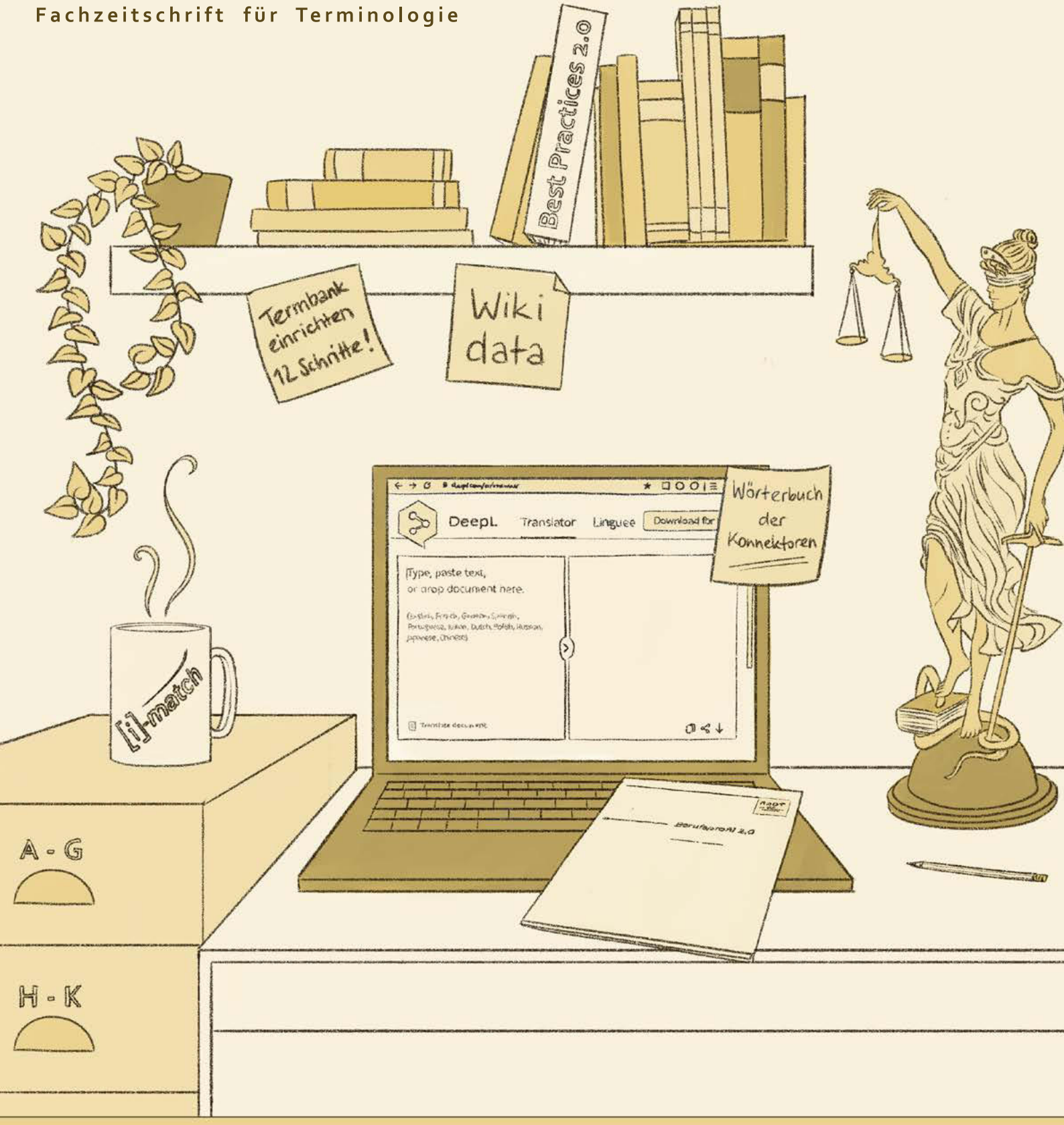
edition

Fachzeitschrift für Terminologie

1|20



Optimierung der Suchanfrage bei Swisslex

Christian Kriele

This article is about a project that aimed at optimizing the search results of a legal information platform. It 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.

Keywords: search engine optimization, recall, precision, semantic relations, thesaurus

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

Das Information Retrieval, das heißt die Gewinnung von Informationen bei Swisslex, basiert derzeit auf einer Volltextsuche mit diversen auf einer Nomenklatur basierenden Filtermöglichkeiten und Wortvorschlägen, die in einer Drop-down-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 Abbildung 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 bzw. 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“ [1]. 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) [2, S. 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 [3, S. 8].
- Für einen Begriff, auf den mit einer sprachlichen Bezeichnung verwiesen werden soll, gibt es oft mehrere Benennungen (Synonymie) [2, S. 16]. So existiert im Rechtskontext für die Benennung „Gebietsabrede“ beispielsweise das Synonym „Gebietsabsprache“.

Abb. 1: Suchoberfläche von Swisslex

> 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 [4, S. 208].

Optimierung der Suchergebnisse mit linguistischen Mitteln

Die übergreifende Zielsetzung des Projektes bestand wie eingangs erwähnt darin, Möglichkeiten der Optimierung der Suchanfrage mit Mitteln der Angewandten Linguistik 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 korpuslinguistischen Methoden. Dazu gehörte die Berechnung von Kollokationen (Wörter, die signifikant häufig in der sprachlichen Umgebung eines Wortes vorkommen) und von Word Embeddings (Berechnung von potenziell synonym verwendeten 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 [5, S. 120]. In DIN 1463-1 werden dabei analog zu ISO 25964-1 (2011) folgende semantische Relationen unterschieden und definiert [6, S. 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"

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.

Thesaurus

In DIN 1463-1 [6, S. 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.

Nachfolgend findet sich eine Liste mit Abkürzungen für Relationen aus DIN 1463-1, die auch im für Swisslex erstellten Thesaurus Anwendung finden [6, S. 11].

DIN 1463-1

OB Oberbegriff
UB Unterbegriff
BF Benutzt für
BS Benutzes Synonym
VB Verwandter Begriff

Ontologie

In der Informatik und der KI-Forschung werden unter Ontologien computerlesbare Wissensmodellierungen verstanden [7, S. 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.

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 angeben 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, sodass 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[2] einen Pilotthesaurus zu erstellen. Zum Einsatz kam dabei SKOS (Simple Knowledge Organisation System)[3], eine auf dem Resource Description Framework (RDF)[4] und RDF-Schema (RDFS)[5] 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.

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[6], SynchroTerm[7], Antconc[8] und IMS Open Corpus Workbench[9]. Die Ergebnisse wurden in einer Liste vereint und durch Swisslex validiert.

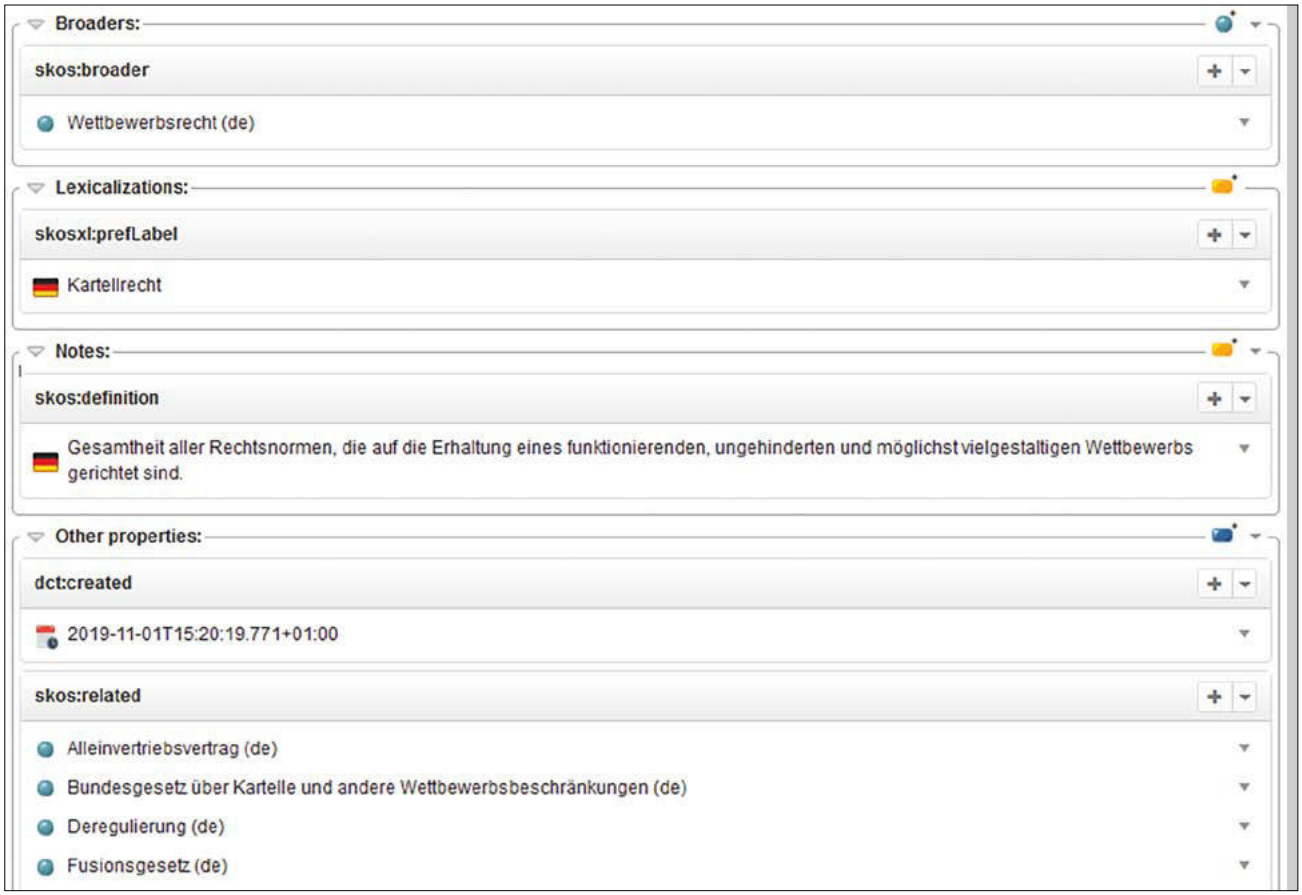


Abb. 2: Beispieleintrag im Pilotthesaurus

Bei diesem Validierungsschritt sortierten Swisslex-Mitarbeiter irrelevante Termini aus und 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 Korpuslinguistik (Kollokationen und Word Embeddings) erneut Analysen im oben genannten Textkorpus durchgeführt. Die bei diesen Analysen ermittelten Termini wur-

Synonyme	**Ober- bzw. Unterbegriff**	**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. 1: Informationen im Thesaurus zu „Gebietsschutz“

den 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. Abbildung 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).

Anwendung 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 Abbildung 3.

Für alle drei Benennungen wurde untersucht, welche Wortvorschläge der Thesaurus in Form von Synonymen, Ober- und Unterbegriffen und verwandten Begriffen bieten würde. Nachfolgend 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 theoretischer 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 tabellarisch aufgeführt. Damit die Darstellung übersichtlich bleibt, sind nur diejenigen Ober- und Unterbegriffe aufgeführt, die sich im Thesaurus jeweils auf der nächsten Ebene befinden (vgl. Tabelle 1 auf Seite 8).

Wie aber könnten die Suchergebnisse nun anhand von diesen Informationen optimiert werden? Eine Erhöhung des Recalls könnte bei der Suche nach „Gebietsschutz“ beispielsweise erreicht werden, 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

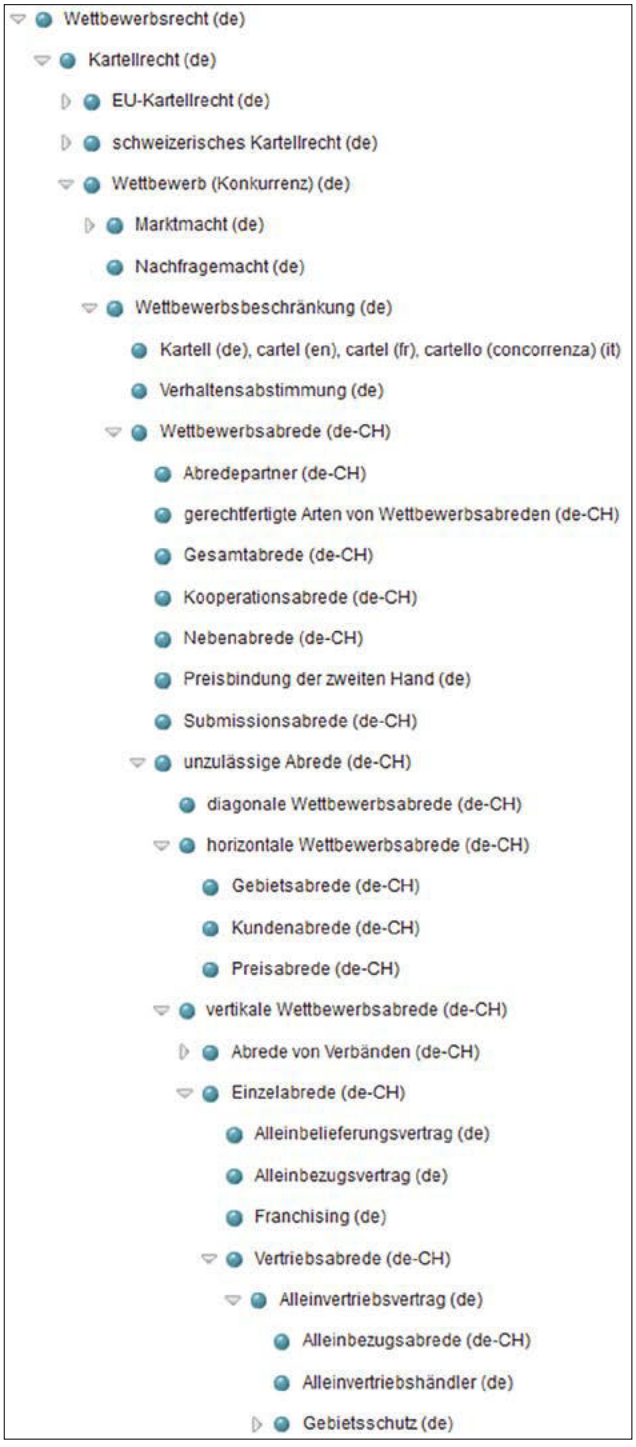


Abb. 3: Ausschnitt aus der Hierarchie des Pilotthesaurus

werden, indem dem Suchbegriff ein verwandter Begriff wie beispielsweise „Alleinbelieferungspflicht“ oder „Alleinbezug“ hinzugefügt wird.

Fazit

Nach einer entsprechenden Implementierung des Thesaurus und der korpuslinguistischen Tools wäre es denkbar, dass nach bzw. während der Eingabe eines Suchbegriffs in zusätzlichen Fenstern bzw. in einem Drop-down-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, wären für die Nutzer von Swisslex weitere Rechercheoptionen verfügbar: Den Nutzern würden mit der Implementierung von Thesaurus, Word Embeddings bzw. Kollokationen neue Rechercheverfahren und Suchstrategien angeboten werden. Die Recherche könnte durch das Einbeziehen von zusätzlichen Optionen in dem Sinne erweitert werden, dass bisher nicht bekannte oder nicht bedachte Benennungen (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 könnten. Die Möglichkeit zur breiteren Recherche würde die Selektion und Fokussierung auf Erwartetes (und mögliche Filterblasen-Effekte) kompensieren, indem in der Suche auch unerwartete Aspekte aufgezeigt werden.

[1] https://www.swisslex.ch/
[2] http://vocbench.uniroma2.it/
[3] https://www.w3.org/2004/02/skos/
[4] https://www.w3.org/2001/sw/wiki/RDF
[5] https://www.w3.org/2001/sw/wiki/RDFS
[6] https://extraterm.org/index.html
[7] https://terminotix.com/index.asp?content=item&item=7&lang=en
[8] http://www.laurenceanthony.net/software.html
[9] http://cwb.sourceforge.net/

Literatur

[1] Identität und Geschichte der Informationswissenschaft | Informationserschließung und Information Retrieval. Abgerufen am 09.04.2020 von https://saar.infowiss.net/projekte/ident/themen/info_aufbereitung/recall/
[2] Drewer, P. / Schmitz, K.-D. (2017): Terminologiemanagement: Grundlagen – Methoden – Werkzeuge. Berlin: Springer.
[3] Kießling, W. (2016): Suchmaschinen. Vorlesungsskript. Abgerufen am 09.04.2020 von https://www.informatik.uni-augsburg.de/lehrstuehle/dbis/db/lectures/ss16/se/scripts/script/SEKap02_2.pdf
[4] Spremann, K. / Bartmann, D. (2013): Informationstechnologie und strategische Führung. Berlin: Springer.
[5] Henrich, A. (2008): Information Retrieval 1. Grundlagen, Modelle und Anwendungen. Abgerufen am 09.04.2020 von https://www.uni-bamberg.de/fileadmin/uni/fakultaeten/wiai_lehrstuehle/medieninformatik/Dateien/Publikationen/2008/henrich-ir1-1.2.pdf
[6] DIN 1463-1:(1987, 11): Erstellung und Weiterentwicklung von Thesauri; Einsprachige Thesauri. Berlin: Beuth.
[7] 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.). Abgerufen am 09.04.2020 von http://dttev.org/images/img/abbildungen/DITeV_org_Terminologie_und_KI_2017_03_22_v2.pdf

Weiterführende Literatur

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

Christian Kriele studierte Übersetzen an der Universität des Saarlandes und arbeitet seit 2011 als Dozent an der ZHAW Zürcher Hochschule für Angewandte Wissenschaften. Er vertritt dort den Bereich Terminologie in der Lehre, in der Forschung und in Form von internen und externen Dienstleistungen. Seit Januar 2014 ist er stellv. DTT-Vorsitzender und zuständig für das Ressort „Fortbildungen".

Kontaktadresse
christian.kriele@zhaw.ch
www.zhaw.ch