Bilingwis: ein statistikbasiertes Konkordanzsystem für die Systematische Rechtssammlung des Bundes

Manuela Weibel | *In diesem Bericht wird Bilingwis beschrieben, ein am Institut für Computerlinguistik der Universität Zürich entwickeltes statistikbasiertes Konkordanzsystem, das eine parallele Wortsuche in der deutschen und der französischen Version der Systematischen Rechtssammlung des Bundes ermöglicht. Mithilfe von Bilingwis können die verschiedenen Übersetzungsvarianten eines Wortes im jeweils dazugehörigen Kontext gefunden und direkt miteinander verglichen werden. Vor Kurzem wurde Bilingwis um das Sprachenpaar Deutsch–Rätoromanisch erweitert.*

Inhaltsübersicht

1 Einleitung

Im Rahmen eines Forschungsprojekts am Institut für Computerlinguistik der Universität Zürich ist ein statistikbasiertes Konkordanzsystem mit dem Namen «Bilingwis» entstanden, das die Suche nach Übersetzungsvarianten von Wörtern in zweisprachigen, parallelen Textkorpora ermöglicht (Volk et al. 2011).

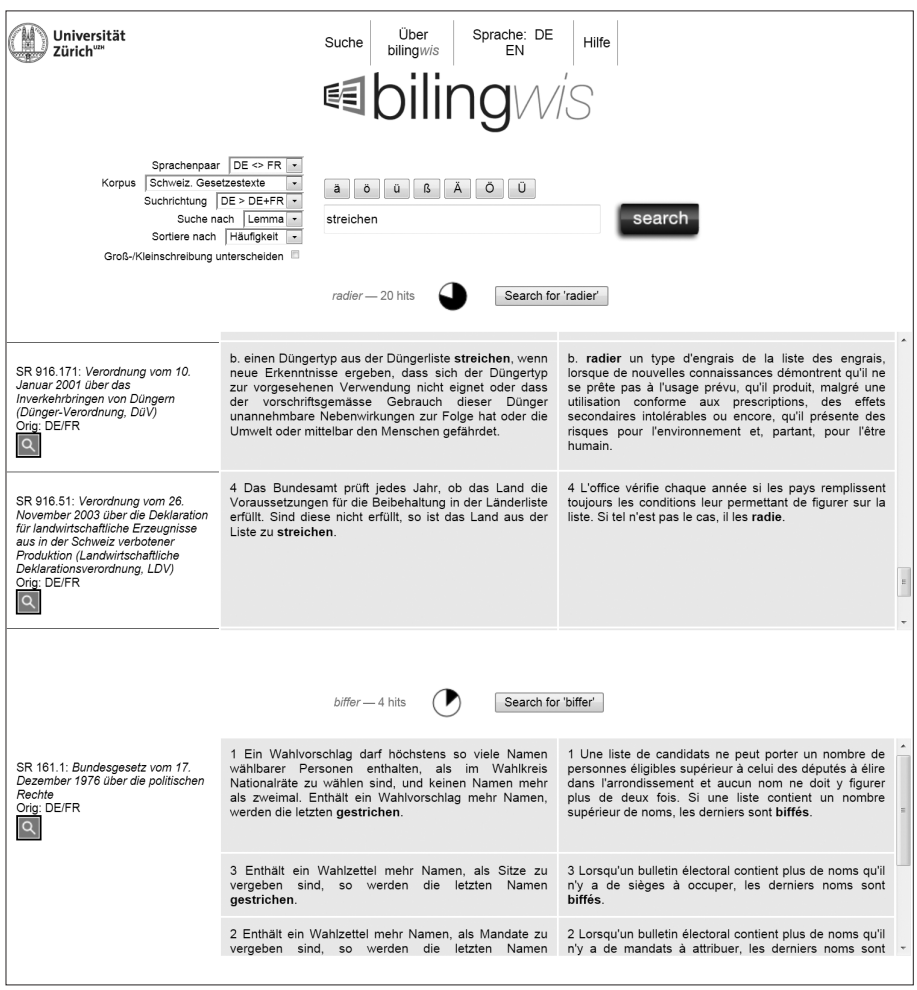
Bilingwis wurde ursprünglich für die Suche in der linguistisch aufbereiteten Sammlung der deutschen und französischen Ausgaben der Jahrbücher des Schweizerischen Alpenclubs (SAC) entwickelt.[1] Seit Anfang 2013 bietet Bilingwis auch die Suche innerhalb der bis zum 16. Oktober 2012 online publizierten französischen und deutschen Versionen der Systematischen Rechtssammlung (SR) an.[2] Ende 2013 ist das System zudem im Rahmen einer Masterarbeit um die ins Rätoromanische übersetzten Erlasstexte der SR erweitert worden.

Im Folgenden werden zunächst die Verwendung von Bilingwis und sein potenzieller Nutzen für Übersetzerinnen und Übersetzer von Erlasstexten dargestellt (Abschnitt 2). Anschliessend werden die Funktionsweise des Systems erklärt und die sich daraus ergebenden Vor- und Nachteile diskutiert (Abschnitt 3). In einem letzten Teil wird schliesslich die Erweiterung des Systems für das Rätoromanische vorgestellt (Abschnitt 4).

2 Verwendung

Bilingwis ermöglicht das Durchsuchen von parallel in zwei Sprachen vorliegenden Texten nach den verschiedenen Übersetzungsvarianten eines Wortes.

Wie ein Online-Wörterbuch bietet Bilingwis die Suche nach einem bestimmten Suchbegriff an. Anders als ein Wörterbuch liefert es jedoch als Ergebnis nicht bloss isolierte Übersetzungsvarianten, sondern gibt für jede dieser Varianten an, (a) wie häufig und (b) in welchen Kontexten sie in den zugrunde liegenden Texten – hier also in den Texten der SR – vorkommt. Abbildung 1 zeigt einen Ausschnitt der Benutzeroberfläche von Bilingwis bei der Suche nach dem Verb *streichen*. An 20 Stellen wurde *streichen* mit *radier* übersetzt, an 4 weiteren Stellen mit *biffer*. Sämtliche Treffer werden im Kontext des Satzes, in dem sie vorkommen, aufgeführt. Die Quellenangabe links der Trefferauflistung gibt Aufschluss über die Herkunft jedes Treffers und ermöglicht weitere Recherchen zu einem aufgeführten Textausschnitt.



Die Angabe der Kontexte ermöglicht es der Benutzerin oder dem Benutzer, für jede Teilbedeutung eines Wortes die adäquate Übersetzung zu ermitteln. Das Verb *streichen* beispielsweise wird in den französischen Texten der SR vorwiegend mit *radier* wiedergegeben. Geht es aber um die Streichung von Namen auf Wahlzetteln, wird im Französischen das Verb *biffer* verwendet.

Ähnlich verhält es sich mit dem Adjektiv *alt*. So hat dieses Wort nicht dieselbe Bedeutung, wenn es in den Ausdrücken «Alter Rhein», «alt Bundesrat» und «zwischen 16 und 22 Jahre alt» verwendet wird. Der Bedeutungsunterschied manifestiert sich in der Tatsache, dass im Französischen in jedem dieser drei Fälle ein anderes Adjektiv verwendet wird: *vieux* («von hohem Alter»), *ancien* («ehemalig») und *âgé (de)* («im Alter von»). Während aus einem einfachen Wörterbuch nicht ersichtlich wird, welche der drei französischen Übersetzungen in welchem Kontext angebracht ist, kann die Benutzerin oder der Benutzer von Bilingwis anhand von konkreten Satzbeispielen den entsprechenden Zusammenhang erschliessen.

Die Angabe der Häufigkeitsverteilung lässt Rückschlüsse darüber zu, wie geläufig eine bestimmte Variante in der entsprechenden Textdomäne ist. So scheint beispielsweise *biffer* gegenüber *radier* eine deutlich restringiertere Verwendung zu haben. Die Häufigkeitsverteilung der Übersetzungsvarianten kann zudem Aufschluss geben über die Zuverlässigkeit eines Resultats.

Weil Bilingwis ein System ist, das auf rein statistischen Methoden basiert (siehe Abschnitt 3), kommt es vor, dass in den entsprechenden Kontexten die falschen Wörter markiert werden. So markiert Bilingwis in zwei Fällen zum Beispiel mit *alors radié* ein zu umfangreiches Segment als Übersetzung von *streichen*. Dies hat zur Folge, dass die beiden Treffer unter einer falschen Übersetzungsvariante eingeordnet werden, in diesem Fall unter dem Wort *alors*, obwohl die aufgeführten Übersetzungsbeispiele eigentlich korrekt sind.

Die Benutzeroberfläche ermöglicht neben der Suche nach einer bestimmten Wortform auch die Suche anhand der Grundform, dem sogenannten Lemma: Wenn man beispielsweise nach dem Wort *alt* sucht, so werden lediglich diejenigen Treffer aufgeführt, in denen das Suchwort in exakt der eingegebenen Form auftritt (*agé* – 48 Treffer, *ancien* – 0 Treffer). Wählt man jedoch die Option «Suche nach: Lemma» aus, so finden sich auch flektierte Formen des Suchworts, also zum Beispiel *alte, alten, älter* oder *älteste*. Auf diese Weise erhöht sich die Trefferzahl, was bedeutet, dass auch die Ergebnisse zuverlässiger werden.

3 Funktionsweise

Die Übersetzungen, die Bilingwis liefert, beruhen nicht auf einem manuell erstellten Wörterbuch, sondern basieren ausschliesslich auf mit statistischen Methoden automatisch berechneten Wortentsprechungen (sog. Wortalignierungen). Statistische Methoden haben gegenüber regelbasierten oder gar manuellen Methoden den Vorteil, dass sie grosse Mengen an Textdaten in kurzer Zeit erfassen und verarbeiten können.

In einem ersten Schritt ordnet das System die einzelnen Texte und innerhalb der Texte die einzelnen Artikel und Absätze einander zu. Weil die Alignierung in Erlasstexten auf diesen Gliederungsebenen bereits inhärent vorhanden ist, erübrigt sich hierfür der Einsatz statistischer Methoden. Statistische Methoden werden jedoch benötigt, um in einem zweiten Schritt die Entsprechungen auf der Satz- und Wortebene berechnen zu können. Diese sind weitaus variabler als jene auf den übergeordneten Gliederungsebenen. So kann es vorkommen, dass zwei deutsche Sätze im Französischen zu einem Satz zusammengefasst werden:

(1) dt. *[1]Der Gewinnsteuer unterliegt der gesamte Reingewinn. Dazu gehören auch:*
fr. *[1]L'impôt sur le bénéfice a pour objet l'ensemble du bénéfice net, y compris:*
(SR 642.14)

Im Vergleich zur vereinzelt auftretenden Asymmetrie auf der Satzebene sind die Entsprechungen auf Wortebene zwischen zwei Sprachen deutlich komplexer, da sie mehr Interpretationsraum zulassen. So kann es vorkommen, dass ein Wort der Ausgangssprache keinem oder gleich mehreren Wörtern der Zielsprache entspricht. Auch die Reihenfolge der einzelnen Satzelemente kann zwischen den beiden Sprachen stark variieren. Diese Schwierigkeit lässt sich anhand von Beispiel (1) illustrieren: Dem deutschen Substantiv *Gewinnsteuer* entspricht der Ausdruck *impôt sur le bénéfice*, während das Verb *unterliegen* im Französischen mit der Paraphrase *avoir pour objet* ausgedrückt wird.

Aufgrund dieser asymmetrischen Entsprechungen können die Übersetzungen auf Satz- und Wortebene einander nicht mehr eins zu eins zugeordnet werden, sondern bedürfen einer komplexeren Alignierung. Diese Aufgabe übernehmen statistische Werkzeuge, welche die wahrscheinlichsten Entsprechungen berechnen.

Um die Qualität dieser Berechnungen zu erhöhen, wird für die Wortalignierung nicht mit den im Text auftretenden Wortformen gearbeitet, sondern mit

ihrer Grundform. Flektierte Wortformen wie *Haus, Hause, Hauses* oder *Häuser* werden unter dem Lemma *Haus* zusammengefasst. Dadurch verbessert sich die Zuverlässigkeit der statistischen Berechnung: Je mehr Vorkommen ein Wort aufweist, desto zuverlässiger wird seine Alignierungsberechnung. Die Bestimmung der Grundform jedes Wortes findet ebenfalls mithilfe statistischer Werkzeuge statt.

Statistische Methoden werden umso zuverlässiger, je umfangreicher das Textmaterial ist, auf dem sie trainiert werden. Der Effizienz dieser Methoden steht der Nachteil von möglichen Falschberechnungen aufgrund automatisierter Algorithmen gegenüber. Falsche Alignierungen sind in Bilingwis deshalb möglich, jedoch eher selten.

4 Erweiterung für Rätoromanisch

Ende 2013 wurde das Bilingwis-Angebot für die Systematische Rechtssammlung im Rahmen einer Masterarbeit (Weibel 2014) um das Sprachenpaar Deutsch–Rätoromanisch erweitert.[3] Integriert sind sämtliche parallelen Erlasse in Deutsch und Rumantsch Grischun, die am 13. September 2013 in der Onlineversion der SR verfügbar waren. Der Entscheid zugunsten der rätoromanischen gegenüber der verfassungsmässig priorisierten italienischen Sprache hatte verschiedene Gründe. In erster Linie war er von der Motivation geleitet, den medialen Zugang zu dieser wenig beachteten Landessprache zu stärken. Unterstützend waren zudem der Forschungsfokus am Institut für Computerlinguistik der Universität Zürich sowie die Forschungsfragen der Masterarbeit.

Der Forschungsfokus des Instituts verlagerte sich im Frühling 2013 in die Richtung dieser wenig beachteten Sprache, als im Rahmen eines Programmierprojekts ein Morphologieanalysesystem für die rätoromanische Sprache entwickelt wurde (Baumgartner et al., 2013). Dieses System bestimmt für jedes eingegebene Wort sowohl dessen Grundform als auch morphologische Eigenschaften wie Kasus, Numerus und Genus. Wie oben erwähnt, ist die Verfügbarkeit eines statistischen Werkzeugs, das diese Analyseaufgabe übernimmt, eine wichtige Voraussetzung für das Funktionieren von Bilingwis. Da nun zum ersten Mal überhaupt ein solches Werkzeug für das Rätoromanische vorhanden war, bot es sich an, dieses in der Praxis einzusetzen.

Für die rätoromanische Sprache liegen vergleichsweise wenige Textdaten vor. Der Nutzen statistischer Methoden bedingt jedoch grosse Mengen an Textdaten: Je umfangreicher die Ressourcen, desto zuverlässiger kann ein statistisches Werkzeug arbeiten. Spannend ist deshalb die Frage, wie sich die Tatsache, dass für eine Sprache nur vergleichsweise kleine Mengen an parallelen Daten vorhanden sind, auf die Zuverlässigkeit von Bilingwis auswirkt. Diese Fragestellung konnte im

Rahmen der Masterarbeit dahingehend beantwortet werden, dass bereits ein paralleles Korpus mit knapp 500 000 Wörtern pro Sprache ausreicht, um ein zuverlässiges Wortsuchsystem zu erstellen. Trotz der gegenüber dem deutsch–französischen Korpus zehnmal kleineren Textmenge funktioniert dieses System sehr gut und zeigt die Vielseitigkeit der Sprache auf. So ist aus Bilingwis zum Beispiel ersichtlich, dass die Systematische Rechtssammlung für das deutsche Wort *Ausweis* sieben verschiedene Entsprechungen (*document, attestat, carta, cumprova, mussament, certificat* und *legitimaziun*) kennt.

5 Fazit

Gesetzestexte haben für die Erstellung eines bilingualen Wortsuchsystems den grossen Vorteil, dass die Texte bereits zu weiten Teilen inhärent aligniert sind. Aufgrund der häufigen Verwendung von konventionalisierten Phrasen (z. B. *Mit Busse wird bestraft, wer …*) ist das Vokabular von Gesetzestexten zudem stark repetitiv und ermöglicht dadurch auch auf der Wortebene eine zuverlässige Alignierung.

Gegenüber einem gängigen Wörterbuch haben korpusbasierte Suchsysteme mit Kontextausgabe wie Bilingwis zahlreiche Vorteile. So bietet die Kontextangabe semantische Zusatzinformation, die in Wörterbüchern oft vermisst wird. Wird das System auf einem Korpus mit neuen Texten erstellt, kennt das System sogar jüngere Ausdrücke und Übersetzungen. Dadurch ist es sehr flexibel und ermöglicht einen Einblick in die aktuelle Verwendung einer Sprache. Daneben kennt das System auch domänenspezifisches Vokabular wie Fachbegriffe, Eigennamen oder Akronyme, die dem Wörterbuch unbekannt sind. Des Weiteren gibt die Häufigkeitsangabe in Bilingwis Aufschluss über die Geläufigkeit einer Übersetzung.

Bilingwis kann nicht nur für das Auffinden von Übersetzungsmöglichkeiten im juristischen Kontext verwendet werden. Darüber hinaus ermöglicht es dank der Auflistung sämtlicher Treffer im Korpus die monolinguale Recherche innerhalb der Gesetzestexte nach konkreten Themenbereichen. Diese wird zusätzlich unterstützt durch die direkt mit der Onlineversion der Systematischen Rechtssammlung verlinkten Quellenangabe auf der linken Seite des jeweiligen Treffers. Nicht zuletzt kann das System deshalb auch von den Sprachdiensten des Bundes für den Zweck der Sicherstellung terminologischer Kohärenz verwendet werden.

Bilingwis ist online abrufbar unter: http://kitt.cl.uzh.ch/kitt/bilingwis

Manuela Weibel, MA, ehemalige Mitarbeiterin am Institut für Computerlinguistik der Universität Zürich, E-Mail: manuela.weibel@hotmail.com

Anmerkungen

1 Siehe Digitalisierungsprojekt «Text+Berg digital»: http://textberg.ch.

2 Die Daten wurden von Stefan Höfler aufbereitet und von Roger Wechsler ins Bilingwis-System integriert.

3 Die Masterarbeit ist online abrufbar unter: www.cl.uzh.ch/studies/theses/lic-master-theses/MLTA_Masterarbeit_Manuela_Weibel.pdf. Im Rahmen der Masterarbeit wurde auch das Korpus der zweisprachig verfassten Pressemitteilungen der Standeskanzlei Graubünden in das Bilingwis-System integriert.

Literatur

Baumgartner, Reto / Bachmann, Martina / Badat, Rolf / Hegglin, Daniel / Tron, Susanna / Widmer, Melanie, 2013, Morphologieanalyse für Rumantsch Grischun. Institut für Computerlinguistik der Universität Zürich.

Volk, Martin / Göhring, Anne / Lehner, Stéphanie / Rios, Annette / Sennrich, Rico / Uibo, Heli , 2011, World-aligned parallel text: A new resource for contrastive language studies, in: Proceedings of the Conference on Supporting Digital Humanities, Copenhagen, abrufbar unter: www.zora.uzh.ch/51481.

Weibel, Manuela, 2014, Aufbau paralleler Korpora und Implementierung eines wortalignierten Suchsystems für Deutsch – Rumantsch Grischun, Institut für Computerlinguistik der Universität Zürich.

Résumé

Bilingwis est un concordancier fondé sur des statistiques développé par l'université de Zurich (Institut für Computerlinguistik). Il permet de faire des recherches parallèles dans les versions allemande et française du Recueil systématique, de trouver les différentes traductions d'un mot dans leur contexte et de les comparer. Depuis peu, Bilingwis permet également de faire des recherches parallèles dans des textes en allemand et en romanche.