen langen Zeitraum konstant. Wörter, die weniger in digitalisierten historischen Dokumenten auftauchen und Studien über Dialekte profitieren davon. Gleichzeitig liegt darin das Risiko des Nicht-Vergessens: In Zeiten, in denen die Assoziation zwischen einem Kontextwort und dem untersuchten Wort gering ist, jedoch beide in den selben Dokumenten vorkommen,

---

können Kontextworte eine größere Rolle spielen, als sie sollten [Tahmasebi et al., 2021, S. 32].

Die meisten Studien im Bereich zeitlicher Kookkurrenzen nutzen „pointwise mutual information (PMI) scores“ anstelle von einfachen frequency scores für Kookkurrenz-Stärke ([Bullinaria and Levy, 2012], [Levy et al., 2015], [Turney and Pantel, 2010]). Die meisten Studien nutzen Kosinus-Ähnlichkeit für die Messung der Ähnlichkeit. Ausnahmen sind Rodda et al. [2017], welche second-order similarity nutzen und Kahmann. et al. [2017], die die Variation von Kookkurrenz-Rängen einzelner Zielworte zu verschiedenen Zeitpunkten untersuchten. Die Ränge basierten auf den zuvor berechneten Signifikanzen. Basile et al. [2016] nutzen Vektoren mit Werten in  $\{-1,0,1\}$  für jeden Term, wobei Elemente, die nicht null sind, selten und zufällig in den Dimensionen der Vektoren zusammen mit Kontextinformationen verteilt sind. Tang et al. [2013] nutzen kontextuelle Entropie [Tahmasebi et al., 2021, S. 18f].

Im Vergleich mit Word Embeddings argumentieren Levy et al. [2015], dass Kookkurrenzen keine grundsätzlich schlechtere Performanz aufweisen, sondern bestimmte Designentscheidungen und Hyperparameter-Optimierungen für bessere Resultate verantwortlich sind. Diese Optimierungen sind nicht inhärent in den Embedding-Algorithmen, sondern können auf traditionelle distributionale Modelle wie Kookkurrenzen übertragen werden.

Eine Schwäche der Kookkurrenz-Untersuchung besteht darin, dass Bedeutungen in den Ergebnissen nicht reflektiert werden. Es ist also fast nicht möglich zu entdecken, was sich genau verändert hat. Topic Modelle und Modelle, die auf Word Sense Induction (WSI) basieren, machen es möglich, auch anzusehen, was sich an einer Wortbedeutung verändert hat. „Common to all topic-based models is that the topics are interpreted as senses. [...] Topics are not in a one-to-one correspondence to word senses (Blei and Lafferty [2006], Wang and Mccallum [2006]) and hence newer induction methods aim at inferring sense and topic information jointly (Wang et al. [2015])“ [Tahmasebi et al., 2021, S. 47].

Lau et al. [2012] vergleichen eine simple, auf Latent Dirichlet Allocation (LDA) basierende Methodologie mit einer bestimmten Anzahl an Wortbedeutungen und einer nicht-parametrische Methode basierend auf einem Hierarchical Dirichlet Process (HDP), welche automatische die Anzahl der Bedeutungen eines Wortes lernen soll. Auf ihrem Test-Datensatz sind die Ergebnisse mit HDP besser als mit LDA. Wang et al. [2015] trennen dagegen Topics von Wortbedeutungen als zwei unterschiedliche Variablen im Unterschied zu LDA und HDP, welche diese Unterscheidung nicht treffen. Weiterhin werden hier die Daten durch Word Embeddings von neuronalen Sprachmodellen und externen Corpora erweitert.

---

Diese Arbeit nutzt LDA. Dabei wird die behandelte Zeitperiode in einzelne Abschnitte aufgeteilt und darauf Topic Modelle berechnet, bevor die Zeitabschnitte miteinander verglichen werden, um semantischen Wandel statt einzelnen Wortbedeutungen zu untersuchen.

Eine andere Möglichkeit bieten die Arbeiten von [Frermann \[2016\]](#), [Lau et al. \[2012\]](#) und [Gulordava and Baroni \[2011\]](#), die dagegen Dynamic Topic Modelle nutzen, bei welchen Topic Modelle über die gesamte Zeitperiode erstellt werden.

Der größere Teil dieser Arbeit beschäftigt sich mit Dynamic Word Embeddings. Die meisten der Arbeiten in diesem Feld haben gemein, dass sie einige Daten über alle Zeitpunkte teilen und dass die daraus resultierenden Embeddings im selben Raum für alle Zeitpunkte existieren. Vektoren müssen also nicht aligniert werden. Allerdings benutzen unterschiedliche Arbeiten unterschiedliche Embedding-Techniken. [Bamler and Mandt \[2017\]](#) nutzen dynamische Word Embeddings, die gemeinsam über alle Zeitperioden trainiert wurden. [Barkan \[2017\]](#) nutzt eine Bayesianische Version des Skip-Gram Modells, bei dem Informationen über alle Zeitpunkte geteilt werden. [Yao et al. \[2017\]](#) nutzen Dynamic PPMI<sup>1</sup> Embeddings, bei denen die Embeddings sich, während sie lernen, schon angleichen. [Rudolph \[2018\]](#) benutzen Dynamic Exponential Family Embeddings, welche Exponential Family Embeddings zusammen mit einem Gauss'schem Random Walk Drift nutzen.

Diese Arbeit stützt sich in großen Teilen auf die Arbeit von [Shoemark et al. \[2019\]](#), in welchem ein neues Evaluations-Framework für die Erkennung von Bedeutungswandel mit Word Embeddings vorgestellt wird. Dabei argumentieren die Autoren, dass das Nutzen der Daten aus einer gesamten Zeitfolge besser ist als nur ein Vergleich des ersten und letzten Zeitpunktes. Unabhängig trainierte Daten auf Word Embeddings funktionieren für längere Zeiträume generell besser und der Referenz-Zeitpunkt für einen Vergleich sollte beachtet werden [[Shoemark et al., 2019](#), S. 2].

---

<sup>1</sup> Positive Pointwise Mutual Information Matrix

## 3. Daten

Wichtig für die Untersuchung von semantischem Wandel ist die Auswahl brauchbarer Korpora. Der Datensatz darf nicht zu klein sein, sollte jedoch das Thema Nachhaltigkeit als Inhalt enthalten. Eine Möglichkeit dafür ist ein großer Datensatz von Zeitungsmaterial. Weitere Möglichkeiten wären Transkripte von Nachrichtensendungen oder Fachmaterial. Reden und Vorträge, die das Thema enthalten, sind ebenso möglich. Allerdings sollten sich die Korpora nicht zu sehr einschränken, um nicht nur eine bestimmte Bedeutung von „nachhaltig“ zu entdecken, sondern den Bedeutungswandel in so vielen Kontexten wie möglich zu untersuchen.

### 3.1. Word Embeddings - Uni Leipzig Korpora

Für die Untersuchung von Word Embeddings konnten Daten aus dem Wortschatz-Portal der Universität Leipzig genutzt werden. Von der Leipzig Corpora Collection wurden aus der News-Sektion von 1995 bis 2021 jeweils die Datensätze über 1M Sätze heruntergeladen, welches Quellenmaterial von Online-Nachrichtenseiten stammt. Die Sätze sind automatisch zusammengestellt und nicht auf das Vorkommen von Nachhaltigkeit bzw. nachhaltig vorsortiert<sup>2</sup>.

Die durchschnittliche Anzahl der Worte pro Text sind ca. 16. Interessant für das Vorkommen des Wortes „nachhaltig“ ist der Anstieg der Nutzung zwischen 1995 und 2021, mit einer kleinen Spitze 2012 aufgrund der Konferenz der Vereinten Nationen über nachhaltige Entwicklung in Rio de Janeiro<sup>3</sup>.

### 3.2. Zeitungen - TAZ

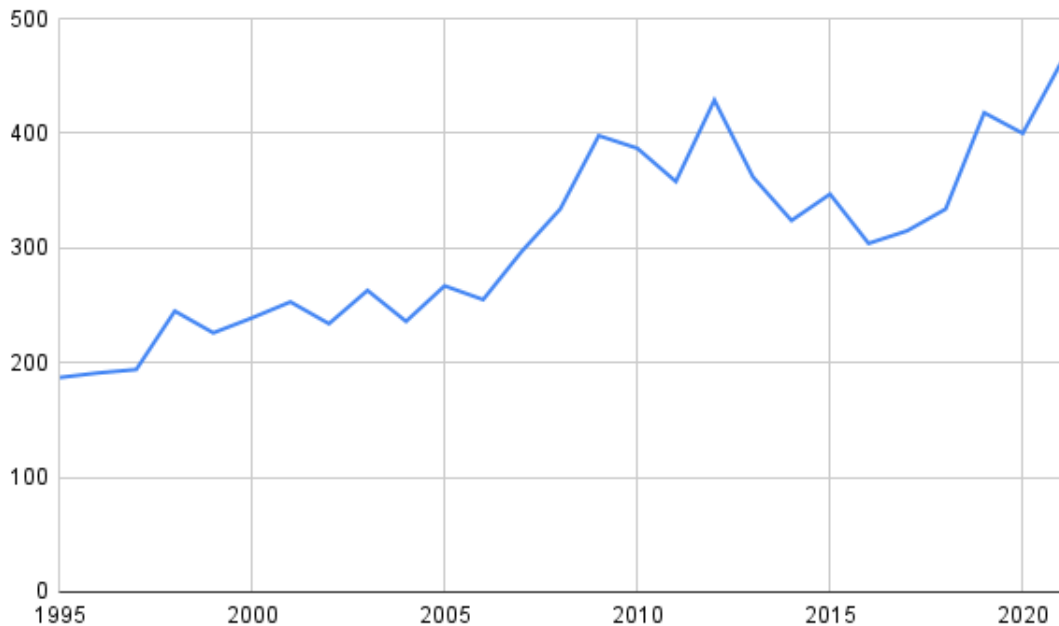
Topic Modelle und zeitliche Kookkurrenzen benutzten im voraus nach den Suchtermen gefilterte Datensätze unterschiedlicher Quellen. Beide Untersu-

<sup>2</sup> <https://wortschatz.uni-leipzig.de/en/download/German>

<sup>3</sup> Abbildung 3.1

### 3.3 Bundestag

---



**Abbildung 3.1.:** Häufigkeit des Wortes „nachhaltig“ im Datensatz der Uni Leipzig von 1995 bis 2021

chungen nutzten dieselben Daten. Der erste Datensatz besteht aus Artikeln der Tageszeitung (TAZ) aus den Jahren 1999 bis 2018. Die Auswahl der Artikel wurde nach dem Vorkommen der Worte „nachhaltig“ bzw. „Nachhaltigkeit“ gefiltert, so dass ein Gesamt-Datensatz von 14676 Dokumenten entstand.

### 3.3. Bundestag

Der zweite Datensatz besteht aus einer Zusammenstellung von auf „nachhaltig“ bzw. „Nachhaltigkeit“ gefilterten Reden aus dem Deutschen Bundestag zwischen 1995 und 2018. Er besteht aus 150854 Dokumenten.

Sowohl der Datensatz Bundestag, als auch der Datensatz TAZ beginnen nicht mit vollen Jahren, sondern nur den letzten Monaten der Jahre 1995 bzw. 1999. Daher ist eine Analyse dieser Jahre weniger aussagekräftig, als die der Folgenden.

## 4. Methoden

Wie in [Kapitel 2](#) vorgestellt, sind eine große Anzahl unterschiedlicher Methoden zur Bestimmung semantischen Wandels möglich. Eine der Schwierigkeiten dabei existiert in der Zuordnung zu zeitlichem Kontext. Eine quantitative Analyse, die Messung z.B. von einzelnen Kookkurrenzen, ist alleine nicht ausreichend. Eine qualitative Bewertung der Ergebnisse muss dabei mitgedacht werden. Es ist daher von Interesse, eine Kombination beider Arten der Analyse durchzuführen.

In dieser Arbeit wurden drei unterschiedliche Methoden genutzt: Word Embeddings, Kookkurrenzen und Topic Modelle. Die erste Analyse, Word Embeddings, wurde durch eigenständige Programmierung unter Bezug auf [[Shoemark et al., 2019](#)] und Nutzung der von Gensim bereitgestellten Word2vec Algorithmen<sup>4</sup> durchgeführt. Die Analyse von zeitlichen Kookkurrenzen und Topic Modellen nutzen dagegen den an der Universität Leipzig entwickelten iLCM<sup>5</sup>. Nach jeder Methode wurde versucht, die Ergebnisse qualitativ zu analysieren - vor allem in Hinblick auf die Themenbereiche, in denen „nachhaltig“ in den letzten 25 Jahren genutzt wurde.

Durch die Kombination qualitativer und quantitativer Ansätze sollte es möglich sein, eine Lösung der Fragestellung, zumindest im Hinblick auf die vorhandenen Daten, zu finden.

### 4.1. Word Embeddings

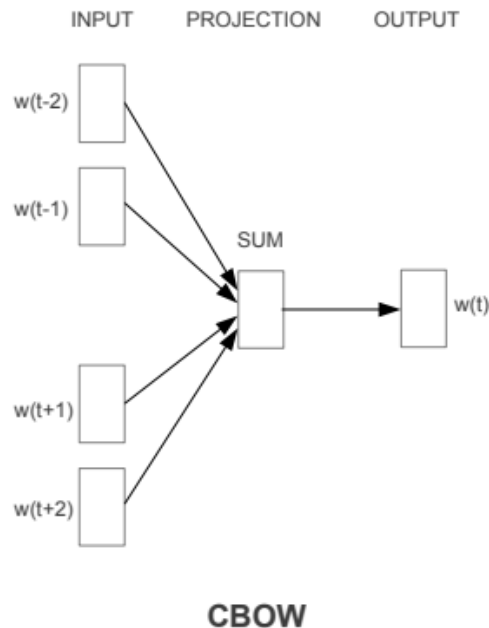
#### 4.1.1. Preprocessing

Nach dem Ansatz von [[Hamilton et al., 2016](#)] ist der Datensatz in einzelne Zeitabschnitte aufgeteilt, in diesem Fall ein Zeitabschnitt  $t$  für ein Jahr. Für jeden Zeitabschnitt  $t$  werden die Daten vorprozessiert, indem Sonderzeichen entfernt,

---

<sup>4</sup> <https://radimrehurek.com/gensim/models/word2vec.html>

<sup>5</sup> [iLCM](#)



**Abbildung 4.1.:** Darstellung der CBOW-Modell-Architektur von [Mikolov et al. \[2013\]](#)

Lemmatisierung vollzogen und alle Worte in Kleinschreibung konvertiert werden. Zwei Ansätze werden trainiert, einmal mit und einmal ohne Named Entity Recognition. Dann werden Word Embeddings berechnet, bei denen jedes Wort  $w$  als  $d$ -dimensionaler Vektor dargestellt wird.

Für das Training der Word Embeddings wurde Gensim genutzt<sup>6</sup>, die Einstellungen sind in diesem Fall: `min_count=2`, `window=10` und `vector_size=200`, die anderen Einstellungen bleiben auf den Default-Werten der Gensim-Bibliothek.

Die Berechnung der Word Embeddings mit `word2vec` nutzt einen Text-Korpus als Input und produziert die Wortvektoren als Output. Dabei wird ein Vokabular aus den Trainingsdaten erstellt und die Vektorrepräsentationen der einzelnen Wörter mit dem Continuous Bag of Words (CBOW) Modell gelernt. Aus den Kontextwörtern  $w(t-n)..w(t-1), w(t+1)..w(t+n)$  eines gesuchten Wortes  $w(t)$  an der Stelle  $t$  im Korpus wird ein Durchschnitt berechnet, der unabhängig von der Reihenfolge ist und genutzt wird, um das Wort  $w(t)$  vorauszusagen. Das vorausgesagte Wort wird mit dem tatsächlichen Zielwort verglichen, dann werden die Embeddings aufgrund der Fehlerrückführung neu gewichtet und der Prozess mehrfach wiederholt [[Mikolov et al., 2013](#)].

<sup>6</sup> <https://radimrehurek.com/gensim/>



### 4.1.2. Model

Für die Quantifizierung des semantischen Wandels zwischen zwei Zeitpunkten  $t$  und  $t+1$  werden zwei Arten der Messung verwendet. Erstere ist die Kosinus-Distanz. Die zweite Methode vergleicht die Nachbarschaften der Embeddings.

Für jeden Zeitpunkt  $t$  werden für jedes Wort  $w$  die  $k$ -nächsten Nachbarn gesucht, basierend auf Kosinus-Ähnlichkeit. Standardmäßig ist  $k = 25$ . Als nächstes werden für  $t$  und  $t+1$  Ähnlichkeitsvektoren der „zweiten Ordnung“ erstellt: Für alle zwei Zeitpunkte wird die Vereinigung  $S$  der Mengen der  $k$ -nearest neighbours und auf dieser Vereinigung der Vektor  $v^t$  berechnet. Jeder Eintrag  $v^t(i)$  enthält die Kosinus-Ähnlichkeit zwischen dem Zielwort  $w$  und dem benachbarten Wort  $S^i$  zum Zeitpunkt  $t$ . Die Idee dahinter ist, dass für den Bedeutungswandel eines Wortes nur die nächsten semantischen Nachbarn relevant sind. Zwischen diesen Vektoren wird daraufhin die Kosinus-Distanz berechnet, den Neighborhood Shift [[Shoemark et al., 2019](#), S. 68].

[Shoemark et al. \[2019\]](#) vergleichen die Effektivität der beiden Ansätze. Es wurden sieben Schemata erstellt, wie sich die distributionale Statistik eines Wortes ändern kann. Drei davon haben mit Bedeutungswandel zu tun - Generalisierung, Veränderung der Bedeutung und die Eingrenzung einer Begriffsbedeutung. Die anderen vier Schemata sind kein Bedeutungswandel, können aber trotzdem den Kontext eines Wortes verändern. Darunter fallen eine Häufung des Wortes, Wörter mit zwei konsistenten Bedeutungen, bei denen die eine nur sporadisch und zufällig auftritt, Wörter, die saisonal in neuem Kontext auftreten und Wörter mit vielen unterschiedlichen Bedeutungen, bei denen eine Eingrenzung nicht stattfindet [[Shoemark et al., 2019](#), S. 69-71].

Die für diese Arbeit relevanten Ergebnisse der von [Shoemark et al. \[2019\]](#) durchgeführten Evaluation sind wie folgt:

- Ein Vergleich nur des ersten und des letzten Zeitpunkts ist nicht so effektiv wie die Nutzung der gesamten Reihe an Zeitpunkten.
- Ein unabhängiges Training einzelner Zeitpunkte ist effektiver als ein stetiges Training über alle Zeitpunkte und ergibt konsistent bessere Ergebnisse.
- Unterschiedliche Zeit-Serien-Ansätze sind besser mit unterschiedlichen Ähnlichkeitsberechnungen. Das Neighborhood Measure zeigte bei Substantiven größere Unterschiede im Bedeutungswandel, während Adjektive effektiver besser mit der Kosinus-Distanz berechnet wurden. Neighborhood ist für die Erkennung von Zeitpunkten, an denen sich Wortbedeutungen geändert haben (change point detection), während Korrelationen

besser durch die Kosinus-Distanz erkannt wurden [Shoemark et al., 2019, S. 71-73].

## 4.2. Kurze Einführung in den iLCM

Für einen Teil der Analyse, die Berechnung von Kookkurrenzen und Topic Modellen, wurde der an der Universität Leipzig entwickelte iLCM verwendet. Dabei handelt es sich um eine integrierte Forschungsumgebung für die Analyse von strukturierten und unstrukturierten Daten mit Methoden des Text Minings. Der iLCM besteht aus zwei größeren Komponenten: Der *Leipzig Corpus Miner* (LCM) und *Open Research Computing* (ORC). Der LCM bietet eine Server-Infrastruktur, bestehend aus einer Dokumentendatenbank (MariaDB<sup>7</sup>), einer NLP pipeline für das Preprocessing von Textdaten (spaCy<sup>8</sup>), einem Volltext-Index (Solr<sup>9</sup>), einer Sammlung an Text-Mining-Prozessen und eine graphische Nutzeroberfläche (GUI mit R Shiny<sup>10</sup>). Das System ist in eine Virtual-Machine-Umgebung (Docker<sup>11</sup>) eingebunden. Dadurch ist es auch ohne die Notwendigkeit zu programmieren möglich, Corpora mit mehreren Millionen Dokumenten zu verwalten und Analysen durchzuführen. Für die Verwaltung der Dokumente bietet der LCM die Möglichkeit, Daten in unterschiedlichen Textformaten mit ihren Metadaten zu importieren und darauf Volltext- und Metadaten-Suchen anzuwenden. Für die Analyse sind neben dem Preprocessing und graphischen Repräsentationen der Metadaten auch Methoden der Analyse, wie Kookkurrenzanalysen, Kontextvolatilität, Topic Modelle etc. möglich.

Die ORC-Umgebung erlaubt die Erweiterung der Analyse-Prozeduren des LCMs durch die Eingabe von Programm-Skripten im Webbrowser, die direkt ausgeführt und im selben Dokument visualisiert werden können. Diese ausführbaren Dokumente, die Kombination aus Skript-Code, Dokumentation und den Resultaten, werden „Notebooks“ genannt und können in einem Repository geteilt, verändert und ausgeführt werden [Niekler et al., 2018]. Für diese Arbeit wurden die Methoden des LCM genutzt, darunter die Dokumentenverwaltung und die Analyse von Corpora mit Kookkurrenzen und Topic Modellen, inklusive ihrer Visualisierung.

---

<sup>7</sup> <http://mariadb.org>

<sup>8</sup> <http://spacy.io>

<sup>9</sup> <https://solr.apache.org/>

<sup>10</sup> <https://shiny.rstudio.com/>

<sup>11</sup> <http://www.docker.com>

### 4.3. Zeitliche Kookkurrenzen

Die Berechnung der Kookkurrenzen erfolgte mit dem iLCM an der Universität Leipzig<sup>12</sup>. Sowohl der Datensatz TAZ, als auch der Datensatz Bundestag wurden dabei in zweijährige Zeitschritte aufgeteilt. Dann wurden Sonderzeichen und Zahlen sowie Stoppwörter entfernt, die Wörter in Kleinschreibung transformiert und lemmatisiert, bevor Kookkurrenzen berechnet wurden. Die zwei für die Berechnung genutzten Signifikanzmaße sind der Dice-Koeffizient und Log-Likelihood.

Um die zeitliche Ebene auszurechnen, wurden die Ergebnisse der jeweiligen Zwei-Jahresschritte, die Rangliste der kookkurrenten Wörter von "nachhaltig" zusammen mit ihren Signifikanzmaßen als vieldimensionale Vektoren angesehen und die Kosinus-Distanz zwischen den Zwei-Jahresschritten berechnet. Dabei wurde sichergestellt, dass die Vektoren nach den kookkurrenten Wörtern sortiert wurden, um die korrekte Distanz zu berechnen. Einen Sonderfall bilden dabei die Wörter, die nur zu einem Zeitpunkt auftauchen. Hier wurde für den Zeitpunkt, an dem das Wort nicht auftaucht, standardmäßig der Wert 0 angenommen, da ein berechnetes Skalarprodukt ebenfalls den Wert 0 annehmen würde.

### 4.4. Topic Modelle

Auch die Berechnung der Topic Modelle erfolgte mit dem iLCM. Die Datensätze Bundestag und TAZ wurden ebenso in zweijährige Zeitschritte aufgeteilt, bereinigt und lemmatisiert, bevor Topic Modelle mit Latent Dirichlet Allocation (LDA) berechnet wurden [Blei et al., 2003; Blei, 2011]. Alle Wörter, die weniger als dreimal in Texten vorkamen, wurden entfernt.

Die Anzahl der Topics wurde auf 10 begrenzt. Dies wurde als Kompromiss zwischen qualitativer und quantitativer Analyse gewählt - oft werden in Analysen die Anzahl zwischen 10 und 20 Topics gesetzt (vgl. Deveaud et al. [2014]). Für die qualitative Analyse über die große Anzahl von Jahren war es notwendig, die Anzahl der Topics zu begrenzen, um die meisten Topics konsistent zu halten und die Anzahl der Topics, welche keine inhaltlichen Zusammenhänge besaßen, zu begrenzen.

---

<sup>12</sup> iLCM

Neben der Analyse mit LDA wurde auch der Relevanz-Wert einzelner Begriffe berechnet, um ein Überlappen der Themen möglichst zu reduzieren. Schließlich wurde die Anzahl der Begriffe eines Topics auf 50 gesetzt. Im iLCM wurden diese mit LDAVis dargestellt und zur Analyse heruntergeladen [Sievrt and Shirley, 2014].

Für die quantitative Analyse der zeitlichen Veränderung wurden ebenfalls jeweils die Topics heruntergeladen und die Relevanz mit dem Wert  $\lambda = 0.5$  berechnet. Als Nächstes wurde für jeden Zeitabschnitt berechnet, wie viele der 10 Topics einen Prozentsatz  $p$  Wörter aus einem Topic des vorherigen Zweijahresabschnittes übernahmen. Wenn zum Beispiel der Wert  $p = 0,5$  betrug, wurden alle Topics, die ein korrespondierendes Topic mit mindestens 50% gleichen Wörtern im vorigen Zeitabschnitt  $t$  besaßen, als „nicht verändertes Topic“ definiert und nicht gezählt.  $n$  ist die Anzahl Wörter, die aus jedem Topic genommen wurden.

$$test\_topic(t_1, T_2, n, p) = \begin{cases} 0 & \text{if } \exists t_2 \in T_2. |t_1 \cap t_2| \geq np \\ 1 & \text{else} \end{cases}$$

Das Ergebnis ist die Anzahl der veränderten bzw. neuen Topics, die auf einem Zeit-Diagramm dargestellt wurden.

$$changed\_topics(T_1, T_2, n, p) = \sum_{t_1 \in T_1} test\_topic(t_1, T_2, n, p)$$

# 5. Ergebnisse

## 5.1. Word Embeddings

Die Ergebnisse für die Berechnungen von Kosinus-Distanz und Neighbourhood-Shift lassen Trends erkennen. Die Ergebnisse sind nach Startjahr dargestellt, so umfasst zum Beispiel in [Abbildung 5.1](#) der Wert bei 2020 die Kosinus-Distanz des Wortes nachhaltig zwischen 2020 und 2021.

Die höchste Kosinus-Distanz mit Named Entity Recognition existiert zwischen den zwei Jahren 2004 und 2005<sup>13</sup>. Sie beträgt ca. 0,5095. Zum Vergleich: Die Wörter mit dem größten Wert sitzen zwischen 1,1456 und 1,0272. Die 100 ähnlichsten Worte zu „nachhaltig“ in den zwei Jahren zeigen einen Unterschied in den Themenfeldern, die diese Beobachtung unterstützen. So sind die Worte 2004 hauptsächlich im wirtschaftlichen Bereich anzusiedeln, zum Beispiel mit „konjunkturell“, „Wettbewerbsfähigkeit“ oder „ökonomisch“. 2005 ist dieser wirtschaftliche Bereich zwar noch immer sichtbar, doch fallen auch mehr Umwelt-Aspekte auf, wie die Worte „ökologisch“ und „Umweltschutz“<sup>14</sup>.

Im Gegensatz zur Kosinus-Distanz gibt es beim Neighbourhood Shift mit NER zwei Spitzen: 2000 auf 2001 und 2006 auf 2007<sup>15</sup>. Der Neighbourhood Shift zwischen 2000 und 2001 beträgt ca. 0,0888. Ähnliche Worte lassen sich 2000 mehr dem Themenfeld der Volkswirtschaft zuordnen, während 2001 Finanzwirtschaft einen höheren Stellenwert bekommt. Zwischen 2006 und 2007 beträgt der Wert des Neighbourhood Shifts ca. 0,0879. 2006 und 2007 stehen ökologische Themen schon mehr im Vordergrund, jedoch findet sich 2006 mehr Ähnlichkeit mit Worten wie „Stabilität“, während 2007 „entgegenwirken“ und „Transformation“ größere Bedeutung erlangen.

Die größte Spitze der Kosinus-Distanz ohne Named Entity Recognition ist zwischen den Jahren 1999 und 2000 mit dem Wert 0,5273, mit einer weiteren Spitze zwischen 2001 und 2002 mit dem Wert 0,5182<sup>16</sup>. Danach fällt der mögliche se-

<sup>13</sup> [Abbildung 5.1](#)

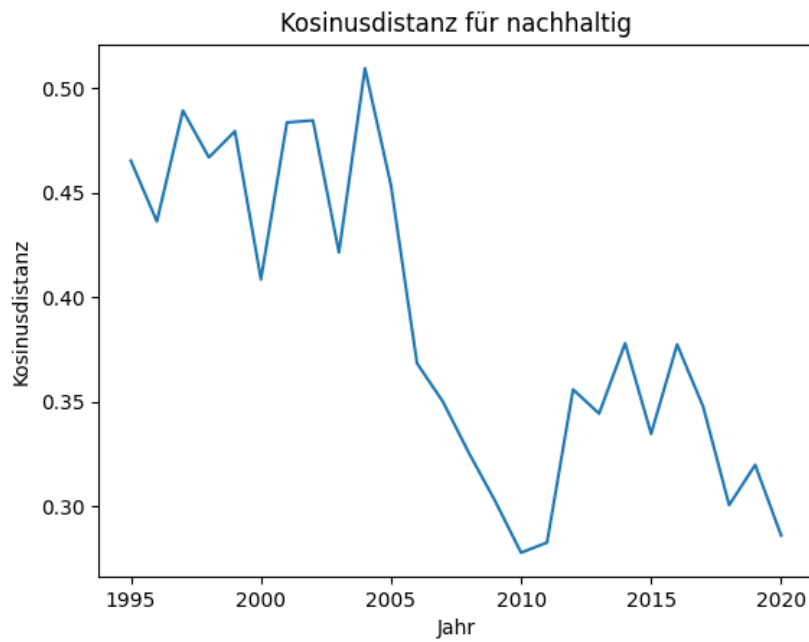
<sup>14</sup> Siehe [Abbildung 5.1](#)

<sup>15</sup> [Abbildung 5.2](#)

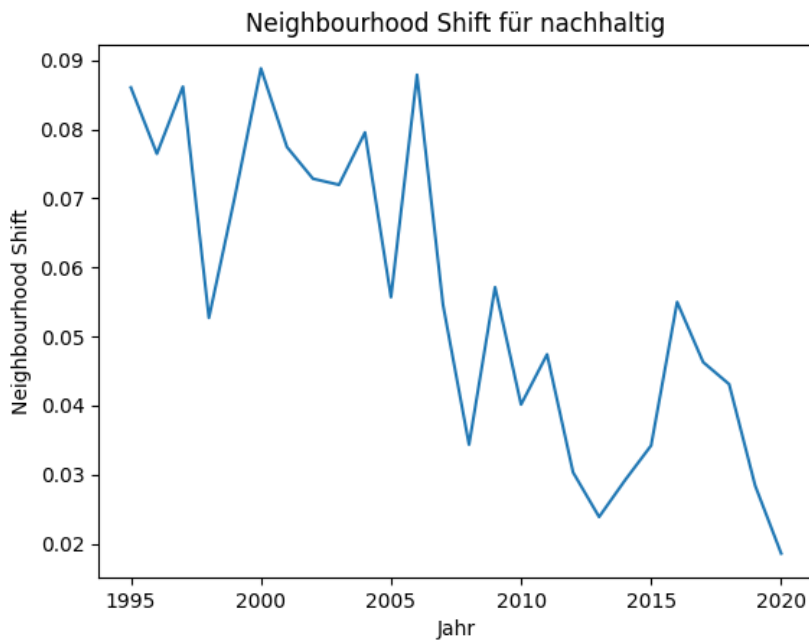
<sup>16</sup> [Abbildung 5.3](#)

## 5.1 Word Embeddings

---



**Abbildung 5.1.:** Kosinus-Distanz der Word Embeddings zwischen 1995 und 2021 mit Named Entity Recognition, Datensatz Uni Leipzig Korpora



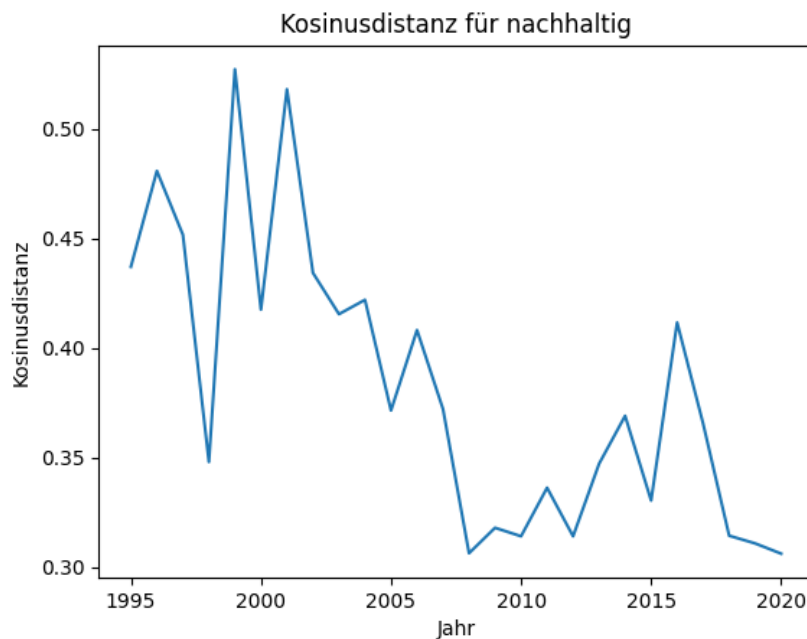
**Abbildung 5.2.:** Neighborhood Shift der Word Embeddings zwischen 1995 und 2021 mit Named Entity Recognition, Datensatz Uni Leipzig Korpora

## 5.1 Word Embeddings

---

**Tabelle 5.1.:** Vergleich der ähnlichsten Worte zu „nachhaltig“ zwischen 2004 und 2005 anhand der Kosinusdistanz, mit Named Entity Recognition, Datensatz Uni Leipzig Korpora

Ähnlichste Worte 2004		Ähnlichste Worte 2005	
Wort	Kosinusdistanz	Wort	Kosinusdistanz
konjunkturell	0.48153117299079895	langfristig	0.44494330883026123
wettbewerbsfähigkeit	0.439864844083786	spürbare	0.44336438179016113
internationalisierung	0.419548898935318	realwirtschaft	0.43882665038108826
strukturell	0.4176228642463684	konjunkturell	0.42783600091934204
demographisch	0.4167070984840393	wirtschaftlich	0.4105398654937744
ökonomisch	0.41635891795158386	demographisch	0.4064999222755432
anziehend	0.41427111625671387	außenhandel	0.40625181794166565
mittelstand	0.4138456881046295	ökologisch	0.38993918895721436
demografischen	0.4121209383010864	staatsfinanzen	0.38520702719688416
unbewusste	0.40870222449302673	wettbewerbsfähigkeit	0.3805069327354431
zukäufe	0.4027814567089081	optimierung	0.38049474358558655
konjunktur	0.397081583738327	hemmen	0.3755541741847992
binnenwirtschaft	0.3960671126842499	strukturell	0.37495049834251404
erholung	0.395352840423584	dauerhaft	0.3722543716430664
stabile	0.3945170044898987	spürbar	0.37089285254478455
investitionsklima	0.39449700713157654	demografische	0.37028563022613525
strukturreformen	0.3919374346733093	arbeitsmarktes	0.36911535263061523
langfristig	0.3904595971107483	ökonomisch	0.36874693632125854
leitbild	0.39043989777565	preisstabilität	0.36716675758361816
stabilisierung	0.3903023302555084	nachhaltigkeit	0.36570122838020325



**Abbildung 5.3.:** Kosinus-Distanz der Word Embeddings zwischen 1995 und 2021 ohne Named Entity Recognition, Datensatz Uni Leipzig Korpora

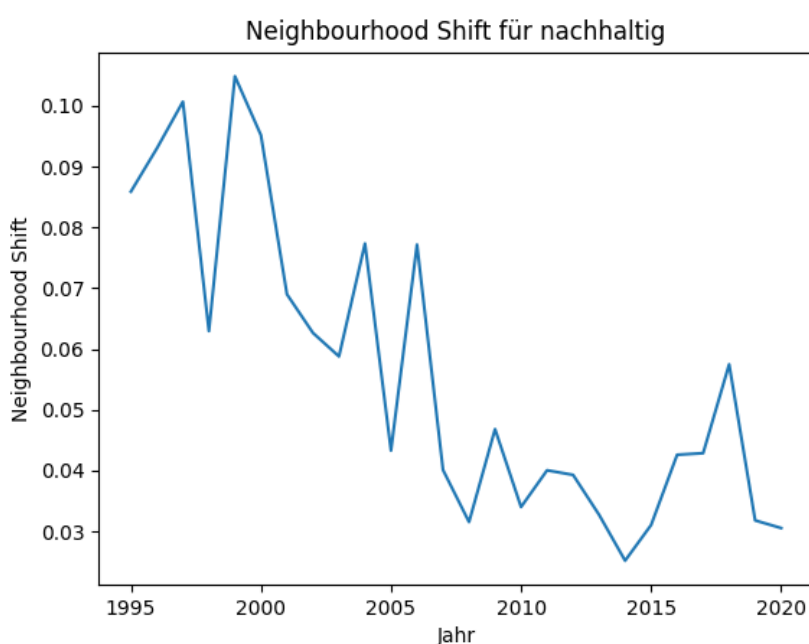
mentische Wandel stark ab, im Gegensatz zu der Spitze zwischen 2004 und 2005, die mit NER noch existiert.

Wie die Kosinus-Distanz zeigt auch der Neighbourhood Shift den größten Bedeutungswandel zwischen 1999 und 2000, mit einem Wert von ca. 0,1049. Die Kurve zeigt sich ähnlich, wobei es zwischen 2001 und 2002 keine Spitze gibt<sup>17</sup>. Themengebiete sind weniger eindeutig, so sind 1999 sowohl Wirtschaftlichkeit, als auch Soziales und Umwelt in den ähnlichsten Worten präsent. 2000 sind Arbeit und Volkswirtschaft präsenter als soziale Themen.

Wenn man alle Ergebnisse miteinander vergleicht, ist spätestens nach 2007 ein weitaus geringerer semantischer Wandel von „nachhaltig“ zu erkennen. Ohne Named Entity Recognition liegt die größte Veränderung um das Jahr 1999, während die Spitze mit NER zwischen 2003 und 2007 liegt. Es lässt sich eine konstantere Verwendung des Begriffs nach 2007 feststellen, mit einem Trend zu ökologischeren Themen. Das Thema Wirtschaft bleibt weiterhin wichtig für den Sprachgebrauch, während Soziales eher in den Hintergrund rückt.

<sup>17</sup> Abbildung 5.4





**Abbildung 5.4.:** Neighborhood Shift der Word Embeddings zwischen 1995 und 2021 ohne Named Entity Recognition, Datensatz Uni Leipzig Korpora

## 5.2. Zeitliche Kookkurrenzen

Die Berechnungen zeitlicher Kookkurrenzen auf dem Bundestag-Datensatz zeigen trotz unterschiedlichem Datensatz und Aufteilung in Zweijahresschritte einen grob ähnlichen Trend wie Word Embeddings, vor allem ohne Named Entity Recognition. So sinkt die Kosinus-Distanz bei beiden Signifikanzmaßen im Bundestag-Datensatz signifikant nach dem zwischen den Jahren 2000 und 2002.<sup>18</sup> Die größte Distanz findet sich sowohl bei Dice, als auch bei Log Likelihood zwischen den Zweijahresschritten ‚2001/2002‘ und ‚2003/2004‘.

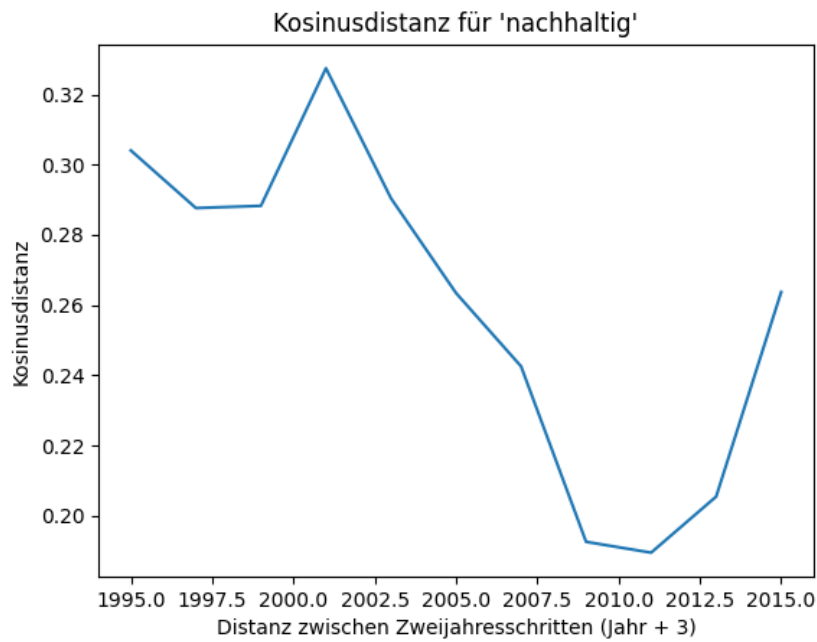
Vergleicht man die Jahre qualitativ miteinander, fällt der Grund schnell auf. So fand 2002 der „Weltgipfel für nachhaltige Entwicklung“ in Johannesburg statt<sup>19</sup>, welcher sich eindeutig in den Kookkurrenzen für 2001/2002 findet, danach nicht mehr<sup>20</sup>. Die Konferenz der Vereinten Nationen über nachhaltige Entwicklung in Rio de Janeiro 2012 findet sich im Bundestag nicht wieder, aber ab dem Weltgipfel für nachhaltige Entwicklung 2015 und insbesondere die Übernahme

<sup>18</sup> Abbildung 5.5 u. Abbildung 5.6

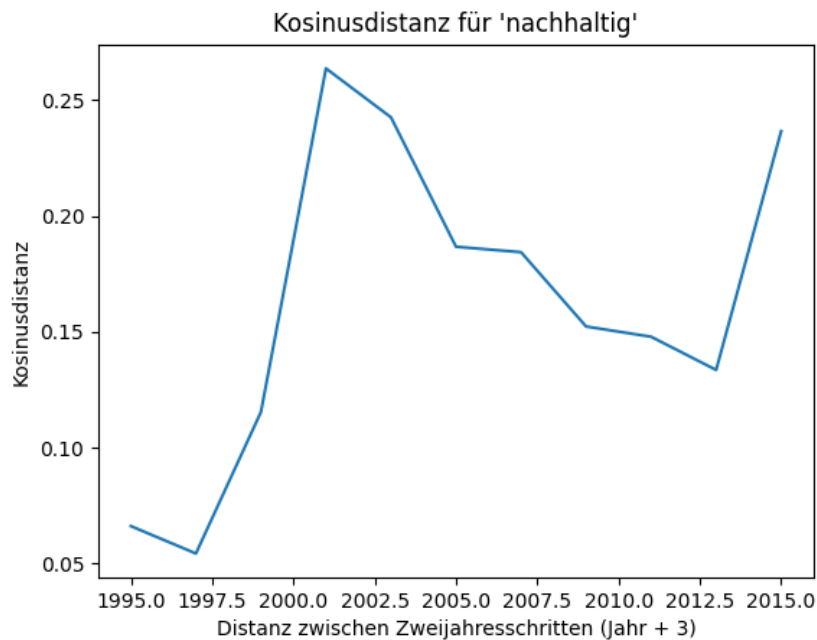
<sup>19</sup> <https://www.bmz.de/de/service/lexikon/weltgipfel-fuer-nachhaltige-entwicklung-14944>

<sup>20</sup> Abbildungen A.1, A.2, A.3, A.4

## 5.2 Zeitliche Kookkurrenzen

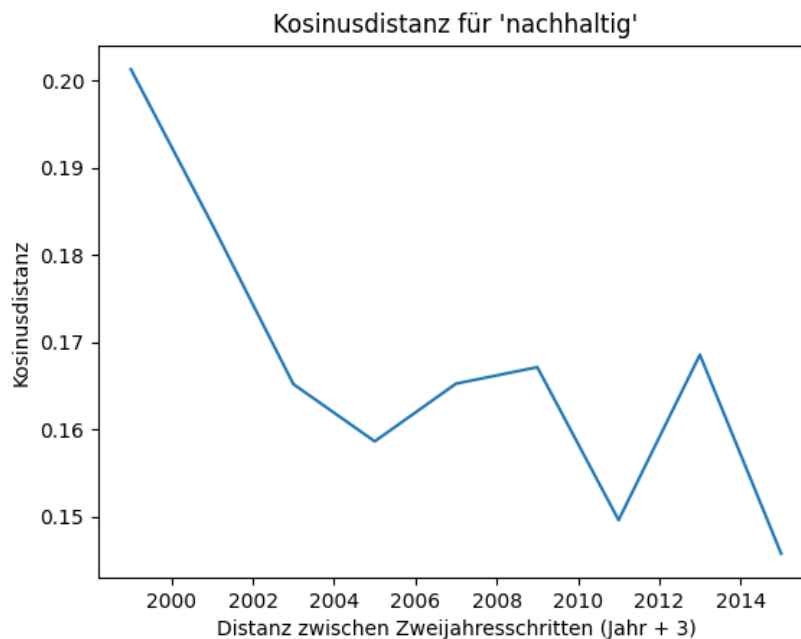


**Abbildung 5.5.:** Kosinus-Distanz der zeitlichen Kookkurrenzen über Dice-Signifikanz zwischen 1995 und 2018 für Bundestag-Datensatz



**Abbildung 5.6.:** Kosinus-Distanz der zeitlichen Kookkurrenzen über Log-Likelihood-Signifikanz zwischen 1995 und 2018 für Bundestag-Datensatz

## 5.2 Zeitliche Kookkurrenzen



**Abbildung 5.7.:** Kosinus-Distanz der zeitlichen Kookkurrenzen über Dice-Signifikanz zwischen 1999 und 2018 für TAZ-Datensatz

der dabei beschlossenen 2030-Agenda als Deutsche Nachhaltigkeitsstrategie ist sichtbar<sup>21</sup>.

Die Kosinus-Distanzen für den Datensatz der TAZ sehen dagegen stark unterschiedlich aus. So startet die Dice-Signifikanz mit einem hohen Wert zwischen den Zweijahresschritten 1999/2000 und 2001/2002, sinkt danach aber signifikant nach unten<sup>22</sup>. Ein Blick auf die Kookkurrenzen zeigt einen recht konstant bleibenden Fokus auf Wirtschaft und Entwicklung.

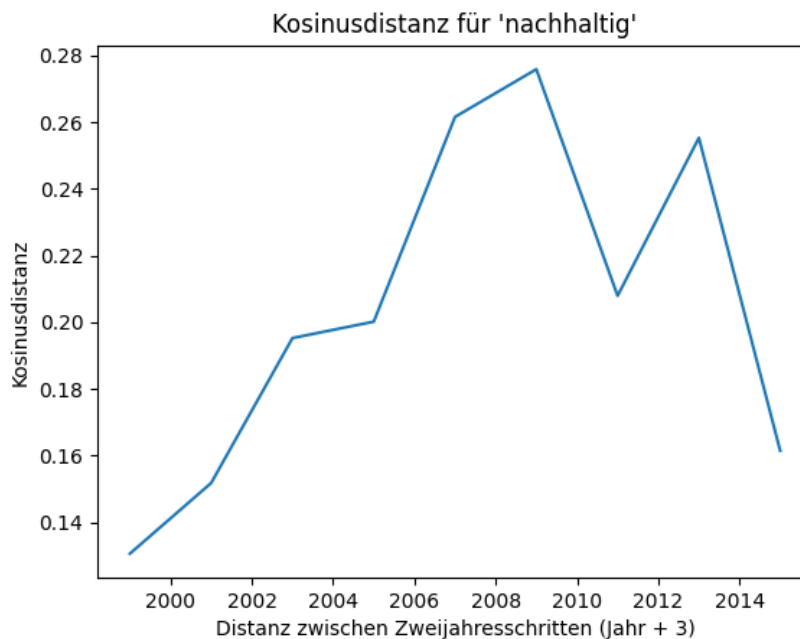
Die Kosinus-Distanz bei Log-Likelihood zeigt einen starken Unterschied zum Rest des Datensatzes.<sup>23</sup> Durch eine qualitative Analyse kann dieser Unterschied nicht eindeutig erklärt werden. Es gibt aber einige Erklärungsansätze: Während die Kookkurrenz-Werte sich bei Dice zwischen 0 und 1 bewegen, kann eine Kookkurrenz bei Log-Likelihood auf einen Wert >1000 kommen. Allerdings finden sich in den Jahren mit den größten Distanzen keine so großen Werte, als dass sie den Unterschied alleine erklären würden. Stattdessen bewegen sich die Themen in der TAZ, die im Zusammenhang mit „nachhaltig“ aufkommen, in

<sup>21</sup> <https://www.bundesregierung.de/breg-de/themen/nachhaltigkeitspolitik/eine-strategie-begleitet-uns/deutsche-nachhaltigkeitsstrategie>

<sup>22</sup> Abbildung 5.7

<sup>23</sup> Abbildung 5.8

## 5.2 Zeitliche Kookkurrenzen



**Abbildung 5.8.:** Kosinus-Distanz der zeitlichen Kookkurrenzen über Log-Likelihood-Signifikanz zwischen 1999 und 2018 für TAZ-Datensatz

mehreren Bereichen, die sich zwischen den Zweijahresschritten stark verändern können, auch wenn sie keine langfristigen Veränderungen in der Nutzung des Wortes darstellen.

Zum Beispiel spielt die Fischerei in den Jahren 2011 bis 2014 eine größere Rolle, was mit der Reform der europäischen Fischereipolitik 2013 zusammenhängt. Allerdings ist nachhaltige Fischerei ein über einen längeren Zeitraum existierendes Thema. Weitere Spitzen korrelieren mit dem Fortschrittsbericht 2008 zur nationalen Nachhaltigkeitsstrategie, der Überprüfung der EU Nachhaltigkeitsstrategie 2009 und Konferenz der Vereinten Nationen über nachhaltige Entwicklung in Rio 2012. Es ist möglich, dass ein größerer und in kleinere Teile aufgeteilter Datensatz detailliertere Resultate erzielen könnte.

Durch die nach Datensatz und Signifikanz unterschiedlichen Ergebnisse ist eine Bewertung für den semantischen Wandel schwierig. Sehr langfristig können Kookkurrenzen als Kontext eines Wortes Informationen geben, aber kurzfristig gibt es zu viele Unterschiede, vor allem bei einem so breit genutzten Wort wie „nachhaltig“. Stattdessen geben Kookkurrenzen Informationen über Ereignisse, die im Kontext mit dem Wort stehen. So konnten der „Weltgipfel für nachhaltige Entwicklung“ in Johannesburg, sowie die Fischerei-Reform 2013 und die davor stattfindenden Diskussionen in den Kookkurrenzen gefunden werden. Vor

allem im Kontext der Tageszeitung TAZ werden eher einzelne Ereignisse näher beleuchtet, statt dass eine Wortveränderung nachvollzogen werden kann.

### 5.3. Topic Modelle

Selbst mit den Einschränkungen, die Topic Modelle haben, zeigt sich ein sehr ähnlicher Trend zu Word Embeddings und Kookkurrenzen. So verändert sich auch hier der Kontext - die Topics - bis zu den frühen 2000ern, gehen Mitte der 2000er auf ein Mindestmaß hinunter, zwischen 2012 und 2018 allerdings wieder hoch, wenn auch nicht auf die selbe Stufe, wie in den späten 90ern und frühen 2000ern.

Beim qualitativen Bewerten der Ergebnisse zeigen sich allerdings die Einschränkungen, die eine Analyse speziell auf Wortveränderung von „nachhaltig“ mit dieser Methode aufzeigen. Benennt man zum Beispiel die Topics des Bundestags der Jahre 1999-2001, kann die folgende Liste für 1999-2000 erstellt werden: Haushalt, Europa/Bundeswehr, Ausbildung, Gesetze, Landwirtschaft/Energie, Arbeit/Wohnungen, Rente/Versicherung, DDR-Geschichte, Rest1, Rest2<sup>24</sup>.

Schon hier fällt auf, dass mehrere Topics als Kombination unterschiedlicher Topics angesehen werden können. Es stellt sich auch die Frage, wie viele der Topics wirklich mit „nachhaltig“ zu tun haben.

Vergleicht man die Liste der Topics mit einer benannten Liste für 2000-2001, werden die Probleme noch deutlicher:

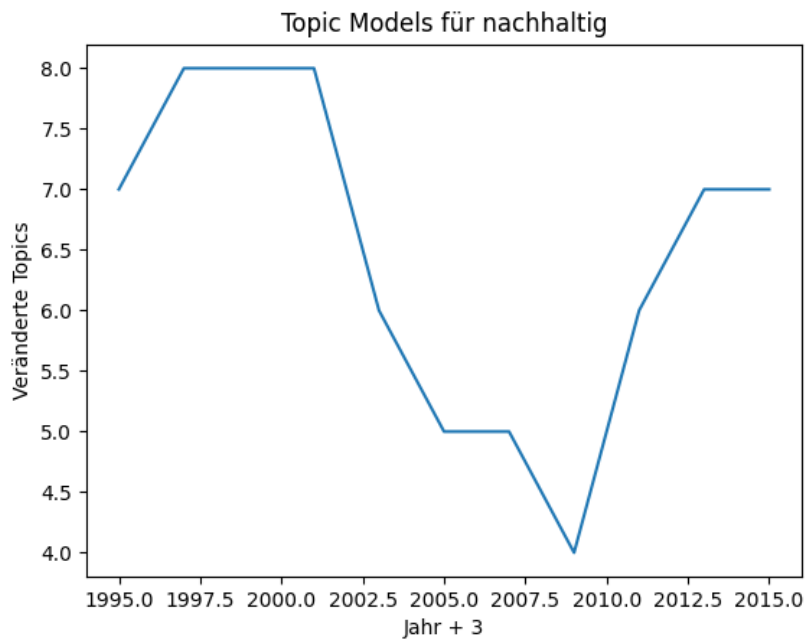
Haushalt, Gesetz, Familie/Arbeitsmarkt, Bildung/Klimaschutz, Europa/-Deutschland, Bundeswehr/Afghanistan, Landwirtschaft/Verkehr, Vornamen, Rest1, Rest2.

Mit  $p = 0.6$  haben sich zwischen den zwei Listen an Topics 8 Topics verändert (vgl. Abbildung 5.9). Schaut man sich die qualitative Veränderung an, kommt man auf leicht unterschiedliche Ergebnisse. Vollständig neu ist nur ein Topic, Vornamen, welches ausschließlich aus Vornamen besteht und damit keinen Mehrwert besitzt. Eine Aufteilung und Zusammenführung unterschiedlicher Topics fand in 5 Fällen statt. So wurde Europa/Bundeswehr zu Europa/Deutschland und Bundeswehr/Afghanistan, mit dem Start des Afghanistan-Krieges 2001. Diese Veränderung kann dabei aber nicht wirklich einer Veränderung der Wortbedeutung von „nachhaltig“ zugeordnet werden.

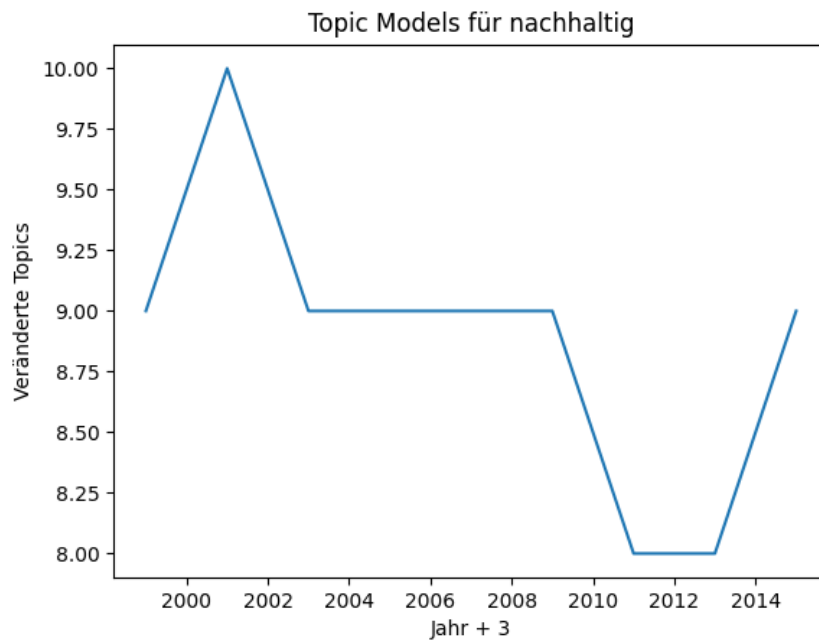
---

<sup>24</sup> Rest1 und Rest2 sind Topics aus Restwörtern, die keinem Thema richtig zugeordnet werden konnten.

### 5.3 Topic Modelle



**Abbildung 5.9.:** Anzahl veränderter Topics für den Bundestag-Datensatz zwischen 1995 und 2018,  $p=0.6$



**Abbildung 5.10.:** Anzahl veränderter Topics für den TAZ-Datensatz zwischen 1999 und 2018,  $p=0.6$

## 5.4. Vergleich der Ergebnisse

Die Ergebnisse der Word Embeddings sind detailliert und aufschlussreich, was durch den weitaus größeren Datensatz und die Einjahresschritte zu erwarten war. So lässt sich hier eine eindeutige Entwicklung der Verwendung von „nachhaltig“ erkennen. Trotz unterschiedlichem Datensatz lässt sich dieser Trend jedoch grob sowohl durch Kookkurrenzen mit dem Bundestag-Datensatz, als auch durch Topic Modelle nachvollziehen.

Word Embeddings geben Hinweise auf den semantischen Wandel, jedoch zeigt sich vor allem im Vergleich mit Kookkurrenzen eine Einschränkung: So weisen Kookkurrenzen stärker auf den Einfluss punktweiser Ereignisse hin, als auf eine tatsächliche Veränderung in der Bedeutung des Wortes. Vor allem der Kontext, in dem das Wort genutzt wird, bleibt ohne Großereignisse recht konstant.

Auch Topic Modelle legen zwar eine große Veränderung im Kontext dar, bieten aber bei genauerer, qualitativer Betrachtung ein uneindeutigeres Bild. „Nachhaltig“ wird lange im Kontext der Landwirtschaft genutzt, jedoch auch im Kontext von Umweltschutz und Wirtschaft allgemein. Die größten Veränderungen finden auch hier bei Ereignissen statt.

Großereignisse an sich ändern zwar den Kontext eines Wortes, aber nicht automatisch seine Bedeutung. Vor allem die hier untersuchten Zeitabschnitte zeigen keine signifikante Änderung in der Bedeutung des Wortes, nur in seiner Verwendung. Es kann weder von einer Eingrenzung noch einer Generalisierung gesprochen werden.

Trotz dieses Ergebnisses bieten die Word Embeddings das größte Potential, Wortbedeutungsveränderungen nachzuvollziehen. Kookkurrenzen sind nützlich und sogar notwendig, um Veränderungen qualitativ zu analysieren, reagieren aber in der quantitativen Analyse zu stark auf einzelne Ereignisse. Dagegen bieten Topic Modelle ein größeres Potential für die quantitative Analyse, zeigen aber andere Probleme auf, wie die Kombination und Aufteilung von Themen zu unterschiedlichen Zeitpunkten, die als Veränderung angesehen werden, jedoch qualitativ wenig Unterschiede zeigen, und die Existenz von Topics, die wenig mit dem untersuchten Wort gemein haben. Ein qualitativer Vergleich der Ergebnisse ist dabei ebenfalls sehr aufwendig.

Es bietet sich also vor allem an, das Potenzial von Word Embeddings für semantischen Wandel weiter zu erforschen, sowie weitere Methoden damit zu vergleichen.

## 6. Diskussion

In dieser Arbeit wurde versucht, den semantischen Wandel des Wortes „nachhaltig“ durch computergestützte quantitative Methoden nachzuvollziehen. Dabei wurde sich auf die These gestützt, dass die Bedeutung eines Wortes zu einem bestimmten Zeitpunkt unabhängig der genutzten Methode vom Kontext des Wortes im geschriebenen Korpus abhängig ist [Tahmasebi et al., 2021, S. 17].

Drei Methoden wurden angewandt und miteinander verglichen: Word Embeddings, zeitliche Kookkurrenzen und Topic Modelle. Für die Quantifizierung des Wandels zwischen einzelnen Zeitpunkten wurden für Word Embeddings sowohl die Kosinus-Distanz, als auch der Neighbourhood Shift berechnet. Für zeitliche Kookkurrenzen wurden die Kookkurrenzen als Vektorraum betrachtet und ebenfalls die Kosinus-Distanz berechnet. Für Topic Modelle wurde berechnet, wie viele der Topics sich zwischen zwei Zeitpunkten um mehr als 60% veränderten.

Trotz unterschiedlichen Datensätzen ließ sich sowohl bei Word Embeddings, als auch bei Topic Modellen eine ähnliche Kurve erkennen. Kontext veränderte sich bis in die frühen 2000er-Jahre stark, ließ dann etwas nach, bevor ein geringerer Kontextwandel nach 2015 einsetzte. Kookkurrenzen zeigten je nach Datensatz unterschiedliche Ergebnisse, allerdings ließen sich an ihnen vor allem Großereignisse ablesen.

Es wäre möglich, die Ergebnisse von Word Embeddings und Topic Modellen als semantischen Wandel zu interpretieren. Da „nachhaltig“ jedoch in einigen Themengebieten (wie z.B. Landwirtschaft) über den gesamten Zeitraum durchgehend genutzt wurde und die größten Veränderungen vor allem um Jahre sichtbar sind, in denen Großereignisse stattfanden (wie der Weltgipfel für nachhaltige Entwicklung in Johannesburg), ist eine so eindeutige Aussage schwierig. Ein Wandel von Kontexten ist so ablesbar, es benötigt jedoch einer linguistischen, qualitativen Analyse, um festzustellen, ob es sich tatsächlich um semantischen Wandel handelt.

Das Gebiet der computergestützten quantitativen Analyse semantischen Wandels ist jung. Es lohnt sich, weitere Methoden, wie z.B. Kontextvektoren, andere Varianten von Embeddings, dynamische Topic Modelle oder Word Sense Induction genauer zu untersuchen und mit den hier vorgestellten Methoden zu



---

vergleichen. Dabei bietet es sich an, näher mit Linguisten zusammenzuarbeiten und gemeinsam Kriterien für die Untersuchung der Qualität einer Methode zu entwickeln. Ein Vergleich unterschiedlicher Methoden kommt ansonsten selten vor. Oft unterscheiden sich die Methoden der Evaluation selbst so stark, dass ein Vergleich schwierig ist [Tahmasebi et al., 2021, S. 70ff].

Mit Standards der Evaluation können Methoden des Text Minings sehr gut als Werkzeuge genutzt werden, um Thesen über den semantischen Wandel bestimmter Worte zu unterstützen und mit quantitativen Argumenten zu untermauern.

## 7. Fazit

Die Bedeutung von Wörtern ändert sich über Jahre hinweg. Der Kontext der Nutzung verändert sich. Die Bedeutung kann sich erweitern, oder auf ein bestimmtes Verständnis einschränken, wenn sie sich nicht sogar komplett verändert. Diese Arbeit untersuchte den semantischen Wandel des Wortes „nachhaltig“ von den 90er bis zu den späten 2010er Jahren.

Durch die Untersuchung und den Vergleich von Word Embeddings, Kookkurrenzen und Topic Modelle auf Zeitungsdaten und Reden aus dem Bundestag konnten Veränderungen im Kontext von „nachhaltig“ untersucht werden, die auf einen semantischen Wandel hinweisen.

Es konnten sowohl mit Word Embeddings, als auch mit Topic Modellen eine starke Veränderung im Kontext des Wortes vor 2005 festgestellt werden, sowie ein leichter Anstieg der Veränderung nach 2015. Auch die Kookkurrenzen der Reden aus dem Bundestag zeigten ein ähnliches Ergebnis, auch wenn die größte Veränderung dort recht eindeutig mit dem „Weltgipfel für nachhaltige Entwicklung“ in Johannesburg korreliert. Die Kookkurrenzen der TAZ unterschieden sich dagegen stark, die größte Veränderung korrelierte mit der Fischereireform der EU 2013.

Die ursprünglich gestellten Forschungsfragen können also so beantwortet werden:

1. Die Untersuchung des Kontext-Wandels des Wortes „nachhaltig“ durch die vorgestellten Methoden ist sehr gut möglich.
2. Die Frage, ob „nachhaltig“ in diesen Jahren einen substanziellen semantischen Wandel vollzogen hat, lässt sich dagegen weniger eindeutig beantworten.
3. Die besten Ergebnisse wurden durch Word Embeddings erzielt, wobei der detailliertere Datensatz Einfluss gehabt haben könnte. So ließ sich eine hohe Veränderung von den 90er-Jahren bis ca. 2005/2006 feststellen, die später stark nachließ.
4. Zeitliche Kookkurrenzen reagierten sehr empfindlich auf einzelne Ereignisse und Kontext und waren daher schwerer für den semantischen Wan-

---

del zu interpretieren. Topic Modelle erzielten ein ähnliches Ergebnis wie Word Embeddings, oft veränderten sich aber Topics uneindeutig, zum Beispiel indem bei zwei Topics mit drei Themen sich ein Thema in das andere Topic verschob, was eine große Veränderung zweier Topics zur Folge hatte, aber qualitativ keinen Unterschied machte.

5. Insgesamt ließ sich feststellen, dass der Kontext in vielen Feldern gleich blieb oder sich auf einzelne Ereignisse bezog. Selbst Umweltschutz kommt im Zusammenhang mit „nachhaltig“ schon 1995 vor. Kontextveränderungen können also festgestellt werden, aber Arten des Bedeutungswandels (zum Beispiel Eingrenzung oder Generalisierung) können nicht eindeutig zugeschrieben werden.

### 7.0.1. Limitierungen

Neben den bereits genannten Problemen bei der Bestimmung semantischen Wandels durch Methoden des Text Minings gibt es noch weitere Einschränkungen. So wird zum Beispiel von [Tahmasebi et al. \[2021\]](#) notiert, dass die Messung von linguistischen oder kulturellen Veränderungen der Bedeutung eines Wortes dadurch limitiert ist, dass diese Bedeutung nicht von anderen Wörtern unabhängig ist. Jedoch basieren die hier vorgestellten Methoden (und die meisten anderen Methoden des Text Minings) darauf, dass kookkurrente Worte oder nächste Nachbarn stabile, unveränderte Bedeutungen über den gleichen Zeitraum haben. Auch der Kontext im Allgemeinen kann sich je nach Datensatz stark verändern.

Insbesondere, wenn ein Wort älter als die vorhandenen Daten ist, wird es schwierig, eine genauere quantitative Analyse durchzuführen. Zum Beispiel war es anhand der vorhandenen Daten unmöglich herauszufinden, wann die Generalisierung von „nachhaltig“ von der Forstwirtschaft auf eine allgemeinere Bedeutung stattfand. Auch die Nutzung der Worte in unterschiedlichen Datensätzen können unterschiedliche Ergebnisse erzielen. Dies war hier schon durch den Unterschied zwischen dem Datensatz des Bundestags und der TAZ in zeitlichen Kookkurrenzen erkennbar. Insgesamt wurden in dieser Arbeit nur die Bereiche der Politik und des Journalismus betrachtet, mit der Hoffnung, damit genügend Informationen über den allgemeinen Gebrauch zu finden. Trotzdem bieten diese Korpora keinen Überblick über die Nutzung des untersuchten Wortes in der gesamten deutschen Sprache und insbesondere nicht über den Nutzen ähnlicher Worte in anderen Sprachen, zum Beispiel „sustainable“ aus dem Englischen.

[Tahmasebi et al. \[2021\]](#) schlägt daher auf S. 68 eine Reihe weiterer Konditionen vor, die für eine Evaluation notwendig sind. Darunter fallen unter anderem

---

auch die Überprüfung von Ergebnissen anhand Datensatz-unabhängigen Wörterbüchern und die Suche nach Art und Zeitpunkt von semantischem Wandel selbst in Studien, die diese nicht direkt messen sollen.

### **7.0.2. Ausblick**

Es lässt sich hoffen, dass die Zusammenarbeit von Linguisten, die semantischen Wandel untersuchen, und Informatikern, die Modelle für die Messung von semantischem Wandel erstellen, weiter erstarkt. Durch Beiträge beider Wissenschaften können besser Methoden entwickelt werden, welche die Forschungsergebnisse möglichst objektiv beurteilen. Während das Forschungsgebiet in den letzten Jahren enorm wuchs, wurden noch keine wirklich allgemeingültigen Standards für die Evaluation aller Werke entwickelt. Diese Arbeit soll einen Beitrag dazu geleistet haben, Methoden zu beurteilen, die semantischen Wandel von spezifischen, ausgesuchten Wörtern messen können und dabei sowohl Schwierigkeiten, als auch Chancen für diesen Ansatz dargestellt haben.

# Anhang A.

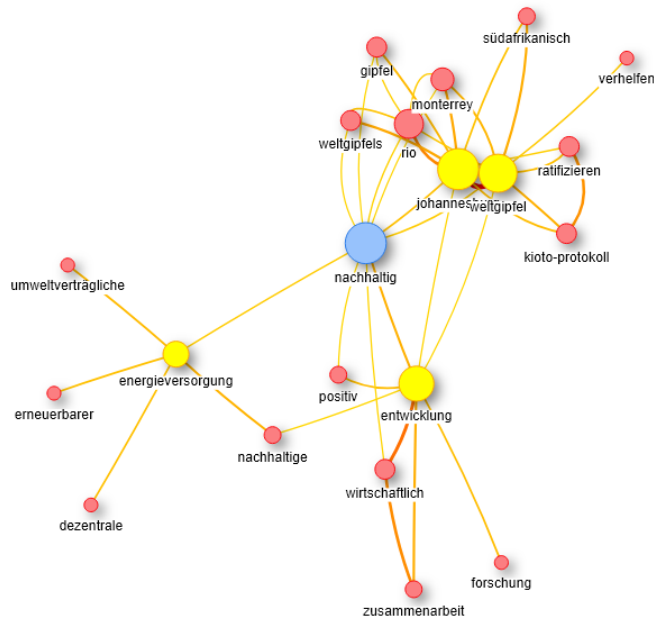
## Anhang

### A.1. Links zum Quellcode

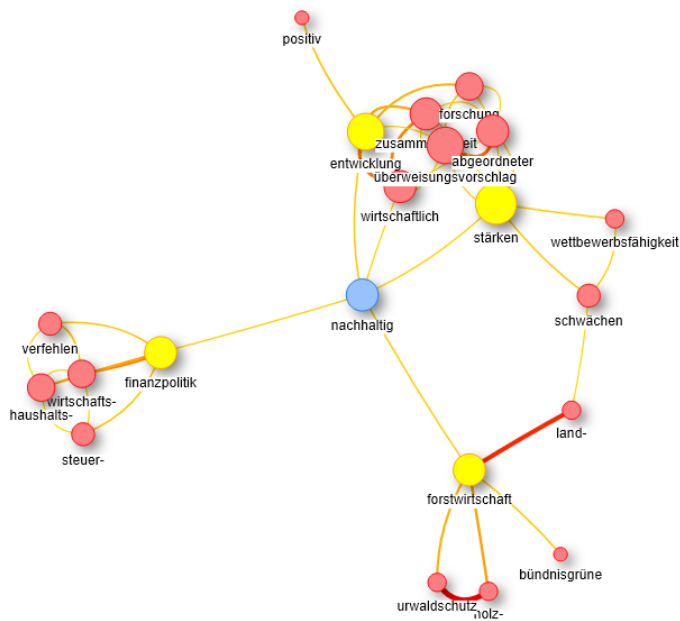
1. Word2Vec:  
<https://git.informatik.uni-leipzig.de/gh67kagu/word2vec>
2. Kookkurrenzen (Auswertung):  
<https://git.informatik.uni-leipzig.de/gh67kagu/kookkurrenzen>
3. Topic Modelle (Auswertung und Berechnung veränderter Topics):  
<https://git.informatik.uni-leipzig.de/gh67kagu/topic-models>

### A.2. Zusätzliche Grafiken

## A.2 Zusätzliche Grafiken

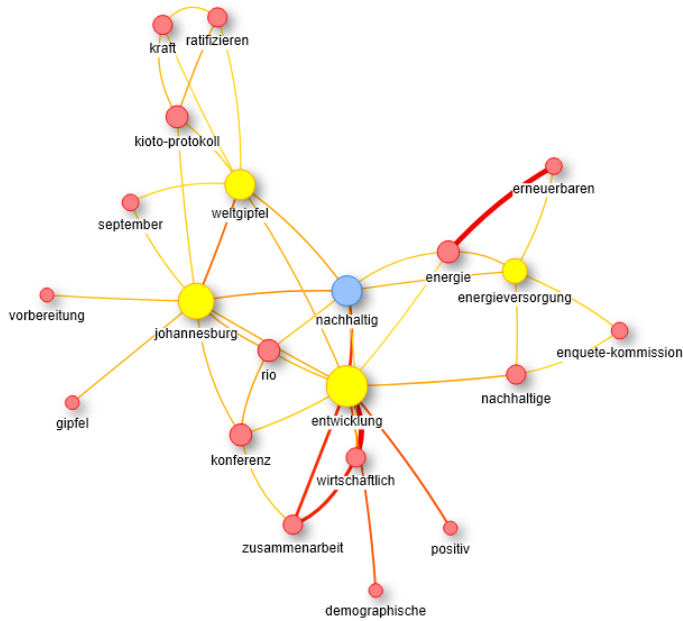


**Abbildung A.1.:** Kookkurrenzen der Dice-Signifikanz für den Bundestag-Datensatz 2001/2002

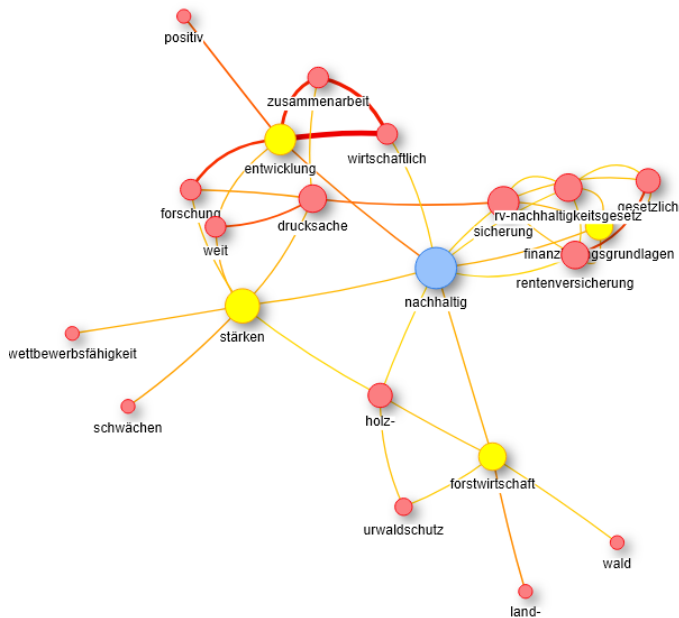


**Abbildung A.2.:** Kookkurrenzen der Dice-Signifikanz für den Bundestag-Datensatz 2003/2004

## A.2 Zusätzliche Grafiken



**Abbildung A.3.:** Kookkurrenzen der Log-Likelihood-Signifikanz für den Bundestag-Datensatz 2001/2002



**Abbildung A.4.:** Kookkurrenzen der Log-Likelihood-Signifikanz für den Bundestag-Datensatz 2003/2004

## A.2 Zusätzliche Grafiken

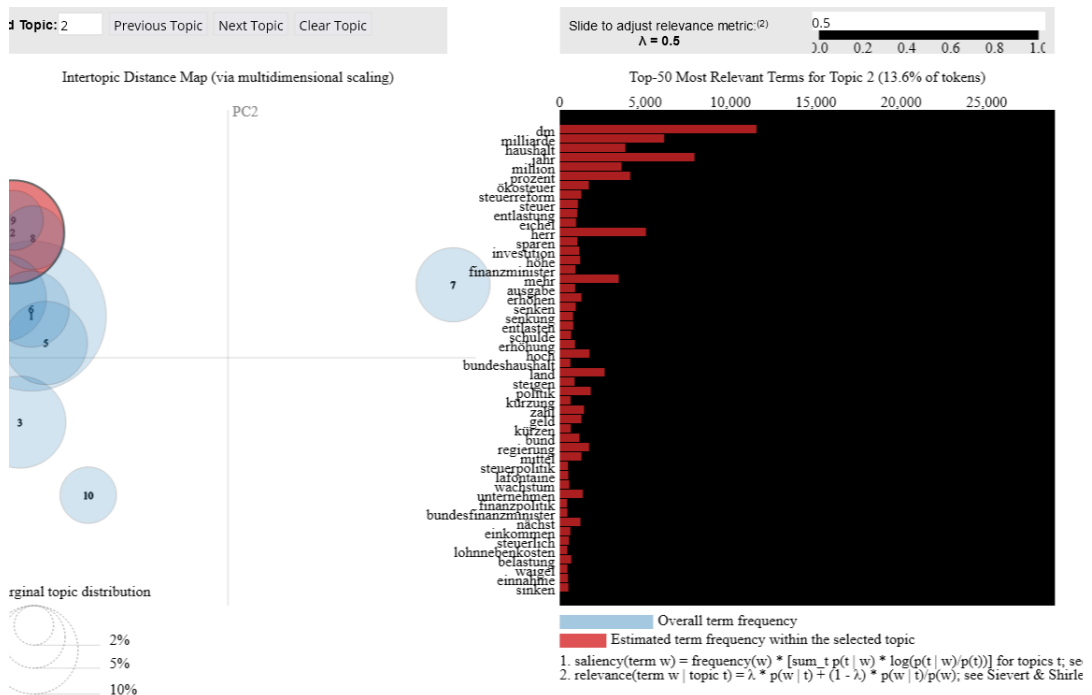


Abbildung A.5.: Beispiel Topic „Haushalt“ Bundestag-Datensatz 1999/2000



# Literaturverzeichnis

- [1] Robert Bamler and Stephan Mandt. Dynamic word embeddings. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 380–389. PMLR, 06–11 Aug 2017. URL <https://proceedings.mlr.press/v70/bamler17a.html>. 8
- [2] Oren Barkan. Bayesian neural word embedding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb. 2017. URL <https://ojs.aaai.org/index.php/AAAI/article/view/10987>. 8
- [3] Pierpaolo Basile, A. Caputo, Roberta Luisi, and Giovanni Semeraro. Diachronic analysis of the italian language exploiting google ngram. In *CLiC-it/EVALITA*, 2016. 7
- [4] D Blei. Introduction to probabilistic topic models. *Communications of the ACM*, 55, 2011. 15
- [5] D Blei and J Lafferty. Correlated topic models. inweiss, y., schölkopf, b., and platt, j., editors. *Advances in neural information processing systems*, 18, 2006. 7
- [6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003. 15
- [7] John A Bullinaria and Joseph P Levy. Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and svd. *Behavior research methods*, 44(3):890–907, 2012. 7
- [8] William Croft. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press, 10 2001. URL <https://doi.org/10.1093/acprof:oso/9780198299554.001.0001>. 6
- [9] Romain Deveaud, Eric SanJuan, and Patrice Bellot. Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, 17(1):61–84, 2014. 15

- [10] Elizabeth Closs Traugott. Semantic change, 2017. URL <https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-323>. 1, 3, 4
- [11] Mirella Frermann, Lea und Lapata. A Bayesian Model of Diachronic Meaning Change. *Transactions of the Association for Computational Linguistics*, 4: 31–45, 2016. URL [https://doi.org/10.1162/tacl\\_a\\_00081](https://doi.org/10.1162/tacl_a_00081). 8
- [12] Ulrich Grober. Urtexte – carlowitz und die quellen unseres nachhaltigkeitsbegriffs. *Natur und Landschaft*, 2, 2013. 1
- [13] Kristina Gulordava and Marco Baroni. A distributional similarity approach to the detection of semantic change in the Google Books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK, July 2011. Association for Computational Linguistics. URL <https://aclanthology.org/W11-2508>. 8
- [14] William L. Hamilton, Jure Leskovec, and Dan 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, August 2016. Association for Computational Linguistics. URL <https://aclanthology.org/P16-1141>. 11
- [15] Martin Hilpert. *Germanic Future Constructions: A usage-based approach to language change*. John Benjamins, 2008. URL <https://www.jbe-platform.com/content/books/9789027291035>. 6
- [16] Christian Kahmann., Andreas Niekler., and Gerhard Heyer. Detecting and assessing contextual change in diachronic text documents using context volatility. In *Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KDIR,,* pages 135–143. INSTICC, SciTePress, 2017. 7
- [17] Daphné Kerremans, Susanne Stegmayr, and Hans-Jörg Schmid. *The Neo-Crawler: identifying and retrieving neologisms from the internet and monitoring ongoing change*, pages 59–96. De Gruyter Mouton, Berlin, Boston, 2012. URL <https://doi.org/10.1515/9783110252903.59>. 6
- [18] Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. Word sense induction for novel sense detection. In *Proceedings*

- of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 591–601, Avignon, France, April 2012. Association for Computational Linguistics. URL <https://aclanthology.org/E12-1060>. 7, 8
- [19] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the association for computational linguistics*, 3:211–225, 2015. 7
- [20] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. URL <https://arxiv.org/abs/1301.3781>. 12
- [21] Andreas Niekler, Arnim Bleier, Christian Kahmann, Lisa Posch, Gregor Wiedemann, Kenan Erdogan, Gerhard Heyer, and Markus Strohmaier. ILCM - a virtual research infrastructure for large-scale qualitative data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL <https://aclanthology.org/L18-1209>. 14
- [22] Michael Rödel. Die invasion der nachhaltigkeit": Eine linguistische analyse eines politischen und ökonomischen modeworts. *Deutsche Sprache : ds ; Zeitschrift für Theorie, Praxis, Dokumentation*, 2:115–141, 2013. URL <http://www.dsdigital.de/ce/die-invasion-der-nachhaltigkeit/detail.html>. 2
- [23] Martina A Rodda, Marco Senaldi, and Alessandro Lenci. Panta rei: Tracking semantic change with distributional semantics in ancient greek. *Italian Journal of Computational Linguistics*, 3:11–24, 06 2017. 7
- [24] David Rudolph, Maja und Blei. Dynamic embeddings for language evolution. Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. ISBN 9781450356398. URL <https://doi.org/10.1145/3178876.3185999>. 8
- [25] Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. Room to Glo: A systematic comparison of semantic change detection approaches with word 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 66–76, Hong Kong, China, November 2019. Association for Computational Linguistics. URL <https://aclanthology.org/D19-1007>. 8, 11, 13, 14

- [26] Carson Sievert and Kenneth Shirley. Ldavis: A method for visualizing and interpreting topics. 2014. 16
- [27] Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen. *Computational approaches to semantic change*. Language Science Press, Berlin, June 2021. URL <https://doi.org/10.5281/zenodo.5040241>. 4, 6, 7, 28, 29, 31
- [28] Xuri Tang, Weiguang Qu, and Xiaohe Chen. Semantic change computation: A successive approach. In Longbing Cao, Hiroshi Motoda, Jaideep Srivastava, Ee-Peng Lim, Irwin King, Philip S. Yu, Wolfgang Nejdl, Guandong Xu, Gang Li, and Ya Zhang, editors, *Behavior and Social Computing*, pages 68–81, Cham, 2013. Springer International Publishing. 7
- [29] Peter D Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188, 2010. 7
- [30] Jing Wang, Mohit Bansal, Kevin Gimpel, Brian D. Ziebart, and Clement T. Yu. A sense-topic model for word sense induction with unsupervised data enrichment. *Transactions of the Association for Computational Linguistics*, 3: 59–71, 2015. URL <https://aclanthology.org/Q15-1005>. 7
- [31] Xuerui Wang and Andrew McCallum. Topics over time: A non-markov continuous-time model of topical trends. volume 2006, pages 424–433, 01 2006. 7
- [32] Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. Discovery of evolving semantics through dynamic word embedding learning. *CoRR*, abs/1703.00607, 2017. URL <http://arxiv.org/abs/1703.00607>. 8

# Erklärung

Ich versichere, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe, insbesondere sind wörtliche oder sinngemäße Zitate als solche gekennzeichnet. Mir ist bekannt, dass Zuwiderhandlung auch nachträglich zur Aberkennung des Abschlusses führen kann. Ich versichere, dass das elektronische Exemplar mit den gedruckten Exemplaren übereinstimmt.

Ort:

Datum:

Unterschrift: