Anlage 1

Ein Ingenieurskonzept zur Herstellung von Wörterbüchern
(Robert Neumann, SAFIR GmbH)

Zweisprachige Wörterbücher sind Hilfsmittel beim Übersetzen von Texten aus einer Quellsprache in eine Zielsprache. Maßgeblich für den Gebrauchswert eines Wörterbuchs ist, wie gut es die Anfragen der Übersetzer nach Wörtern abdeckt. Dabei ist nicht der Umfang der Einträge im Wörterbuch entscheidend, sondern es gelten drei Kriterien:

- Je mehr der angefragten Wörter aus dem zu übersetzenden Quelltext als Einträge im Wörterbuch vorhanden sind, umso brauchbarer ist das Wörterbuch.
- Je präziser das angegebene zielsprachliche Übersetzungsäquivalent die Bedeutung des quellsprachlichen Wortes in der quellsprachlichen Domäne beschreibt, umso nützlicher ist das Wörterbuch.
- Mit je weniger Einträgen das Wörterbuch die beiden genannten Kriterien erfüllt, umso wirtschaftlicher ist es hinsichtlich der Erstellungskosten und der nachfolgenden Pflegekosten.

Das Herstellungskonzept

Das Konzept (Entwurf, Design) legt fest, welche Worteinträge in das Wörterbuch aufzunehmen sind. Wie kann ein solches Wörterbuchkonzept gewonnen werden?

Die alte Schule

Traditionell haben sich Wörterbuchverlage auf Sprachkenner und Gewährspersonen und deren Zettelkästen verlassen, um das Wörterbuchkonzept zu entwerfen. Aber auch für erfahrene Lexikographen ist der Wortschatz einer Sprache nur unzuverlässig abschätz- und bewertbar; deren Entscheidungen hängen ab von deren Erfahrungsraum, Intuition und Introspektion.

Verlegerische Entscheidungen liefen deshalb in der Regel darauf hinaus, unter Einbezug vieler Gewährspersonen möglichst umfängliche Universalwörterbücher anzustreben.

Die Wirtschaftlichkeit solcher kostenintensiven Unternehmungen wurde durch gering veränderte Wiederauflagen über lange Zeit hinweg gesichert. Daneben verbesserte die Publikation von Auszügen als Spezialwörterbücher (für Schüler, Reisewörterbücher, Fachwörterbücher, usw.) die Amortisation der Investition.

Language Engineering

Computergestütztes Language Engineering ermöglicht es, dem Erfahrungsraum der Lexikographen eine Sprachdokumentensammlung der Quellsprache als primäre Instanz zur Seite zu stellen. Der Intuition und Introspektion der Lexikographen wird die statistische Analyse dieser Sammlung beigestellt – der fachsprachliche Terminus „korpusbasierte Lexikographie“.

Der Wortschatz der Dokumentsammlung wird statistisch beschrieben. Es werden die Worthäufigkeiten, Wortverteilungen und Wortkollokationen ausgerechnet. Das Wörterbuchkonzept wird

nach messbaren Kriterien berechnet. Z. B. werden die 60.000 häufigsten Wörter als Einträge des zu erstellenden Wörterbuchs festgelegt.

Die Zuordnung eines zielsprachlichen Übersetzungsäquivalents zu einem quellsprachlichen Wort geschieht nach wie vor durch den Lexikographen.

Eine Variante des beschriebenen Verfahrens kann zum Design der zu verwendenden Übersetzungsäquivalente angewendet werden:

Der Lexikograph hat mit der Sprachdokumentensammlung die Instanz an der Hand, die er konsultieren und nach der er entscheiden kann. Er trifft seine Entscheidung in Kenntnis des statistischen Verhaltens des jeweiligen Wörterbucheintrags in einer großen Textsammlung. Den Zugang zur Dokumentensammlung und deren statistischer Analyse ermöglicht ein spezieller lexikographischer Computer-Arbeitsplatz.

Der Mehrwert

Das Wörterbuch ist hinsichtlich seiner Qualität für den in der Sprachdokumentsammlung repräsentierten Sprachraum beschrieben. Es kann gemessen werden, ob die erwünschte Qualität erreicht ist.

Der voraussichtliche Abdeckungsgrad des Wörterbuchs in der operationellen Anwendung kann proaktiv mit Hilfe von Kontroll-Sprachdokumentensammlungen geprüft werden.

Ist in den lexikographischen Arbeitsplatz ein zielsprachliches Textkorpus eingebunden, das derselben Domäne zuzurechnen ist, so gibt der Abgleich der relativen Häufigkeit der quellsprachlichen Wörter mit denen der Übersetzungsäquivalente im Korpus ein Maß für die Korrektheit der Übersetzung selbst.

Im Betrieb eines Wörterbuchs durch den Wörterbuch-Server im operationellen Feld werden Qualitätsminderungen des Wörterbuchs unmittelbar beim Auftreten festgestellt und an den lexikographischen Arbeitsplatz gemeldet. Dadurch werden unmittelbar und zeitnah Rücksteuerungen als Korrekturen/Erweiterungen des Wörterbuchs möglich.

Die Arbeit der Lexikographen geschieht computergestützt und ist weitgehend durch in den lexikographischen Arbeitsplatz integrierte Verfahren gesteuert. Die Notwendigkeit Gewährsleute einzubinden entfällt. Dies reduziert sowohl Personal- wie Zeitaufwand.

Voraussetzungen des Verfahrens

Um das beschriebene Modell des Language Engineering für Wörterbücher erfolgreich durchzuführen, sind Voraussetzungen zu erfüllen:

Dokumentensammlung

Die Sprachdokumentsammlung soll den Wortbestand der Domaine, die das Wörterbuch abdecken soll, widerspiegeln.

Die Sprachdokumentsammlung soll eine ausreichende Textmenge enthalten, so dass darin alle linguistischen Eigenschaften der Domaine enthalten sind. Dies kann durch automatisierte statistische Verfahren sichergestellt werden.

Die Sprache einer Domaine entwickelt sich ständig. Deshalb ist die Sprachsammlung vorzugsweise so anzulegen, dass sie kontinuierlich an einem einlaufenden Sprachdatenstrom gemessen und – wenn Abweichungen erkannt – gezielt gepflegt werden kann. Sie stellt kein ein für alle Mal abgeschlossenes Gebinde dar; der Verstoß gegen ein solches Monitoring leistet der Veralterung des aus der Sammlung abgeleiteten Wörterbuchs Vorschub.

Wortform – Grundwort

Ein Wörterbuch enthält in jedem Wörterbucheintrag das Grundwort. In Sprachsammlungen finden wir jedoch die einzelnen Wortformen. Wortformen sind unterschiedliche Ausprägungen von Wörtern. Sie sind in morphosyntaktischen Grammatiken beschrieben.

Ein relevanter Teil der Bedeutung eines Wortes (gemeinhin als „inhaltliche Bedeutung“ bezeichnet) gilt aber, von überschaubaren Ausnahmen abgesehen, für alle Formen eines Wortes. Die Wortform fügt zur „inhaltlichen“ Wortbedeutung einen von Sprachwissenschaftlern „morphosyntaktisch“ genannten Bedeutungsanteil hinzu.

Wie kann der Bezug zwischen den Wortformen in der Dokumentsammlung und den Grundwörtern im Wörterbuch hergestellt werden, um den Wörterbuchaufbau wirtschaftlich und fehlerresistent zu gestalten?

Lemmatisierung – Dabei sind zwei Verfahren im Einsatz

Die Wortformen werden den Grundwörtern (Lemmata) zugeordnet. Das Verfahren ist computergestützt organsiert durch „Tokenizing“ der vorhandenen Textsammlungen und Reduzierung der ermittelten „Tokens“ zu „Types“.

Es steht ein Set von morphosyntaktischen Regeln bereit. Mit diesen Regeln kann auf zweierlei Weise verfahren werden:

Analytisch: Angefragte Wörter werden vor dem Abgleich mit dem Wörterbuch anhand dieser Regeln analysiert und durch die morphosyntaktisch möglichen Grundwörter ersetzt.

Generativ: Zu den Wörtern des Wörterbuchs werden alle morphosyntaktisch korrekten Wortformen generiert und mit den zugehörigen Einträgen des Wörterbuchs verbunden. Die Generierung der morphosyntaktisch möglichen Wortformen für das Arabische (und je nach ihren morphologischen Eigenschaften auch andere neu zu bearbeitende Sprachen) kann von den Wortwurzeln aus gestartet werden, was den Effekt erzeugt, bereits die Wortformen aller in der zukünftigen Entwicklung des Arabischen möglichen Wortformen mit Wörterbuchenträgen zu verbinden.

Beispiele:

Die Generierung von Wortstämmen zur arabischen Wurzel **ءبَيدي**:

ءبَيدٌ ءبَيديا ءبَيدٍ ءبَيدُ ءبَيدَ ءبَيدِ ءبَياديتا ءبَيدونَ ءبَيديَنَ ءبَيديانِ ءبَيدَيْنِ ءبَياديتا ءبَيدو ءبَيديا

Die Generierung von Wortformen zum somalischen Grundwort **aabbe**:

aabbaha aabbahaa aabbahaan aabbahaas aabbahan aabbehee aabbeheer aabbihii aabbohoo aabbuhu aabbayaal aabbayaasha aabbhayga aabbhaaga aabbhiisa aabbheeda aabbhayaga aabbheenna aabbhiinna aabbhooda aabbhaygaa aabbhaagaa aabbhiisaa aabbheedaa aabbhayagaa aabbheennaa aabbhiinna aabbhooda aabbhaygaa aabbhaagaa aabbhiisaa aabbheedaa aabbhayagaa aabbhiinnaa aabbhoodaa aabbhaygii aabbhaagii aabbhiisii aabbheedii aabbhayagii aabbhayagii aabbhiinnii aabbhoodii aabbeed aabbhaygu aabbhaygu aabbhusu aabbheedu aabbhayagu aabbheennu aabbhunnu aabbhoodu aabbayaashayda aabbayaashaada

Das Verfahren setzen ein auf die Grundwörter normiertes Wörterbuch voraus. Liegt kein auf die Grundwörter hin normiertes Wörterbuch vor, so kann dies aus einer Wortliste computergestützt mit Hilfe des oben beschriebenen aber iterativ angewandten Verfahrens „selbstlernend“ erzeugt werden.

Mehrdeutigkeit von Wörtern

Wörter sind – außerhalb eines konkreten Kontextes – meistens mehrdeutig (polysem). Das Wörterbuch hat für mehrdeutige Wörter pro Bedeutung, soweit in der zu bearbeitenden Domaine aufgetreten, einen Wörterbuch-Eintrag und legt dem Wörterbuchbenutzer (Mensch oder Computerprogramm) alle Wörterbucheinträge zum betroffenen Wort vor.

Mehrdeutigkeit eines Wortes einer Sprache kommt zustande, weil es in sprachlichen Kontexten in verschiedenen Wortnachbarschaften Unterschiedliches bedeutet. Ein Wort ist in ein „Bedeutungsgeflecht“ mit anderen Wörtern verwoben. Je nach den Wörtern im Kontext wird ein Teil dieses Geflechts als Bedeutung realisiert. Die Darstellung solcher Bedeutungsbeziehungen zwischen Wörtern ist eine anspruchsvolle Anforderung, der nur wenige Wörterbücher nachkommen.

Zwischen den verschiedenen Bedeutungen und Bedeutungsnuancen eines Wortes bestehen keine klaren Relationen, sondern es besteht eher eine „Vagheit“ im Bedeutungsgeflecht, erzeugt durch die Wörter im Kontext. Dies begründet die Fähigkeit, in Sprache Neues zu formulieren. Die Mitglieder einer Sprachgemeinschaft nutzen die Vagheit, die Bedeutungen zugrunde liegt, z. B. für Sprachspiele, für Dichtung, aber auch um Neues in der Kultur, der Politik und der Wissenschaft usw. zu beschreiben.

Statistische Beschreibung (Korrelations-, Korrelationsanalysen) des Wortumfeldes eines Wortes in einer Sprachdokumentsammlung erlaubt es, die Bedeutung eines Wortes mit mathematischen Methoden zu beschreiben und auszudifferenzieren. Dies gibt dem Lexikographen bei seiner Übersetzungsaufgabe gemessene, über sein eigenes sprachliches Universum hinausgehende Information, um Mehrdeutigkeiten aufzulösen.

Die Analysen können für das quellsprachliche und für ein ähnliches, zielsprachliches Korpus der Domäne vorgenommen werden.

Der Lexikograph kann am lexikographischen Arbeitsplatz auf die Ergebnisse der Analysen zugreifen und als Entscheidungshilfe nutzen. Ein solchermaßen erweitertes Verfahren im Wörterbuch-Produktionsprozess reduziert die Produktionszeit und erhöht den Gebrauchswert eines Wörterbuchs zusätzlich.