NEUE SUCHE BEI SWISSLEX

Jörn Erbguth[1]

[1] CTO, Swisslex – Schweizerische Juristische Datenbank AG
Rue du Mont Blanc 21, 1201 Genf
jerbguth@swisslex.ch; www.swisslex.ch

Schlagworte: ***Juristische Informationssysteme, Computerlinguistik, Crosslinguale Suche, Lemmatisierung, Wortarterkennung***

Abstract*:* *Die treffgenaue Suche stellt bei juristischen Datenbanken einen entscheidenden Qualitätsfaktor dar. Die neue Suchmaschine bei Swisslex kombiniert Linguistik-Werkzeuge mit dem crosslingualen Thesaurus TDS III und bietet damit eine im juristischen Bereich einzigartige Suchpräzision. Zusätzlich bietet ein Regelframework die Möglichkeit, unerwünschte Mehrdeutigkeiten gezielt auszublenden. Dadurch kann Swisslex die juristische Suche auf einfache Art und Weise kontinuierlich weiter optimieren. Gleichzeitig wird die linguistische Aufbereitung der Suchanfrage transparent dargestellt. Nutzerinnen und Nutzern können hiermit unerwünschte Wortformen oder Übersetzungen gezielt bei der Suche ausklammern. Swisslex setzt damit bei der juristischen Suche nicht nur in Bezug auf Precision und Recall sondern auch in Bezug auf Usability neue Massstäbe.*

1. Suche bei Swisslex

Die *Swisslex – Schweizerische Juristische Datenbank AG* ist die führende Rechtsinformationsplattform der Schweiz (www.swisslex.ch). Sie bietet einen Zugriff auf schweizerische Gesetzestexte, Urteile eidgenössischer und kantonaler Instanzen sowie umfangreiche juristische Literatur. Die wichtigsten Zugriffsmethoden sind die Suche und die Verlinkung der Dokumente untereinander. Die von Swisslex ab der Version 3.0 eingesetzten Applikationen im Back- und Frontend waren bis auf die Suchmaschine Eigenentwicklungen.

Swisslex beschloss 2013, auch eine eigene Suchmaschine auf der Basis von *Apache Lucene / Solr*[1] zu programmieren. Ziel der Neugestaltung der Swisslex-Suche war die Verbesserung von Recall, Precision und Usability. Ein besonderes Augenmerk lag dabei auf der Weiterentwicklung der crosslingualen Suche, die zudem unabhängig von den Flexionen der Wörter sein soll. Störendes Rauschen durch Mehrdeutigkeiten sollte reduziert werden. Gleichzeitig sollte die Suche transparent und manuell anpassbar sein. Im Folgenden werden zunächst die Anforderungen im Einzelnen und dann die Architektur des gewählten Lösungsansatzes sowie ihre Besonderheiten beschrieben.

2. Anforderungen und Herausforderungen

2.1. Crosslinguale Suche

Die Schweiz hat vier Landessprachen: deutsch, französisch, italienisch und räteromanisch. Bei Swisslex sind 70% der Inhalte in deutscher, 26% in französischer und 4% in italienischer Sprache. In diesem Kontext ist es unabdingbar, dass Swisslex eine crosslinguale Suche anbietet, d. h. dass

[1] http://lucene.apache.org/solr.

z. B. mit deutschen Suchbegriffen auch die entsprechenden französisch- oder italienisch-sprachigen Dokumente gefunden werden können.

2.2. Normalisierung der Flexionsformen

Suchbegriffe sollen nicht nur in der eingegebenen Form sondern auch in anderen Flexionsformen gefunden werden. So soll z. B. mit *unberechtigt* auch *unberechtigter*, *unberechtigte* oder *unberechtigtes* gefunden werden. Zudem sollen nicht nur die sich durch Endungen unterscheidenden Flexionsformen, sondern auch im Wortstamm abgewandelte Formen wie *ging* beim Verb *gehen* gefunden werden. Begriffe wie *Oberst* und *oberster* unterscheiden sich zwar nur in ihrer Endung, gehören morphologisch jedoch nicht zusammen und sollen daher unterschieden werden. Bei Begriffen wie *Garten* und *Betrug*, die gleichzeitig Substantive als auch Verbformen von *garen* und *betragen* sind, sind die Verben meist nicht mit gemeint. Die alleinige Anwendung morphologischer Regeln reicht hier nicht.

2.3. Neue Rechtschreibung

Die Suche soll das gleiche Resultat ergeben, unabhängig davon, ob die neue oder alte deutsche Rechtschreibung verwendet wird. Dies gilt zum einen für Schreibweisen wie *Schifffahrt* und *Schifffahrt* als auch für neu auseinander geschriebene Wörter wie *anders denkend* oder *bekannt geben.*

2.4. Suchlogik

Bei Swisslex soll weiterhin mit einer erweiterten booleschen Suchlogik gesucht werden. Neben den Operatoren **AND**, **OR** und **NOT** gibt es dabei die Operatoren **ADJ**, **NEAR** und **SAME**, die bestimmen, dass zwei Suchbegriffe direkt hintereinander, innerhalb einer bestimmten Anzahl von Wörtern oder im selben Absatz stehen müssen.

2.5. Mehrwortbegriffe

Es gibt viele Fachtermini, die aus mehreren Wörtern bestehen, z. B. *ungerechtfertigte Bereicherung*. Werden diese eingegeben, sollen Dokumente gefunden werden, in denen diese Fachtermini ggf. auch in flexierter Form wie z. B. einer *ungerechtfertigten Bereicherung* im Zusammenhang enthalten sind. Gleichzeitig sollen bei der Suche mit *ungerechtfertigte Bereicherung* Dokumente, in denen die einzelnen Wörter des Mehrwortbegriffs verteilt an unterschiedlichen Stellen im Dokument enthalten sind, wie z. B. *ungerechtfertigte Kritik* und *persönliche Bereicherung*, nicht gefunden werden.

2.6. Sortierung und Ranking

Für Juristeninnen und Juristen ist es wichtig, dass die neuesten Dokumente in der Trefferliste nach oben sortiert werden. Zwar bietet Swisslex auch eine Sortierung nach "Relevanz" an, diese wird jedoch nur selten verwendet. Daher ist eine recht unscharfe Suche, verbunden mit einem guten Relevanzranking wie es z. B. *Google* macht[2], für Swisslex weniger geeignet.

[2] Zur Suche bei Google: "*Alles über die Suche*". http://www.google.ch/intl/de/insidesearch/howsearchworks/algorithms.html.

2.7. Transparenz

Die Suche bei Swisslex soll transparent sein. Dies bedeutet, dass die Suche nachvollziehbar ist. Eine Suche mit statistischen Ähnlichkeiten würde diese Anforderung nicht erfüllen. Wichtig in diesem Zusammenhang ist jedoch, dass diese Anforderung nicht zu einer Komplizierung der Suchtechnik führt. Dementsprechend sollen sich Transparenz und Usability in der neuen Swisslex-Suche nicht ausschliessen.

2.8. Fine-Tuning

Auch bei sorgfältiger Konzeption der Suchmaschine wird nicht jede Entscheidung über die Breite oder Präzision der Suche optimal sein. Daher muss es für Swisslex auf einfache Art und Weise möglich sein, nachträglich die Interpretation für einzelne Begriffe oder für bestimmte Fallkonstellationen zu ändern.

Umgekehrt sollen auch die Nutzerinnen und Nutzer die linguistische Interpretation ihrer eigenen Suchanfrage anpassen können.

3. Architektur des Lösungsansatzes

Als Suchmaschine wird *Solr* eingesetzt, welches auf *Lucene* basiert. *Lucene*[3] ist eine open-source Suchmaschine. Sie ist hoch performant und hat die kommerziellen und hochpreisigen Suchmaschinen grösstenteils aus dem Markt verdrängt. Ergänzt wird *Solr* durch die *Language Tools* der Firma *Canoo*[4], die die Lemmatisierung der verschiedenen Flexionsformen übernimmt. Für die crosslinguale Suche wird schliesslich der *TDS III*[5] eingesetzt. Der *TDS III* ist ein im Auftrag des Vereins *eJustice.ch*[6] unter der Leitung des Schweizerischen Bundesgerichts weiterentwickelter Übersetzungsthesaurus.

In der Abbildung 1 wird der Suchablauf dargestellt, der aus drei Suchschritten besteht. Zunächst wird die **direkte Suche** durchgeführt. Dabei wird ein Suchbegriff ohne linguistische Aufbereitung direkt so gesucht, wie er eingegeben wurde.

Die Suchbegriffe werden dann auf ihre linguistischen Grundformen zurückgeführt; aus *Klägers* wird damit z. B. *Kläger*. Dies geschieht gleichermassen bei den Suchbegriffen wie bei den Begriffen in den Dokumenten. Dadurch werden z. B. *Regresses* und *Regresse* auf das Lemma *Regress* normalisiert. Damit kann bei der **Suche mit verschiedenen Flexionsformen** jede Flexionsform eines Begriffs als Suchbegriff eingegeben werden, und es werden alle Dokumente gefunden, die eine der Flexionsformen des Begriffes enthalten.

Die Verwendung eines auf Morphologie und Lexikon basierten Lemmatizers wird gegenüber dem Einsatz eines einfachen Stemming-Algorithmus bevorzugt, da ein Lemmatizer wesentlich genauer arbeitet.[7] Insbesondere bei der Verknüpfung des Lemmatizers mit der Übersetzung der Suchbegriffe würde ein einfacher Stemmer viel Rauschen erzeugt, welche die **Precision** negativ beeinflusst.

---

[3] http://lucene.apache.org.

[4] http://www.canoo.com – eine Version der deutschen Language Tools können unter http://canoo.net getestet werden.

[5] *Cappellano/Bühler*, TDS und Jurivoc: die beiden Schweizer juristischen Thesauri https://www.bj.admin.ch/dam/data/bj/staat/rechtsinformatik/magglingen/2010/03_cappellano_buehler-d.pdf.

[6] http://www.ejustice.ch.

[7] *Koernius/Laurikkala/Kalervo/Martti*, Stemming and Lemmatization in the Clustering of Finnish Text Documents, S. 632.

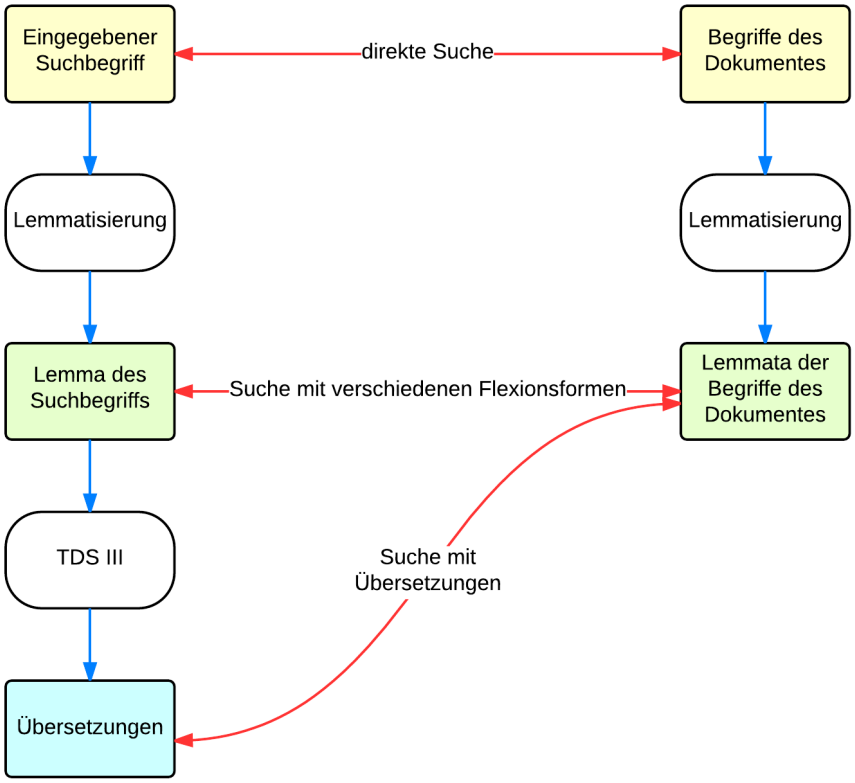

Abbildung 1: Basisarchitektur – Verarbeitung der Suchbegriffe (links) und Indexierung der Dokumente (rechts)

Schliesslich werden die Suchbegriffe bzw. deren Grundformen noch übersetzt. Damit findet z. B. *Regress* auch Dokumente in italienischer Sprache, die *Regresso* enthalten. Dies ist ein gängiger Ansatz für die crosslinguale Suche.[8]

4. Anpassungen des Lösungsansatzes

Der bislang beschriebene Ansatz bietet einen guten **Recall**, d. h. die meisten relevanten Dokumente werden damit gefunden. Allerdings gibt es einige Fälle, in denen durch linguistische Mehrdeutigkeiten oder unzureichende Berücksichtigung linguistischer Besonderheiten die **Precision** leidet. Im Folgenden werden daher entsprechende Anpassungen des oben allgemein beschriebenen Lösungsansatzes erläutert:

4.1. Mehrdeutigkeiten

Mehrdeutigkeiten gibt es in folgenden Bereichen:

4.1.1. Mehrdeutige Lemmatisierungen

Einige Wortformen erlauben mehrere Lemmatisierungen wie z. B. *Betrug* oder *Garten*. Die Suche mit *Betrug* meint meistens das Nomen und nicht das Verb *betragen* oder gar die französische Übersetzung *faire*. Ähnliches gilt für die Suche mit *Garten*, welche meistens nicht das Verb *garen* und erst recht nicht dessen Imperativ *gar* meint, wobei *gar* wiederum viele andere Bedeutungen hat. Wird umgekehrt mit dem französischen Begriff *jardin* gesucht, sollte die Übersetzung *Garten* gefunden werden, nicht jedoch die anderen Formen des Verbes *garen*.

[8] *Nasharuddin/Abdullah/Kadir/Azman*, A Review on the Cross-lingual Information Retrieval, S. 353.

Bei der Indexierung der Dokumente können solche Mehrdeutigkeiten durch eine **Part-of-Speech-Analyse** reduziert werden. Steht beispielsweise *Garten* gross geschrieben nicht am Satzanfang, kann es sich nur um das Nomen und nicht um die Verbform handeln. Dadurch wird bei der Indexierung der Dokumente ein Grossteil der falschen Zuordnungen vermieden.

Bei der Suche funktioniert dieser Ansatz leider nicht, da meistens mit einzelnen Suchbegriffen und nicht mit ganzen Sätzen gesucht wird. Zudem soll bei der Eingabe der Suchbegriffe die Gross- und Kleinschreibung weiterhin keinen Unterschied machen.

Um verbleibende linguistisch zulässige, aber ungewollte Bedeutungen auszuschliessen, wird die durch die Language Tools von Canoo durchgeführte Lemmatisierung um eine selbst entwickelte regelbasierte Korrekturkomponente ergänzt. Mit dieser können im Fall von Mehrdeutigkeiten meistens nicht gemeinte Bedeutungen von der Suche ausgeschlossen werden.

4.1.2. Mehrdeutige Akzentsetzungen

Akzente in der französischsprachigen Suche werden selbst von Muttersprachlern häufig nicht geschrieben. Daher soll die Suche hier tolerant sein. Auf der anderen Seite soll bei der Suche mit richtiger Akzentsetzung präzise gesucht werden: *Élève* findet damit weder *éleve* noch *élevé*. Werden dagegen Akzente vergessen oder falsch gesetzt, so findet das System weiterhin alle Formen. Diese Regeln wurden ebenfalls mit Hilfe der regelbasierten Korrekturkomponente implementiert.

4.1.3. "Falsche Freunde" in den unterschiedlichen Sprachen

Viele Wörter, die z. B. dem Lateinischen entlehnt sind, existieren gleichlautend in den verschiedenen Sprachen und haben identische oder sehr ähnliche Bedeutungen. Allerdings gibt es auch einige gleichlautende Wörter, die unterschiedliche Bedeutungen haben. So würde das französische Wort für *Gerichtsstand for* praktisch alle englischen Dokumente finden. Ebenso hat der deutsche *Donner* nichts mit dem französischen Verb *donner* zu tun.

Um bei diesen "falschen Freunden"[9] die Suchpräzision zu erhöhen, wird eine strikt sprachgetrennte Suche implementiert. Die Sprache der Suchbegriffe ist einstellbar. Nur in dieser Sprache wird lemmatisiert und nur aus dieser Sprache wird übersetzt. Lediglich bei der direkten Suche wird sprachunabhängig in allen Dokumenten gesucht.

4.1.4. Verschiedene Bedeutungen eines Lemmas

Manche Wörter sind bereits in ihrer Grundform in einer Sprache mehrdeutig. So kann *Bank* das Geldinstitut oder die Parkbank meinen. Während gerade bei der Übersetzung eine Zuordnung zu einer Bedeutung die **Precision** erhöhen würde, so wäre die Zuordnung zu einer Bedeutung in den Dokumenten nur nach aufwendiger und fehleranfälliger semantischer Analyse z. B. mit statistischen Verfahren und bei der Suche nur über noch unsicherere statische Annahmen über die Nutzer möglich. Daher wird bei Swisslex auf eine entsprechende Differenzierung verzichtet.

4.2. Mehrwortbegriffe

Die juristische Suche wird besonders effizient, wenn mit juristischen Fachbegriffen gesucht wird. Diese Fachbegriffe bestehen häufig aus mehreren Wörtern wie z. B. *ungerechtfertigte Bereicherung* oder *grobe Fahrlässigkeit*. Hier wäre eine Übersetzung der einzelnen Wörter sehr ungenau. Der

[9] Eine umfangreiche Liste von falschen Freunden findet sich bei Wikipedia http://de.wikipedia.org/wiki/Liste_falscher_Freunde.

TDS III bietet eine grosse Anzahl von Übersetzungen von Mehrwortbegriffen und Phrasen, die Swisslex für die Suche verwenden kann.

Wird ein solcher Mehrwortbegriff wie z. B. *ungerechtfertigte Bereicherung* erkannt, wäre eine Verknüpfung mit **AND** wenig präzise. Schliesslich sind damit eher nicht Dokumente gemeint, die beide Begriffe an weit entfernten Stellen in unterschiedlichen Zusammenhängen enthalten wie z. B. *ungerechtfertigte Kritik* an einer und *persönlichen Bereicherung* an anderer Stelle. Gleichzeitig wäre eine Suche als wortwörtliche Phrase zu eng. Flexionen wie *einer ungerechtfertigten Bereicherung* oder *der ungerechtfertigten Bereicherungen* sollen ebenfalls gefunden werden.

Daher werden Mehrwortbegriffe erkannt und auch ohne eingegebenen Operator automatisch zusammenhängend gesucht sowie übersetzt. Werden Mehrwortbegriffe mit einem Proximity-Operator wie z. B. **SAME** verwendet, so wird der Proximity-Operator in die Übersetzung übernommen. Daher findet *ungerechtfertigte SAME Bereicherung* nicht nur *Die Bereicherung war ungerechtfertigt* sondern auch *l'enrichissement était illégitime*.

4.3. Weibliche und männliche Berufsbezeichnungen

Unterscheiden sich weibliche und männliche Berufsbezeichnungen nur durch ihre Endungen, so können sie ebenfalls gleichgesetzt werden. Die Suche mit *Rechtsanwalt* findet daher auch Dokumente, in denen *Rechtsanwältin* vorkommt. Wird umgekehrt mit *Rechtsanwältin* gesucht, ist weniger klar, ob beide Geschlechter oder gezielt die weibliche Form gemeint ist. Gerade bei der Suche nach Urteilen zur Diskriminierung von Rechtsanwältinnen wäre es wenig hilfreich, alle Dokumente, die das Wort *Rechtsanwalt* enthalten, ebenfalls zu finden. Daher haben wir uns bei Swisslex entschieden, hier die Suche asymmetrisch zu gestalten und mit der weiblichen Form die männliche Form nicht mit zu finden.

4.4. Transparenz und Anpassbarkeit

Ohne spezielle Anpassungen des rein morphologisch-linguistischen Modells würden die Mehrdeutigkeiten zu einer schlechten **Precision** führen. Daher werden, wie in Kapitel 4.1. beschrieben, die Mehrdeutigkeiten durch ein Regelsystem reduziert. Leider leidet unter solchen Anpassungen aber auch die Transparenz des Suchverfahrens. Durch die regelbasierten Anpassungen ist es nicht mehr einfach erklärbar und vorhersehbar, welche speziellen Wortformen und Wortbedeutungen ausgeschlossen werden. Auch Nutzerinnen und Nutzer sollen einstellen können, welche Mehrdeutigkeiten ggf. bei der Suche ausgeschlossen werden sollen. Aus diesen beiden Gründen können sich Nutzerinnen und Nutzer bei Swisslex sowohl die Lemmatisierung als auch die Übersetzung anzeigen lassen und diese für ihre Suche anpassen (vgl. Abbildung 2 sowie Anhang).[10]

Zu jedem eingegebenen Suchbegriff werden die Grundformen angegeben. Für jede Grundform ist ersichtlich, welche Flexionsformen damit gefunden werden. Dabei werden nicht alle theoretisch denkbaren Formen, sondern nur die bei Swisslex in Dokumenten tatsächlich vorkommenden Formen angezeigt. Zu jeder Grundform werden die Übersetzungen in den anderen Sprachen angegeben. Neben den Übersetzungen finden sich wiederum ihre Flexionsformen, die bei Swisslex in Dokumenten vorhanden sind. Zur Anpassung können nun Übersetzungen oder auch komplette Grundformen abgewählt werden. Die neben den Übersetzungen bzw. Grundformen stehenden Flexionsformen werden dann nicht mehr mit gesucht.

[10] Eine solche Funktion hatte es in einer früheren Version von Swisslex schon einmal gegeben. Die aktuelle Funktion geht jedoch einiges darüber hinaus.

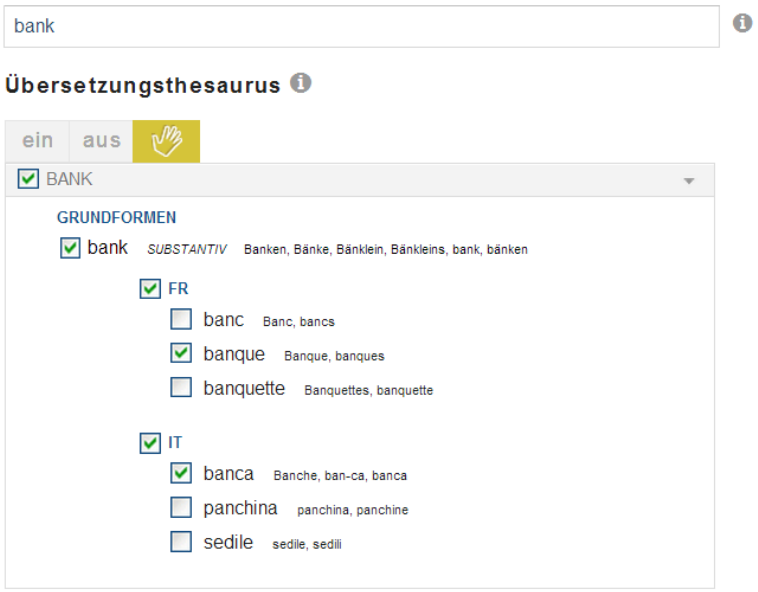

Abbildung 2: Auswahl der passenden Übersetzungen zum Begriff *Bank*

5. Evaluation

Für die Evaluation wurden die 1‘000 häufigsten Suchen bei Swisslex extrahiert und die Suchresultate der Suchen mit den verschiedenen Ansätzen verglichen. Im Ergebnis konnten sowohl **Recall** als auch **Precision** gegenüber der bisherigen Lösung verbessert werden. Das Benutzerinterface wurde in einem Versuchslabor auf seine Usability getestet.

6. Conclusio

Lemmatisierung verknüpft mit einer Lexikon-basierten Übersetzung der Suchbegriffe ermöglicht eine crosslinguale Suche. Ohne Einschränkungen würden jedoch verschiedene Arten von Mehrdeutigkeiten die Suchpräzision deutlich beeinträchtigen. Daher wurde dieser Ansatz durch eine regelbasierte Disambiguierung ergänzt. Das Ergebnis der linguistischen Modifikation der Suche wird den Nutzern angezeigt und kann bei Bedarf modifiziert werden. Damit konnten nicht nur die Anforderungen der Optimierung von Recall und Precision erfüllt werden, sondern die Suche bei Swisslex gewann dadurch zusätzlich an Transparenz und Anpassbarkeit.

7. Literatur

Korenius, Tuomo/Laurikkala, Jorma/Järvelin, Kalervo/Juhola, Martti, Stemming and lemmatization in the clustering of finnish text documents. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management* (pp. 625-633). ACM 2004

Nasharuddin, Nurul Amelina/Abdullah, Muhamad Taufik/Kadir, Rabiah Abdul/Azman, Azreen, A Review on the Cross-lingual Information Retrieval, Information Retrieval & Knowledge Management,(CAMP), 2010 International Conference on. IEEE 2010

Cappellano/Bühler, TDS und Jurivoc: die beiden Schweizer juristischen Thesauri https://www.bj.admin.ch/dam/data/bj/staat/rechtsinformatik/magglingen/2010/03_cappellano_buehler-d.pdf

8. Anhang

Swisslex stellt einen Dialog zur Anpassung der Übersetzung und Lemmatisierung zur Verfügung. Bei den im Text genannten Beispielen stellt sich dies wie folgt dar:

Abbildung 3: Übersetzungen der Suche mit Gerichtsstand

Abbildung 4: Suche mit Rechtsanwältinnen findet nur die weiblichen Formen

Abbildung 5: Suche mit *Rechtsanwalt* findet auch die weiblichen Formen