Optimierung der Suchergebnisse bei Swisslex. Digitale Lösungsansätze der Angewandten Linguistik zur Modellierung von Semantik

Christian KRIELE & Philipp DREESEN
Zürcher Hochschule für Angewandte Wissenschaften
Departement Angewandte Linguistik
Theaterstrasse 15c, 8401 Winterthur, Schweiz
christian.kriele@zhaw.ch; philipp.dreesen@zhaw.ch

This article is about a project that aimed at optimizing the search results of a legal information platform. The project focused on the question of which linguistic phenomena have an influence on the "recall" and "precision" factors and how the search results can be optimized using applied linguistics methods. The article begins by presenting the state of the search query at the start of the project and the resulting shortcomings. Subsequently, the structure of the search query optimization project is outlined: In two subprojects, independent solutions were generated and validated. The first solution was terminologically oriented, the second solution made use of corpus linguistic methods. The terminologically oriented subproject tried to answer the question to what extent the search query can be optimized using a knowledge system. The second solution investigated on how corpus linguistic approaches such as the creation of word embeddings and collocation profiles could be used to achieve this goal. The article concludes by discussing a possible synthesis of the before mentioned two approaches and the way the implementation of the two methods in the search engine could enhance the search experience of the users.

Keywords:
search engine optimization, recall, precision, semantic relations, thesaurus, distributional semantics.

Stichwörter:
Suchmaschinenoptimierung, Ausbeute, Präzision, semantische Relationen, Thesaurus, distributionelle Semantik.

1. Einleitung

Anlass für das in diesem Beitrag vorgestellte Projekt war ein sehr konkretes Vorhaben: die Optimierung der Suchergebnisse der Dokumentensuchmaschine von Swisslex[1], der kommerziellen Anbieterin einer Rechtsinformationsplattform im Markt und Rechtsraum Schweiz. Swisslex stellt Juristen Rechtsinformationen in einer Datenbank zur Verfügung. Dazu gehören Urteilssammlungen, Fachzeitschriften, Gesetzeskommentare und Werke aus der Fachliteratur. Das von Swisslex finanzierte Projekt wurde an der Zürcher Hochschule für Angewandte Wissenschaften in Zusammenarbeit mit dem Büro b3, einem Beratungs- und Dienstleistungsunternehmen im Bereich Übersetzungs-, Terminologie- und Wissensmanagement, durchgeführt.[2] Ziel

1 https://www.swisslex.ch/

2 Wir danken im Projekt Felix Steiner, Noah Bubenhofer und Selena Calleri.

des Projektes war es zu evaluieren, ob die Ergebnisse der Suchmaschine mit Mitteln der Angewandten Linguistik optimiert werden können.

Das Projekt wurde transdisziplinär durchgeführt, d. h. der ausser-wissenschaftliche Praxispartner und die Angewandte Linguistik haben gemeinsam eine Fragestellung der Suchoptimierung bearbeitet (vgl. Perrin & Kramsch 2018). Transdisziplinarität bedeutet, nicht nur für, sondern mit einem Praxispartner zu arbeiten. Entsprechend ist es erforderlich zu verstehen, welche Erwartungen, Möglichkeiten und Grenzen auf beiden Seiten mit dem Projektziel verknüpft werden. Im Projekt wurde vereinbart, die linguistische Suchmaschinenoptimierung zu Testzwecken zu entwickeln und Anwendungs-perspektiven für die bestehende Suchmaschinenumgebung aufzuzeigen. Hierfür wurden Testabfragen simuliert.

Suchoptimierung ist bisher kein ausgewiesenes Einsatzgebiet der Angewandten Linguistik. Es ist jedoch festzustellen, dass *Information Retrieval* (d. h. die Gewinnung von Informationen) ein semiotischer, insbesondere ein stark textbasierter Ansatz der Informationswissenschaft ist (Stock & Stock 2015). Entsprechend hat auch Swisslex bisher insbesondere mit IT-Lösungen versucht, beispielsweise Mehrsprachigkeit, Mehrdeutigkeit, veränderte Rechtsschreibung und Flexion in die Suchabfragen zu integrieren (Erbguth 2015).

Das Information Retrieval bei Swisslex basiert derzeit auf einer Volltextsuche mit diversen auf einer Nomenklatur basierenden Filtermöglichkeiten und Wortvorschlägen, die in einer Dropdown-Liste erscheinen. Bei den Wortvorschlägen handelt es sich um Wörter, die im selben Text wie der Suchbegriff vorkommen und mit demjenigen Buchstaben beginnen, der nach dem Suchbegriff eingegeben wird, siehe Abb. 1.

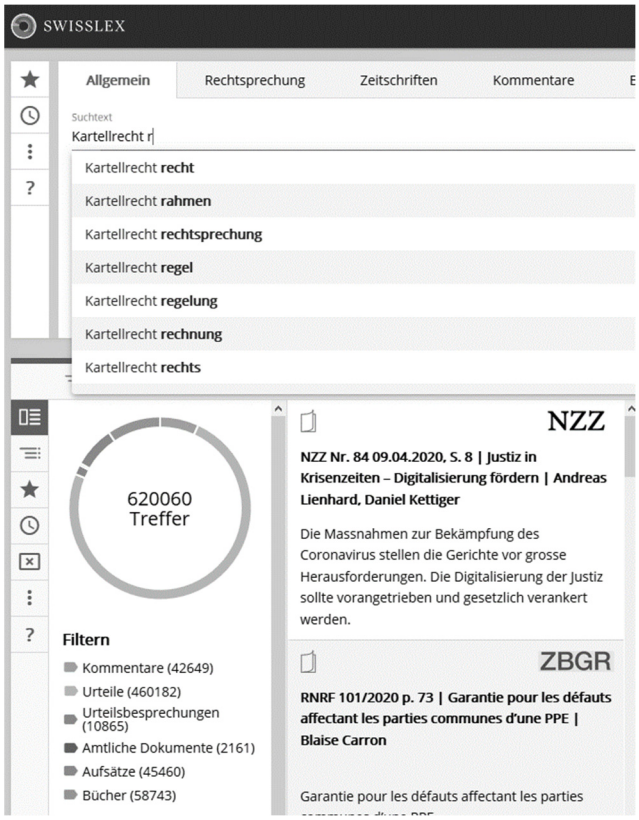

Abb. 1: Suchoberfläche von Swisslex

1.1 Precision und Recall

Für die spezifischen Bedürfnisse der Suche durch Experten bei Swisslex ist die Balance zwischen den beiden voneinander abhängigen Faktoren *Precision* (Präzision) und *Recall* (Ausbeute) besonders relevant. Was jedoch wird in der Informationswissenschaft unter Precision und Recall genau verstanden? Der Recall sagt etwas darüber aus, wie viele der in der Datenbank vorhandenen relevanten Dokumente gefunden wurden – ins Verhältnis gesetzt zur Anzahl aller relevanten Dokumente in der Datenbank. Die Precision setzt jene Zahl ins Verhältnis zur Zahl der insgesamt gefundenen Dokumente, sie gibt an, wie viele der gefundenen Dokumente relevant sind. Die Ergebnisse zu einer Suchanfrage sollten möglichst präzise sein, also möglichst nur relevante Treffer enthalten. Gleichzeitig sollte der Recall möglichst hoch sein. Es sollten also alle Dokumente gefunden werden, die relevant sind.

Aus linguistischer Perspektive sind dabei zwei Phänomene zu beobachten, die sich auf Recall und Precision auswirken können:

- Eine Benennung (bzw. mehrere Benennungen mit gleicher Form) repräsentiert in manchen Fällen unterschiedliche Begriffe (Ambiguität) (Drewer & Schmitz 2017: 17). Ein Beispiel für Ambiguität im Rechtskontext ist die Benennung *Wettbewerb*, unter der einerseits ein sportlicher Wettbewerb und andererseits verschiedene Formen von wirtschaftlichem

Wettbewerb verstanden werden können. Bei einer Suchanfrage führen ambige Benennungen zu einer schlechten Precision, da unter Umständen auch Dokumente gefunden werden, in denen die Benennung einen anderen Begriff repräsentiert (Kießling 2016: 8).

- Für einen Begriff, auf den mit einer sprachlichen Bezeichnung verwiesen werden soll, gibt es oft mehrere Benennungen (Synonymie) (Drewer & Schmitz 2017: 16). So existiert im Rechtskontext für die Benennung *Gebietsabrede* beispielsweise das Synonym *Gebietsabsprache*. Eine Suchanfrage, bei der die Synonyme des Suchbegriffs nicht beachtet werden, führt unter Umständen zu einem schlechten Recall, da diejenigen Dokumente nicht gefunden werden, in denen nur die Synonyme des Suchbegriffs vorkommen (Spremann & Bartmann 2013: 208).

1.2 Optimierung der Suchergebnisse mit linguistischen Mitteln

Die übergreifende Zielsetzung des Projekts bestand wie eingangs erwähnt darin, mit Mitteln der Angewandten Linguistik Möglichkeiten der Optimierung der Suchanfrage zu evaluieren und Swisslex entsprechende Empfehlungen abzugeben.

Konkretes Ziel der Optimierung aus linguistischer Sicht war vor dem oben genannten Hintergrund, Precision und Recall zu verbessern. Die Grundidee dabei war, dass sowohl Precision als auch Recall verbessert werden können, wenn der Nutzer dem ursprünglichen Suchbegriff weitere Suchbegriffe hinzufügt bzw. wenn er diesen durch einen zutreffenderen Suchbegriff ersetzt. Dafür wurden in zwei Teilprojekten unterschiedliche Ansätze verfolgt. Teilprojekt 1 beschäftigte sich mit dem Einsatz einer Ontologie bzw. eines Thesaurus, Teilprojekt 2 mit korpus- und computerlinguistischen Methoden. Dazu gehörte die Berechnung von Kollokationen (Wörter, die signifikant häufig in der sprachlichen Umgebung eines Wortes vorkommen) und von Word Embeddings (hier eingesetzt zur Berechnung von potenziell synonym gebrauchten Wörtern). Das Ziel beider Ansätze war es, 'intelligente' Wortvorschläge zu generieren. 'Intelligent' deshalb, weil es sich bei den vorgeschlagenen Benennungen nicht einfach nur um Benennungen handeln sollte, die zufällig im gleichen Text vorkommen, sondern um Benennungen, die auf der Grundlage korpuslinguistischer Berechnungen als relevant erachtet werden bzw. in einer semantischen Relation zum Suchbegriff stehen (Henrich 2008: 120).

Die linguistische Semantik ordnet die semantischen Relationen in bestimmter Weise (vgl. Cruse 1986: Kap. 4; Lyons 1977: Bd. 1, Kapitel 9), kommt aber nicht ohne eine Kategorie für vage semantische Relationen aus. Letzteres wird häufig mit dem Konzept der "Familienähnlichkeit" von Wittgenstein (1982, § 66) gelöst. Im vorliegenden Fall werden die semantischen Relationen verwendet, die auf der für das Erstellen von Thesauri relevanten Norm beruhen, wobei ebenfalls

eine abschliessende Sammelkategorie verwendet wird. In DIN 1463-1 werden dabei analog zu ISO 25964-1 (2011) folgende semantische Relationen unterschieden und definiert (DIN 1463-1 1987: 5/6):

a) **Äquivalenzrelation**
"Eine Äquivalenzrelation ist die Beziehung zwischen gleichwertigen Bezeichnungen (bedeutungsgleich oder bedeutungsähnlich), die zu einer Äquivalenzklasse zusammengeführt werden."

Beispiel: *Gebietsabrede* und *Gebietsabsprache*.

b) **Hierarchierelationen**
"Hierarchierelationen liegen vor, wenn zwei Begriffe zueinander in einem Verhältnis der Über- bzw. Unterordnung stehen. Dabei sind zwei grundsätzlich unterschiedliche Formen der Hierarchierelationen zu unterscheiden."

- "Eine Abstraktionsrelation (generische Relation) ist eine hierarchische Relation zwischen zwei Begriffen, von denen der untergeordnete Begriff (Unterbegriff) alle Merkmale des übergeordneten Begriffs (Oberbegriff) besitzt und zusätzlich mindestens ein weiteres (spezifizierendes) Merkmal."

  Beispiel: *Anwalt* (Oberbegriff) und *Rechtsanwalt* (Unterbegriff)

- "Eine Bestandsrelation (partitive Relation) ist eine hierarchische Relation zwischen zwei Begriffen, von denen der übergeordnete (weitere) Begriff (Verbandsbegriff) einem Ganzen entspricht und der untergeordnete (engere) Begriff (Teilbegriff) einen der Bestandteile dieses Ganzen repräsentiert."

  Beispiel: *Recht* (Verbandsbegriff) und *Privatrecht* (Teilbegriff)

c) **Assoziationsrelation**
"Eine Assoziationsrelation ist eine zwischen Begriffen bzw. ihren Bezeichnungen als wichtig erscheinende Relation, die weder eindeutig hierarchischer Natur ist, noch als äquivalent angesehen werden kann."
Beispiel: *Rechtsanwalt* und *Mandant*

2. Terminologisch orientierte Verfahren

In der Folge wird auf das terminologisch orientierte Teilprojekt 1 eingegangen. Zunächst galt es dabei zu klären, mit welchen linguistischen Methoden Begriffe bzw. ihre Bezeichnungen anhand der oben genannten Relationen in Bezug zueinander gesetzt werden. Da dies sowohl in Thesauri als auch in Ontologien der Fall ist, war ein erstes Ziel des Teilprojektes zu ermitteln, welche der beiden so genannten Wissensordnungen sich für das Erreichen des Ziels des Projektes besser eignete.

2.1 Thesaurus

In DIN 1463-1 (DIN 1463-1 1987: 2) wird Thesaurus folgendermassen definiert:

"Ein Thesaurus im Bereich der Information und Dokumentation ist eine geordnete Zusammenstellung von Begriffen und ihren (vorwiegend natürlichsprachigen) Bezeichnungen, die in einem Dokumentationsgebiet zum Indexieren, Speichern und Wiederauffinden dient.

Er ist durch folgende Merkmale gekennzeichnet:

a) Begriffe und Bezeichnungen werden eindeutig aufeinander bezogen ("terminologische Kontrolle"), indem
   - Synonyme möglichst vollständig erfasst werden,
   - Homonyme und Polyseme besonders gekennzeichnet werden,
   - für jeden Begriff eine Bezeichnung (Vorzugsbenennung, Begriffsnummer oder Notation) festgelegt wird, die den Begriff eindeutig vertritt,

b) Beziehungen zwischen Begriffen (repräsentiert durch ihre Bezeichnungen) werden dargestellt."

In Thesauri werden in der Regel alle oben genannten Relationsarten verwendet.

Tab. 1 stellt die Abkürzungen für Relationen aus DIN 1463-1 vor, die auch im für Swisslex erstellten Thesaurus Anwendung finden (DIN 1463-1 1987: 11).

DIN 1463-1	
OB	Oberbegriff
UB	Unterbegriff
BF	Benutzt für
BS	Benutze Synonym
VB	Verwandter Begriff

Tab. 1: Abkürzungen für Relationen nach DIN 1463-1

2.2 Ontologie

In der Informatik und der KI-Forschung werden unter Ontologien computerlesbare Wissensmodellierungen verstanden (Drewer et al. 2017: 11). Die wichtigsten Bestandteile von Ontologien sind Klassen und Instanzen, die über Eigenschaften verfügen und miteinander über Relationen verbunden sind. Ebenso wie beim Thesaurus werden dabei sowohl hierarchische als auch assoziative Relationen verwendet. Im Gegensatz zum Thesaurus wird bei Ontologien jedoch explizit angegeben, um welche Art von assoziativer Relation es sich handelt (z. B. X "ist Mitglied von" Y, X "vertritt" Y). Im Unterschied zu Thesauri kommen Ontologien darüber hinaus sowohl bei Mensch-Maschine-Interaktionen als auch bei der Interaktion zwischen verschiedenen Maschinen zum Einsatz und lassen automatisches Schlussfolgern zu.

2.3 Optimierung der Swisslex-Suche: Thesaurus oder Ontologie?

Im Laufe des Projektes beschloss das Projektteam gemeinsam mit Vertretern von Swisslex, einen Thesaurus und keine Ontologie für die Optimierung der Swisslex-Suche einzusetzen. Der Hauptgrund dafür war, dass die Optimierungsidee auch durch einen Thesaurus realisiert werden konnte: Die in einem Thesaurus verwendeten Relationen führen zu Begriffen bzw. Bezeichnungen, die zum Suchbegriff entweder in einer hierarchischen, synonymen oder assoziativen Relation stehen. Dass in einem Thesaurus im Gegensatz zu Ontologien nicht explizit angegeben wird, um welche Art von assoziativer Relation es sich jeweils handelt, spielt für die Nutzergruppe keine wesentliche Rolle. Auch weitere Anwendungsszenarien von Ontologien wie beispielsweise das Ermöglichen von logischen Schlussfolgerungen standen nicht im Fokus von Swisslex, so dass der Nutzen einer Ontologie in keinem adäquaten Verhältnis zum Mehraufwand für das Erstellen und die Pflege einer solchen stand.

Vor diesem Hintergrund wurde beschlossen, mit dem von der Universität Rom entwickelten Tool VocBench[3] einen Pilotthesaurus zu erstellen. Zum Einsatz kam dabei SKOS (Simple Knowledge Organisation System)[4], eine auf dem Resource Description Framework (RDF)[5] und RDF-Schema (RDFS)[6] basierende formale Sprache zur Kodierung von Dokumentationssprachen. Um den Einfluss der linguistischen Methoden auf die Suchergebnisse im gegebenen Zeit- und Kostenrahmen adäquat prüfen zu können, musste auch der thematische Rahmen eingegrenzt werden. Als relativ gut abgrenzbarer Rechtsbereich wurde dabei das Kartellrecht erachtet.

2.4 Erstellen des Pilotthesaurus

Grundlage für das Erstellen des Pilotthesaurus war das Sammeln potenziell relevanter Termini. Zu diesem Zweck wurden aus einer von Swisslex aufbereiteten Textsammlung kartellrechtlich relevanter Texte mit verschiedenen Methoden (Termextraktions-, Konkordanz- und Korpusanalyseprogramme) Termini extrahiert. Zum Einsatz kamen dabei folgende Tools: extraterm[7], SynchroTerm[8], Antconc[9] und IMS Open Corpus Workbench[10]. Die Ergebnisse wurden in einer Liste vereint und durch Swisslex validiert. Bei diesem Validierungsschritt sortierten Swisslex-Mitarbeiter irrelevante Termini aus und

3 http://vocbench.uniroma2.it/

4 https://www.w3.org/2004/02/skos/

5 https://www.w3.org/2001/sw/wiki/RDF

6 https://www.w3.org/2001/sw/wiki/RDFS

7 https://extraterm.org/index.html

8 https://terminotix.com/index.asp?content=item&item=7&lang=en

9 http://www.laurenceanthony.net/software.html

10 http://cwb.sourceforge.net/

fügten weitere relevante Termini hinzu. Im Anschluss modellierte das Projektteam aus den relevanten Termini die hierarchische Grundstruktur des Pilotthesaurus in einer Mindmap. Für die Anreicherung der Grundstruktur um Synonyme, Ober- und Unterbegriffe bzw. verwandte Begriffe wurden dann mit Methoden der Korpus-/Computerlinguistik (Kollokationen und Word Embeddings) erneut Analysen im oben genannten Textkorpus durchgeführt. Die bei diesen Analysen ermittelten Termini wurden anschliessend wiederum durch Swisslex validiert, um sicherzustellen, dass eine juristische Perspektive vor allem in Hinblick auf die Auswahl von verwandten Begriffen gewährleistet war. Schliesslich erstellte das Projektteam den Pilotthesaurus auf Grundlage der oben genannten Grundstruktur und der validierten zusätzlichen Termini in Vocbench. Abb. 2 zeigt einen Beispieleintrag des Thesaurus mit der Vorzugsbenennung *Kartellrecht* (skosxl:prefLabel), dem Oberbegriff *Wettbewerbsrecht* (skos:broader), einer Definition (skos:definition) und diversen verwandten Begriffen (skos:related).

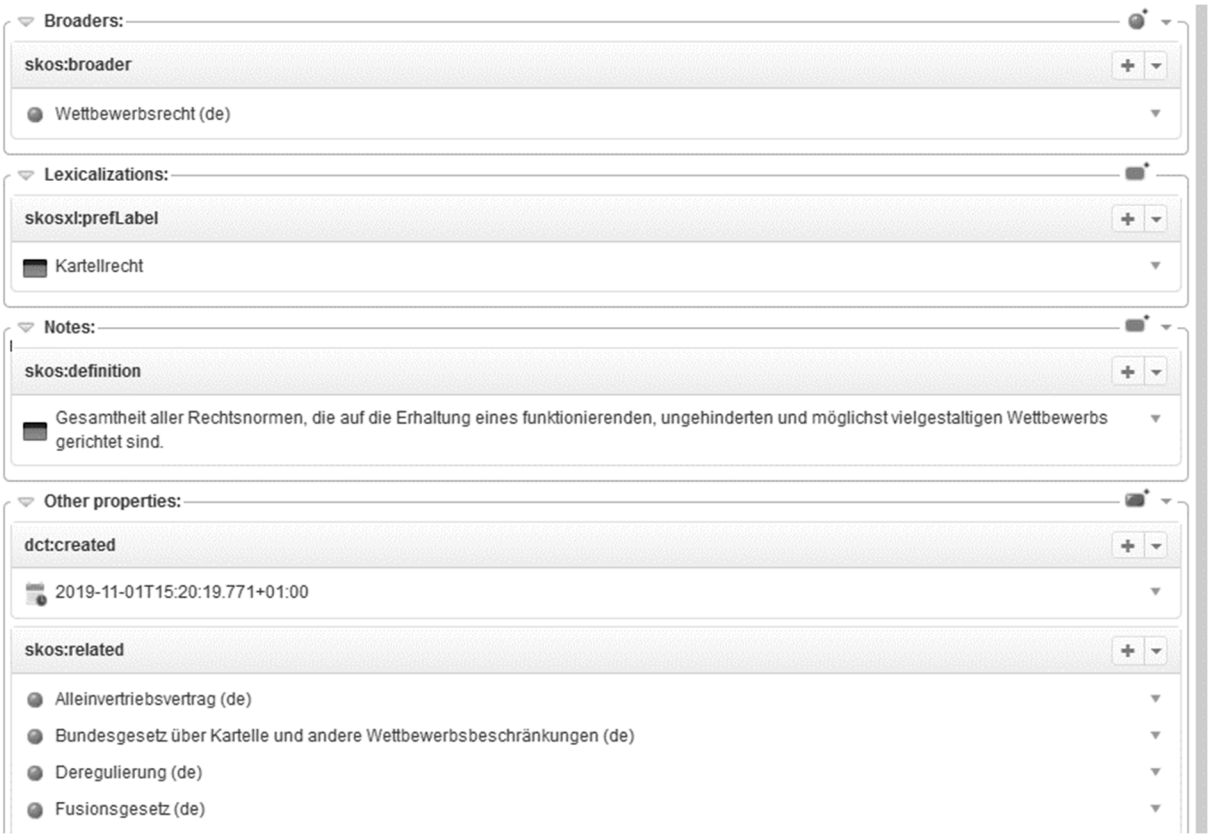

Abb. 2: Beispieleintrag im Pilotthesaurus

2.5 Anwendungsperspektiven des Thesaurus

Um zu prüfen, welche Auswirkungen ein Thesaurus auf die Suche bei Swisslex haben könnte, wurden drei Suchbegriffe ausgewählt, die sich in der Hierarchie des Thesaurus auf drei unterschiedlichen Ebenen befinden. Auf der obersten

Ebene befindet sich die Benennung *Kartellrecht*, auf einer mittleren Ebene *Preisabrede* und auf einer unteren Ebene *Gebietsschutz*, siehe Abb. 3.

Abb. 3: Ausschnitt aus der Hierarchie des Pilotthesaurus

Für alle drei Benennungen wurde in einer simulierten Suche ermittelt, welche Wortvorschläge der Thesaurus in Form von Synonymen, Ober- und Unterbegriffen und verwandten Begriffen bieten würde. In Tab. 2 finden sich beispielhaft die Untersuchungsergebnisse zur Benennung *Gebietsschutz*.

Anzumerken ist an dieser Stelle, dass der Thesaurus bei Abschluss des Projektes nicht erschöpfend befüllt war und bei einer entsprechenden Überarbeitung bzw. Erweiterung weitere Vorschläge in oben genannter Form hinzukämen. Zudem wurde der Thesaurus noch nicht in die Swisslex-Suche integriert. Die nachfolgend aufgeführten Ergebnisse sind also explorativer Natur, aber sie zeigen, welche Möglichkeiten ein vollständig gefüllter Thesaurus bei der Suche bieten würde. Da es in Vocbench derzeit nicht möglich ist, Einträge mit allen Informationen übersichtlich darzustellen, werden die Ergebnisse in der Folge tabellarisch aufgeführt. Damit die Darstellung übersichtlich bleibt, sind nur diejenigen Ober- und Unterbegriffe aufgeführt, die sich im Thesaurus jeweils auf der nächsten Ebene befinden.

Synonyme	**Ober- bzw. Unterbegriffe**	**verwandte Begriffe**
Gebietsabrede Gebietsabsprache	Alleinvertriebsvertrag (OB) absoluter Gebietsschutz (UB) Gebietsschutzklausel (UB)	Alleinbelieferungspflicht Alleinbezug Alleinbezugsbindung Alleinbezugspflicht Alleinbezugsverpflichtung Alleinbezugsabrede Alleinbezugsvertrag Alleinvertrieb Alleinvertriebshändler Alleinvertriebssystem Ausschliesslichkeitsbindung Bezugsbindung Gebietsbeschränkung Kundenbeschränkung Marktaufteilung Mengenvorgaben

Tab. 2: Informationen im Thesaurus zu *Gebietsschutz*

Wie aber könnten die Suchergebnisse nun anhand von diesen Informationen optimiert werden? Das Testbeispiel zeigt, dass eine Erhöhung des Recall bei der Suche nach *Gebietsschutz* beispielsweise erreicht werden könnte, indem dem Suchbegriff die Synonyme *Gebietsabrede* bzw. *Gebietsabsprache* hinzugefügt werden. Die Precision hingegen könnte erhöht werden, wenn der Suchbegriff *Gebietsschutz* beispielsweise durch einen Unterbegriff wie *absoluter Gebietsschutz* bzw. *Gebietsschutzklausel* ersetzt wird, falls dies die Sachverhalte sind, nach denen die entsprechende Person tatsächlich sucht. Die Precision könnte auch erhöht werden, indem dem Suchbegriff ein verwandter Begriff wie beispielsweise *Alleinbelieferungspflicht* oder *Alleinbezug* hinzugefügt wird.

3. Korpuszentrierte Verfahren

In einem komplementären Schritt wurde anhand eines Testdatensets exploriert, wie korpuszentrierte Methoden des Natural Language Processing und der

Korpuslinguistik den Suchprozess optimieren können (Stock & Stock 2015: 167-226, 275-300). Grundsätzlich werden hierfür linguistisch annotierte Textdaten verwendet, d. h. Texte mit grammatischen Informationen, semantischen Annotationen und Meta-Angaben aufbereitet und angereichert etc. (Krasselt et al. 2020). Für die Berechnung von Word Embeddings und Kollokationen ist die Annotation von Wortarten und Grundformen notwendig.

Im Projektverlauf haben sich zwei Methoden als zielführend für die Optimierung von Recall und Precision herausgestellt. Die Verfahren im Fall der Kollokationsberechnung und im Fall des Word Embeddings werden im Folgenden knapp dargestellt.

3.1 Word Embeddings

Im Rahmen der Distributionellen Semantik sind Word Embeddings eine verbreitete Methode, um in grossen Textkorpora ähnlich verwendete Ausdrücke automatisch zu identifizieren; dies wird mittlerweile auch in der Rechtswissenschaft eingesetzt (Landthaler 2016). Im vorliegenden Fall ist von word2vec ausgegangen worden (Mikolov et al 2013). Das Ziel der Berechnung von Word Embeddings ist die Anreicherung eines Ausdrucks mit potenziellen Synonymen, wobei ein weiter Synonymbegriff zugrunde liegt. Dabei ist aber wichtig zu betonen, dass es sich bei der Ausbeute nicht immer um Synonyme handelt, sondern auch um weitere semantische Relationen wie Antonyme, Hyperonyme oder einfach Wörter im ähnlichen semantischen Feld (Bubenhofer et al. 2019). Basis der Berechnung ist eine statistische Auswertung des Kontextverhaltens von Wörtern in Form von Vektorenberechnungen, oft mittels eines Machine-Learning-Algorithmus, der als neuronales Netz implementiert ist.

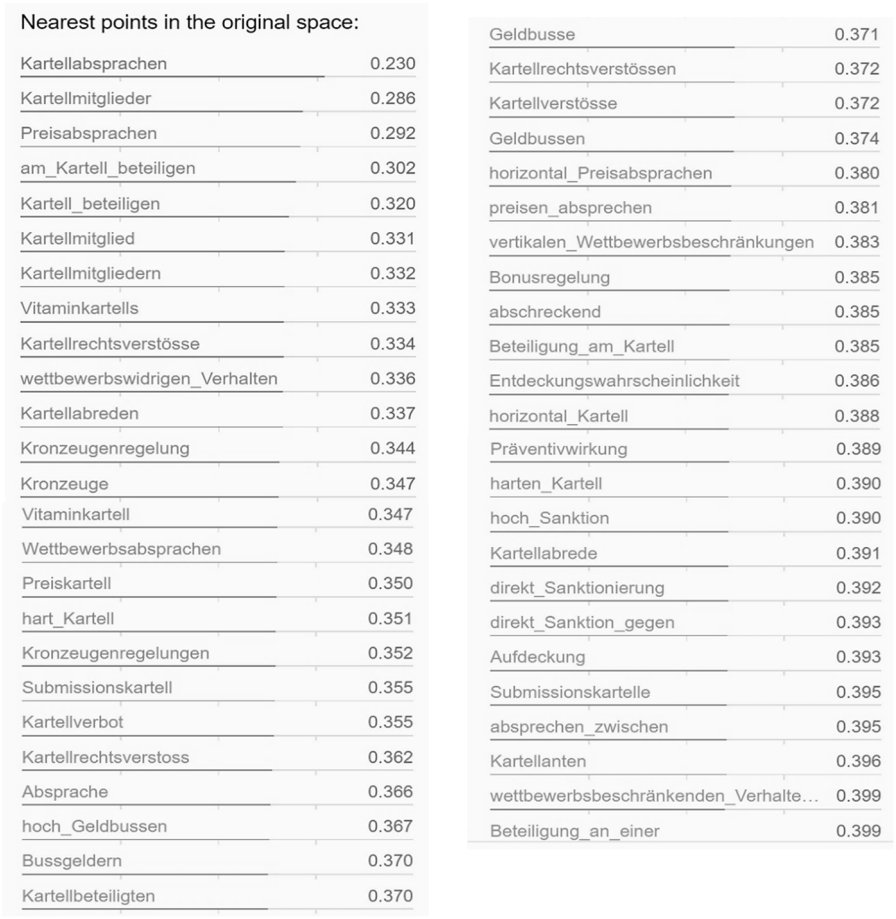

Abb. 4: Top Nearest Neighbours zu *Kartell*

Abb. 4 zeigt absteigend Ausdrücke, die ähnlich verwendet werden wie *Kartell*. Die Darstellung erfolgte auf einem von uns auf Basis des Testkorpus berechneten Modell. Das Modell stellt eine Erweiterung des Suchwortes und damit des Recall dar, da auch Synonyme oder ähnliche Ausdrücke genannt sind: *Kartellabsprachen*, *Preisabsprachen*, *Kartellabreden*, *Wettbewerbsabsprachen*, *horizontale Preisabsprachen* etc. Die ähnlich verwendeten Ausdrücke werden als Nearest Neighbours ausgegeben, weil sie in der zugrundeliegenden Vektorberechnung in der Nähe des Ausgangswortes liegen.

3.2 Kollokationen

Kollokationen sind eine in der Korpuslinguistik bewährte Methode, um die Semantik eines Ausdrucks genauer zu bestimmen (Lemnitzer & Zinsmeister 2015). Es handelt sich um ein Verfahren, mit dem die statistische Überzufälligkeit, mit der zwei Wörter zusammen in Textdaten auftreten, berechnet wird (sog. Assoziation). Bei einer gleichmässigen Verteilung der Wörter in einem Korpus wird davon ausgegangen, dass sie auch gemäss ihrer jeweiligen Häufigkeit zusammen auftreten. Das Messen der Assoziation zwischen zwei Ausdrücken zeigt, ob die beiden Ausdrücke häufiger, als man es bei einer gleichmässigen Verteilung erwarten würde, zusammen auftreten: Dies

ist z. B. der Fall, wenn das Wort *Kartell* überdurchschnittlich oft mit dem Wort *Wettbewerbsabrede* auftritt (vgl. Tab. 2). Trifft dies zu, werden sie als statistisch signifikante Kollokation aufgefasst. Für die Berechnung der Assoziation werden verschiedene statistische Masse verwendet, so z. B. der Log-Likelihood-Test, Mutual Information oder der Fisher Exact Test (Evert 2009). Weiter muss definiert werden, wie gross das Window sein soll, in dem nach potenziellen Kollokaten gesucht werden soll (z. B. im vorliegenden Fall max. fünf Wörter links/rechts des Suchwortes). Tab. 3 zeigt eine Beispielberechnung von Kollokationen mit *Kartell*, geordnet nach statistischer Signifikanz (Log-likelihood value), wobei auch die beobachteten und erwarteten Häufigkeiten angegeben sind.

No.	*Word*	*Total no. in whole corpus*	*Expected collocate frequency*	*Observed collocate frequency*	*In no. of texts*	*Log-likelihood value*
1	Wettbewerbsabreden	4248	1.605	321	301	2793.134
2	treffen	3947	1.491	308	298	2700.329
3	Marktmacht	3765	1.423	273	267	2351.431
4	Rechts	9761	3.688	287	285	1945.649
5	andere	22742	8.593	326	305	1746.01
6	privaten	7269	2.747	239	239	1673.238
7	öffentlichen	14036	5.303	274	272	1633.623
8	oder	248060	93.728	491	387	841.357
9	und	832091	314.401	853	565	642.742
10	Beteiligung	3500	1.323	88	40	568.05
11	am	58564	22.128	163	76	370.711
12	ein	200755	75.854	293	168	360.526
13	einem	92773	35.054	194	101	347.672
14	hartes	37	0.014	25	19	347.399
15	beteiligt	3174	1.199	50	32	276.316
16	Organisation	5237	1.979	55	24	260.388
17	Monopolfragen	16	0.006	16	8	252.449
18	Unternehmen	64697	24.445	135	81	241.157
19	das	395000	149.248	368	190	229.573
20	kartellähnliche	65	0.025	19	11	220.926
21	Immaterialgüterrecht	3589	1.356	38	17	180.47
22	Regulierungsrecht	137	0.052	18	7	177.211
23	beteiligten	15304	5.783	59	38	167.989
24	Fernmelderecht	251	0.095	19	18	165.088
25	Teilnahme	2223	0.84	27	19	135.424
26	Wettbewerbsrechte	50	0.019	12	2	134.064
27	an	141526	53.475	154	84	125.379
28	Kartell	1772	0.67	22	7	111.282
29	Preisüberwachungsrecht	63	0.024	10	7	102.537
30	Preisbildungskommission	120	0.045	11	7	99.961

Tab. 3: Beispiel-Kollokationen mit *Kartell* (Auszug der Top 30)

Die Berechnung lässt sich als ein Kollokationsprofil von *Kartell* lesen: Es zeigt, dass *Kartell* signifikant häufig zusammen mit *Wettbewerbsabreden*, *treffen*, *Marktmacht* etc. auftritt. Es wird damit ein semantisches Profil sichtbar und die Kollokationen geben Antwort zu Fragen der Art:

- Was ist ein Kartell? Wettbewerbsabreden.
- Wie ist ein Kartell beschaffen? Welche Typen gibt es? Kartellähnlich, hart.
- Kartell gilt für? Unternehmen (privaten und öffentlichen Rechts).
- Assoziierte Rechtsbereiche? Immaterialgüterrecht, Regulierungsrecht, Fernmelderecht, Wettbewerbsrecht.

Die Kollokationen können also helfen, die Precision zu *Kartell* zu verbessern, indem z. B. *hartes* oder *kartellähnlich* dem Suchbegriff hinzugefügt werden. Falls diese Art von Kartell gesucht ist, kann so die Suche einfach eingeschränkt werden.

3.3 Anwendungsperspektiven der Word Embeddings

Die Begriffe *Kartellrecht* und *unzulässige Abrede* sind testweise untersucht worden. Bei den Word Embeddings sind N-Gramme berechnet worden, d. h. Einzelwörter oder hochfrequent miteinander vorkommende Wortgruppen (Abb. 5).

Nearest points in the original space:

Kartellrechts	0.189
Wettbewerbsrecht	0.266
europäisch_Kartellrecht	0.289
europäische_Wettbewerbsrecht	0.302
EU-Kartellrecht	0.316
europäisch_Wettbewerbsrecht	0.317
Verhältnis_zwischen_Immaterialgüter	0.323
europäisch_Wettbewerbsrechts	0.334
Kartellverfahrensrecht	0.338
Immaterialgüterrecht	0.342
im_Kartellrecht	0.343
EG-Kartellrecht	0.349
schweizerisch_Kartellrecht	0.350
Kartellzivilrecht	0.351
europäisch_Kartellrechts	0.353
schweizerisch_Kartellgesetz	0.353
Regulierungsrecht	0.355
Patentlizenzverträgen	0.356
Kartellrechtspraxis	0.357
kartellrechtlichen	0.359
Kartellgesetz	0.362
schweizerische_Kartellgesetz	0.363
schweizerisch_Kartellgesetzes	0.363
Marktmachtmissbrauch	0.364
EG-Wettbewerbsrecht	0.365
schweizerische_Wettbewerbsrecht	0.366
Wettbewerbsrechts	0.367
Andrea_Heinemann	0.367

Abb. 5: Top Nearest Neighbours zu *Kartellrecht*

Es zeigt sich, dass sich die Synonymgruppe zu *Kartellrecht* im engeren Sinn vor allem aus unspezifischem Wortgebrauch (*Wettbewerbsrecht*) sowie nationalen und internationalen Bezügen (z. B. *schweizerisches Kartellgesetz*, *europäisches Wettbewerbsrecht*, *EU-Kartellrecht*, *EG-Kartellrecht*) zusammensetzt.

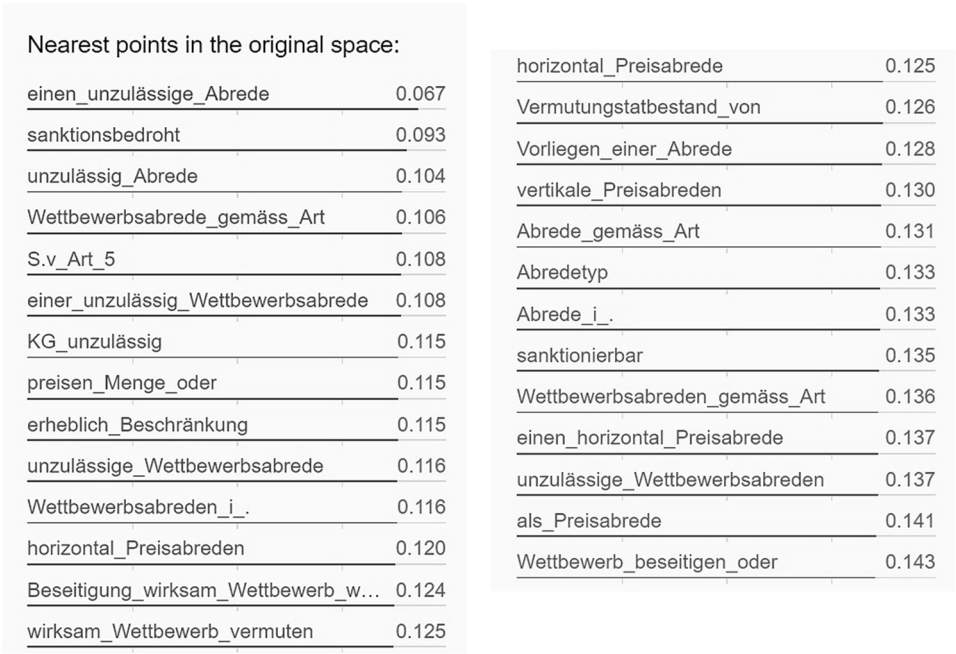

Abb. 6: Top Nearest Neighbours zu *unzulässige Abrede*

Die ähnlich verwendeten Wörter zu *unzulässige Abrede* (vgl. Abb. 6) belegen synonyme Ausdrücke im engeren (*unzulässige Wettbewerbsabrede*) sowie im weiteren Sinne: *Erhebliche Beschränkung* bezieht sich auf die Ursache, *sanktionsbedroht* bezieht sich auf die Folge von unzulässiger Abrede. Es werden auch juristische Prüfungsschritte genannt (*Vermutungstatbestand von*, *Vorliegen einer*). Es kann konstatiert werden, dass bei der simulierten Suche nach *Kartellrecht* als auch nach *unzulässige Abrede* sowohl Precision als auch Recall erhöht werden können, wenn die Suchbegriffe um einzelne Begriffe aus den jeweils berechneten Synonymgruppen ergänzt würden. Der Zweck des Embeddings stellt sich allerdings vor allem bei juristisch kompetenten Nutzern ein, die in der Lage sind, die vorgeschlagenen Wörter als entsprechende Alternativen zu deuten.

3.4 Anwendungsperspektiven der Kollokationen

Die Begriffe *Kartellrecht* und *unzulässige Abrede* (absolute Frequenz im Korpus: 10'208) ergeben ein Kollokationsprofil mit folgenden Merkmalen (vgl. Tab. 4): Auffällig ist, dass juristisch gebräuchliche Abkürzungen (z. B. *Rn, Rz, Band*) im Profil vorkommen, die auf den ersten Blick wenig zur Semantik des Begriffs beitragen. Es wird aber auch sehr deutlich, welche Literatur besonders einschlägig ist, nämlich *Zäch* und *Emmerich* als relevante Lehrbücher sowie der Kommentar von *Langen/Bunte*. Unmittelbar ersichtlich im Profil ist auch, dass es ein europäisches und ein schweizerisches Kartellrecht gibt. Mit diesen Zusatzinformationen kann die Suche präzisiert werden, etwa um schnell einschlägige Zitationen finden oder sich für eine Rechtsebene entscheiden zu können.

No.	Word	Log-likelihood
1	Zäch	18933.956
2	Schweizerisches	5887.357
3	Rn	4314.574
4	Roger	3477.879
5	2005	3291.514
6	Zürich	3164.884
7	Rz	2932.256
8	europäischen	2746.824
9	Kartellrecht	2702.501
10	im	2039.183
11	Bern	2022.568
12	Entwicklungen	1881.272
13	V/2	1685.012
14	das	1631.913
15	S	1606.027
16	N	1565.231
17	Emmerich	1473.477
18	Wettbewerbsrecht	1240.04
19	Aufl	1224.123
20	Band	1219.903
21	,	1207.615
22	Europäisches	1162.38
23	Immaterialgüter	1150.172
24	schweizerischen	1069.495
25	Langen/Bunte	1066.062
26	schweizerische	1003.159
27	Hrsg	945.391
28	LETTL	938.461
29	München	903.088
30	ff	895.39

Tab. 4: Kollokate zu *Kartellrecht* (Auszug der Top 30)

Für *unzulässige Abrede* (absolute Frequenz im Korpus: 327) ergibt sich ein Kollokationsprofil mit folgenden Merkmalen (vgl. Tab. 5): *Unzulässige Abrede* kommt in einem ökonomischen Kontext vor (*Unternehmen*, *Online-Handel*). Auffällig und sicherlich ohne statische Auswertung nicht so leicht festzustellen sind die Signifikanzen der spezifischen Berufsbezüge *Ärztegesellschaften* und *Musikalienhandel*. Für die Anwendung ergibt sich somit eine Perspektive auf thematische Teilgebiete, die die weitere Suchanfrage differenziert leiten können.

No.	Word	Log-likelihood
1	5	958.81
2	einer	693.924
3	an	525.851
4	nach	496.2
5	Unternehmen	388.41
6	3	327.268
7	Absätze	268.849
8	Art	260.537
9	KG	184.136
10	Online-Handel	182.384
11	Artikel	182.239
12	Abs	182.101
13	Abs.	138.351
14	Beteiligung	119.509
15	Verzicht	105.786
16	Berufsverbände	91.462
17	Ärztegesellschaften	90.383
18	eine	86.347
19	Musikalien-Händler	82.475
20	das	77.697
21	Mengensteuerung	75.608
22	möglicherweise	64.764
23	beteiligt	64.45
24	ein	55.713
25	S.v	51.214
26	marktmächtigen	44.301
27	Sinne	36.458
28	Massgabe	34.781
29	vorliegen	33.344
30	Text	32.126

Tab. 5: Kollokate zu *unzulässige Abrede* (Auszug der Top 30)

4. Fazit

Wie eingangs dargelegt, ist im Forschungsprozess von Praxispartner und Angewandter Linguistik gemeinsam eine testweise Entwicklung von linguistischen Suchmaschinenoptimierungen umgesetzt worden. Mithilfe von Abfragen aus dem Bereich des Kartellrechts konnten Anwendungsperspektiven aufgezeigt werden. Im Fall einer Implementierung ergeben sich folgende verbesserte Suchoptionen:

Nach einer entsprechenden Implementierung des Thesaurus und der korpuszentrierten Tools wäre es denkbar, dass während bzw. nach der Eingabe eines Suchbegriffs in zusätzlichen Fenstern oder in einem Dropdown-Menü unter dem Eingabefenster weitere für die Suche potenziell relevante Benennungen angezeigt werden. Diese Benennungen würden auf den oben genannten Relationsarten basieren (Äquivalenz-, Hierarchie- und Assoziationsrelationen). Durch Hinzufügen von verwandten Begriffen könnten somit beispielsweise Ambiguitäten aufgelöst werden, wodurch sich die

Precision erhöhen würde. Zudem bestünde die Möglichkeit, den ursprünglich gewählten Suchbegriff durch einen anderen zu ersetzen (Synonym, Ober- bzw. Unterbegriff oder verwandter Begriff), falls dieser eher zum eigentlich gesuchten Thema führt. Auch dies würde zur Erhöhung der Precision führen, mit der Einschränkung, dass das Ersetzen des Suchbegriffs durch einen entsprechenden Oberbegriff in der Regel zu weniger spezifischen Dokumenten führt. Durch das Hinzufügen von Synonymen zum Suchbegriff könnte darüber hinaus der Recall erhöht werden, vorausgesetzt, die Suche ist so eingestellt, dass nach allen eingegebenen Wörtern gesucht wird. Schliesslich könnten sich die Nutzer durch das Navigieren in einem auf den Thesaurus basierenden Begriffsbaum einem Thema systematisch nähern.

Abgesehen davon, dass durch die Implementierung der beiden oben genannten Ansätze sowohl Precision als auch Recall erhöht werden könnten, würden der typischen Rechercheprаxis der Nutzenden von Swisslex Optionen hinzugefügt, die man als Affordanz bezeichnen kann (vgl. Hopkins 2020): Swisslex würde mit der Implementierung von Thesaurus, Word Embeddings und/oder Kollokationen neue Handlungsweisen der Recherche und des Findens anbieten, ohne dass diese zwingend in jedem Fall genutzt werden müssen. Das Recherchieren kann durch den Einbezug von Optionen breiter werden (vgl. Abb. 7) in dem Sinne, dass bisher nicht bekannte oder nicht bedachte Begriffsverwendungen (z. B. Ober- und Unterbegriffe), thematische Aspekte (z. B. häufige Teilthematisierungen einer Rechtsnorm), einschlägige Fälle/Urteile (z. B. häufig zitierte Fälle/Urteile) oder andere Parameter in den Blick kommen (vgl. Tab. 2 bis 5). Die Möglichkeit zur breiteren Recherche kompensiert die Selektion und Fokussierung auf Erwartetes (und mögliche Filterblasen-Effekte), indem in der Suche auch unerwartete Aspekte aufgezeigt werden, deren Relevanz durch die statistische Signifikanz belegt ist.

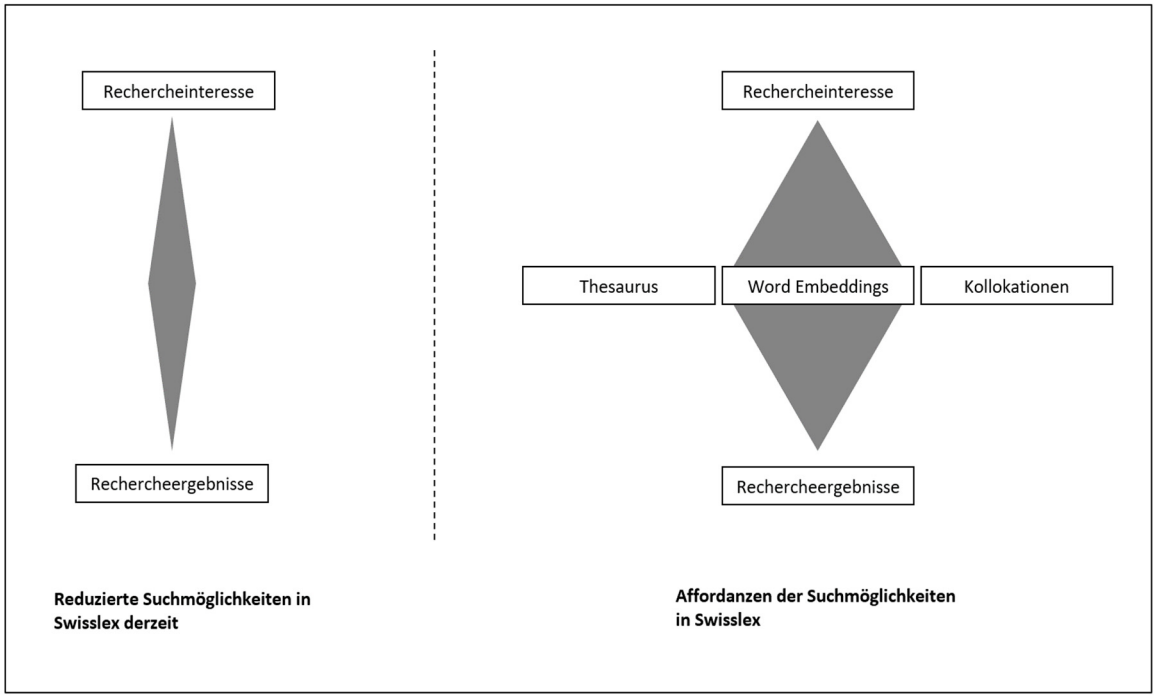

Abb. 7: Affordanzen der Suchmöglichkeiten in Swisslex nach der Optimierung

Mit diesem vorübergehend erweiterten Blick auf den Recherchegegenstand könnte die Suche in hohem Masse reflektiert durchgeführt werden, was sich optimalerweise im Verständnis des Ergebnisses niederschlägt. Den Swisslex-Nutzenden könnte dadurch anschaulich und nachvollziehbar gezeigt werden, was im Zuge ihrer Recherche an Zwischenergebnissen nicht weiterverfolgt und was vom erweiterten Recherchegegenstand genutzt wird. Denn jedem als relevant befundenen Ergebnis geht notwendigerweise ein Selektionsprozess voraus, in dem entschieden wird, was als nicht relevant bewertet wird. Durch die Möglichkeit, den Nutzenden diesen Prozess bewusst zu machen, kann sukzessive Wissen und Kompetenz im Umgang mit Rechercheergebnissen aufgebaut werden. Swisslex und seine Nutzende würden so einen weiteren Baustein im reflektierten und professionellen Umgang mit Suchergebnissen hinzugewinnen.

LITERATUR

Bubenhofer, N., Calleri, S. & Dreesen, P. (2019). *Politisierung in rechtspopulistischen Medien: Wortschatzanalyse und Word Embeddings*. *Osnabrücker Beiträge zur Sprachtheorie (OBST) 95*, 211-241.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge, MA: Cambridge University Press.

DIN 1463-1 (1987, 11). *Erstellung und Weiterentwicklung von Thesauri; Einsprachige Thesauri*. Berlin: Beuth.

Drewer, P., Massion, F. & Pulitano, D. (2017). *Was haben Wissensmodellierung, Wissensstrukturierung, künstliche Intelligenz und Terminologie miteinander zu tun?* DIT (Deutsches

Institut für Terminologie e.V.). http://dttev.org/images/img/abbildungen/DITeV_org_Terminologie_und_KI_2017_03_22_v2.pdf

Drewer, P. & Schmitz, K.-D. (2017). *Terminologiemanagement: Grundlagen - Methoden – Werkzeuge*. Berlin: Springer.

Erbguth, J. (2015). *Neue Suche bei Swisslex*. *Jusletter IT*, 26. Februar 2015. https://jusletter-it.weblaw.ch/issues/2015/IRIS/neue-suche-bei-swiss_867ac3edc2.html.

Evert, S. (2009). *58. Corpora and collocations*. In A. Lüdeling & M. Kytö (Hgg.), *Corpus Linguistics, Bd. 2* (S. 1212-1248). Berlin, New York: Mouton de Gruyter.

Henrich, A. (2008). *Information Retrieval 1. Grundlagen, Modelle und Anwendungen*. https://www.uni-bamberg.de/fileadmin/uni/fakultaeten/wiai_lehrstuehle/medieninformatik/Dateien/Publikationen/2008/henrich-ir1-1.2.pdf.

Hopkins, J. (2020). *The concept of affordances in digital media*. In H. Friese, M. Nolden, G. Rebane & M. Schreiter (Hgg.), *Handbuch Soziale Praktiken und Digitale Alltagswelten* (S. 47-54). Wiesbaden: Springer V.

Kießling, W. (2016). *Suchmaschinen*. Vorlesungsskript. https://www.informatik.uni-augsburg.de/lehrstuehle/dbis/db/lectures/ss16/se/scripts/script/ SEKap02_2.pdf.

Krasselt, J., Dreesen, P., Fluor, M., Mahlow, C., Rothenhäusler, K. & Runte, M. (2020). *Swiss-AL: A multilingual Swiss web corpus for applied linguistics*. *Proceedings of The 12th Language Resources and Evaluation Conference*, 4138-4144. Marseille, France. https://www.aclweb.org/anthology/2020.lrec-1.509 .

Landthaler, J., Waltl, B., Holl, P. & Matthes, F. (2016). *Extending full text search for legal document collections using word embeddings*. In F.J. Bex & S. Villata (Hgg.), *Legal knowledge and information systems*: *JURIX 2016* (S. 73-82). Amsterdam, Netherlands: IOS Press, (= the twenty-ninth annual conference. Frontiers in artificial intelligence and applications, volume 294).

Lemnitzer, L. & Zinsmeister, H. (2015). *Korpuslinguistik. Eine Einführung*. 3., überarb. und erw. Aufl. Tübingen: Narr.

Lyons, J. (1977). *Semantics. Bd. 2*. Cambridge, MA: Cambridge University Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality*. arXiv:1310.4546 [cs.CL]. https://arxiv.org/pdf/1310.4546.

Perrin, D. & Kramsch, C. (2018). Introduction: Transdisciplinarity in applied linguistics. In *AILA Review 31*, 1-13. https://doi.org/10.1075/aila.00010.int.

Spremann, K. & Bartmann, D. (2013). *Informationstechnologie und strategische Führung*. Berlin: Springer.

Stock, W. G. & Stock, M. (2015). *Handbook of information science*. Berlin: Mouton De Gruyter.

Wittgenstein, L. (1982). *Philosophische Untersuchungen*. Frankfurt am Main: Suhrkamp.

WEITERFÜHRENDE LITERATUR

Hedden, H. (2016). *The accidental taxonomist*. Second Edition. Medford, New Jersey: Information Today, Inc.

ISO 25964-1 (2011). *Information and documentation - Thesauri and interoperability with other vocabularies - Part 1: Thesauri for information retrieval*. ISO: Geneva.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). *Efficient estimation of word. Representations in vector space*. In: arXiv:1301.3781 [csCL]. https://arxiv.org/pdf/1301.3781.

Stock, W. G. & Stock, M. (2008). *Wissensrepräsentation. Informationen auswerten und bereitstellen*. München: Oldenbourg.