Statistische Übersetzung von technischen Dokumentationen

Bachelorarbeit im Fach Informatik

vorgelegt von

Katrin Affolter

Münchenstein, Schweiz

Matrikelnummer 10-917-250

Angefertigt am Institut für Computerlinguistik
der Universität Zürich
Prof. Dr. Martin Volk

Betreuer: Dr. Mark Fišel

Abgabe der Arbeit: 30.9.2014

Zusammenfassung

Die vorliegende Bachelorarbeit beschäftigt sich mit der Erstellung eines aktuellen statistischen maschinellen Übersetzungssystems (SMÜ-Systems) für die Übersetzung von technischen Dokumentationen vom Deutschen ins Englische.

Hierfür werden als Erstes die domänenspezifischen Texte der Firma Finnova in einen paralleln Korpus verwandelt, um anschliessend mit Moses ein SMÜ-System zu trainieren. In einem zweiten Schritt werden weitere parallele Korpora mit Hilfe von Moses trainiert, um anschliessend mit dem domänenspezifischen Korpus zu einem besseren SMÜ-System verrechnet zu werden.

Die zweite Komponente meiner Arbeit besteht darin, die am häufigsten auftretenden Übersetzungsfehler zu untersuchen und Vorschläge zur Verbesserung zu machen.

Abstract

This bachelor thesis studies the creation of a state-of-the-art baseline system for translating technical documentation from German to English.

In the first step I transformed the domain specific texts supplied by the company Finnova to a parallel corpus, which I used to train a statistical machine translation using Moses.

In the second step I used the Moses SMT framework to train additional parallel corpora and combine those corpora with the domain specific corpus to compute a better statistical machine translation.

The second component of my thesis was to analyse the most frequent translation errors and to make suggestions for improvement.

Danksagung

Ich bedanke mich ganz herzlich bei Prof. Dr. Martin Volk für die Möglichkeit meine Bachelor Arbeit im Bereich maschinelle Übersetzung zu machen.

Ausserdem bedanke ich mich bei Dr. Mark Fišel für seine Betreuung und Unterstützung.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

BLEU	BiLingual Evaluation Understudy
Hyp	Hypothese; in dieser Arbeit der vom SMÜ-System aus dem deutsche ins englische übersetzte Text/Satz.
MERT	Minimum Error Rate Training
METEOR	Metric for Evaluation of Translation with Explicit ORdering
NLTK	Natural Language ToolKit
Ref	Reference; in dieser Arbeit der englische zur Verfügung gestellte Text/Satz.
SMÜ	Statistische maschinelle Übersetzung
Src	Source; in dieser Arbeit der deutsche zur Verfügung gestellte Text/Satz.
TER	Translation Error Rate
XML	eXtensible Markup Language

Sprachkürzel (ISO 639-1)

DE	Deutsch
EN	Englisch

1 Einführung

Als maschinelle Übersetzung bezeichnet man das automatische Übersetzen von Texten der Ausgangssprache in die Zielsprache mit Hilfe eines Computerpogramms. Bei der statistischen maschinellen Übersetzung, kurz SMÜ, (Brown et al., 1993) beruht die Übersetzung auf Wahrscheinlichkeiten. In meiner Arbeit verwende ich ein phrasenbasiertes SMÜ-System (Koehn et al., 2003). Dafür muss vor dem eigentlichen Übersetzen dem System ein möglichst grosser, zweisprachiger Textkorpus zur Verfügung gestellt werden. Mit diesem Korpus werden anhand von mathematischen Verfahren (beispielsweise Häufigkeit) mögliche Übersetzungen für Wörter und Phrasen berechnet. Die berechneten Informationen werden in einem sogenannten Übersetzungsmodell gespeichert.

Übermittelt man dem trainierten SMÜ-System einen Text in der Ausgangssprache, übersetzt dieses den Text in die Zielsprache anhand der Wahrscheinlichkeiten der einzelnen Phrasen und Wörter.

1.1 Motivation

Statistische maschinelle Übersetzungen sind über die Jahre immer besser geworden. Je grösser die zur Verfügung gestellte Datenmenge ist, desto besser werden die Übersetzungsmodelle. Für die Auswahl der Texte gilt: je spezifischer die Texte sind anhand derer das Übersetzungsmodell trainiert wurde, desto ungenauer wird die Übersetzung bei einem Text sein, der nicht zu dieser Domäne gehört. So würde beispielsweise das Wort ‚*Pass*' in der Domäne „Gesetztestext" mit ‚*passport*' ins Englische übersetzt. Nimmt man aber die Domäne „Gebirge", so wird die Übersetzung ‚*pass*' lauten.

Das heisst die Auswahl geeigneter Texte mit denen trainiert werden soll, kann ausschlaggeben für die Qualität einer Übersetzung sein (3.4). Dafür liegt bei domänenspezifischen Übersetzungen meist keine grosse Textsammlung (Korpus) vor.

Ausserdem weisst jedes Sprachpaar unterschiedliche Problematiken auf. Beispielsweise ist die Übersetzung vom Deutschen ‚*die schwarze Katze*' ins Englische ‚*the black cat*' ohne Neuanordnung der Wörter möglich. Ins Französische muss es aber ‚*le chat noir*' heissen: das Adjektiv kommt nach dem Nomen.

In meiner Bachelorarbeit erstelle ich ein aktuelles (engl.: state-of-the-art) Übersetzungssystem für einen domänenspezifischen parallelen Deutsch-Englisch Korpus. Dieser Korpus wird von der Firma Finnova gestellt, dass heisst er ist in seinem Umfang sehr beschränkt und weist weitere Problemstellungen auf.

Die Firma nutzt für eine andere interne Domäne schon ein ähnliches System. Der Nutzen für sie besteht darin, dass Übersetzer in den meisten Fällen nur noch eine Optimierung vornehmen müssen. Dies verkürzt den Aufwand für die Übersetzer, zum Beispiel kann die Vorabübersetzung dabei helfen Synonymdifferenzen[1] zu minimieren.

Zum Abschluss wird eine Test-übersetzung erstellt, welche ich auf die am häufigsten auftretenden Fehler analysiere. Für diese stelle ich mögliche Vorgehensweisen zur Verbesserung vor.

1.2 Struktureller Aufbau der Arbeit

Meine Bachelorarbeit habe ich chronologisch aufgebaut, entsprechend den Schritten für die Erstellung des SMÜ-Systems und der Analyse.

Im zweiten Kapitel „Domänenspezifischer Korpus“ beschreibe ich die einzelnen Schritte, welche notwendig waren um die zur Verfügung gestellten Texte in einen parallelen Korpus (2.1) zu verwandeln und anschliessend das Grundgerüst für das SMÜ-System (2.2) zu trainieren. Um die späteren Auswirkungen der zusätzlichen Korpora besser zu verstehen, findet eine erste Evaluation (2.3) des SMÜ-Systems statt.

Im dritten Kapitel „Zusätzliche Korpora“ wird die Auswahl (3.1) der Korpora, die als zusätzliches Textmaterial verwendet wurden, erläutert. Anschliessend beschreibe ich das notwendige Aufarbeiten (3.2) und die Trainingsschritte (3.3). Im Anschluss werde ich im selben Kapitel auf die verschiedenen Kombinationsmöglichkeiten (3.4) des domänenspezifischen Korpus und der zusätzlichen Korpora eingehen. Danach findet die endgültige Evaluation (3.5) des SMÜ-Systems statt.

Im vierten Kapitel „Analyse der häufigsten Fehler“ diskutiere ich die am häufigsten auftretenden Fehler und mache Vorschläge zur Verbesserung.

Im letzten Kapitel „Schlusswort“ fasse ich meine Überlegungen zusammen und erläutere die nächstmöglichen Schritte zur Verbesserung des SMÜ-Systems.

[1] Ein Beispiel wäre wie zuvor beschrieben das deutsche Wort ‚*Pass*’ ins Englische ‚*passport*’ oder ‚*pass*’.

[2] Erhalten zusammen mit den XML-Dateien.

2 Domänenspezifischer Korpus

Der domänenspezifische Korpus von Finnova basiert auf Texten von einer technischen Dokumentation. Sie zeichnen sich durch folgende spezielle Merkmale aus:

- Teilsätze, beispielsweise Beschriftungen von Feldern
  ‚*Maximum in Bankwährung*'

- Technische Bezeichnungen
  ‚*Neu wird aufgrund eines* UPDATE*s oder* INSERT*s auf dem Objektstatus (*OM_AUF.OBJ_STAT_CD*) ein Datenbanktrigger ausgelöst* .'

- Englische Bezeichnungen im deutschen Text
  ‚*Es wird nur die aktuelle Transaktion (Bsp :* Global Order *) berücksichtigt* .'

Deshalb ist es notwendig bei der Tokentrennung (S. 12) darauf zu achten, dass beispielsweise technische Bezeichnungen wie ‚OM_AUF.OBJ_STAT_CD' nicht am Punkt getrennt werden.

Zunächst müssen die 398 XML-Dateien in einen paralleln Korpus verwandelt werden. Erst anschliessend kann mit dem effektiven Training des SMÜ-System begonnen werden.

2.1 Aufarbeitung

Während der Aufarbeitung werden aus den vorhandenen XML-Dateien die Textstücke und Sätze extrahiert, bereinigt und anschliessend die Sätze zueinander ausgerichtet (vgl. Abbildung 1). Erst dann spricht man von einem parallelen Korpus mit welchem man das SMÜ-System trainieren kann.

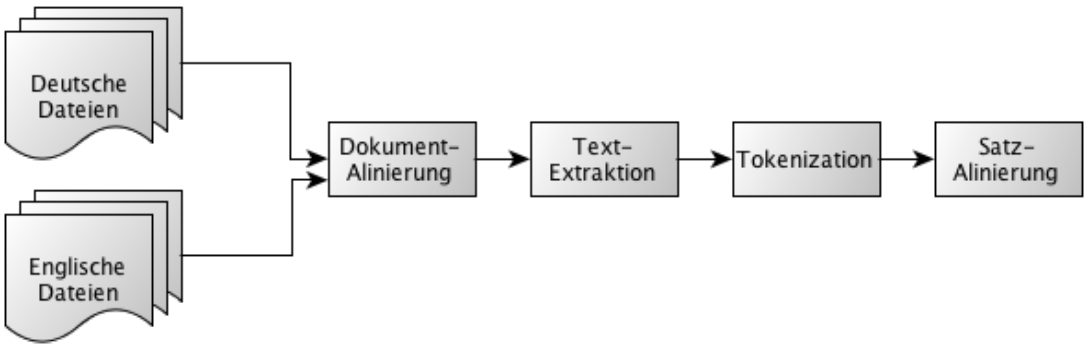


Abbildung 1: Systematische Darstellung zur Aufarbeitung des domänenspezifischen Korpus

2.1.1 Alinierung der Dokumente

Der erste Schritt zur Erstellung des parallelen Korpus besteht darin, die XML-Dateien einander zu zuordnen, so dass jeweils ein Dateipaar aus einer deutschen und einer englischen XML-Datei mit dem inhaltlich gleichen Text entsteht.

Im vorliegenden Fall ist dies eine relativ einfache Aufgabe, da die Dateinamen wie folgt aufgebaut sind:

```
ecl_id_sprache_beschreibung.xml
```

Das heisst ich konnte die Dateipaare anhand der gegebenen ID bilden. Es sind 188 Paare entstanden. Weitere 22 Dateien konnten nicht zugewiesen werden und sind somit nicht für den parallelen Korpus verwendbar.

2.1.2 Text extrahieren

Für die nächsten Schritte wird der Text aus den XML-Dateien extrahiert. Ein Blick in eine der Dateien ermöglicht einem die Struktur zu analysieren.

```
<tei:text xmlns:tei="http://www.tei-c.org/ns/1.0">
   <tei:body>
      <tei:div type="page-orientation" subtype="A4landscape">
         <tei:div type="chapter" subtype="h1">
            <tei:p>
               <tei:hi rend="h1">1 <tei:anchor xml:id="_Toc356478814"/>Zielsetzung dieses Dokumentes</tei:hi>
            </tei:p>
            <tei:p>Diese ECL dient dem Implementierer / Kunden als Drehbuch für seine Aktivitäten. Die ECL enthält
            <tei:p>Dokumenten-Owner ist der Finnova Entwicklungs-Projektleiter.</tei:p>
         </tei:div>
         <tei:div type="chapter" subtype="h1">
            <tei:p>
               <tei:hi rend="h1">2 <tei:anchor xml:id="_Toc356478815"/>Umgebungsmodell</tei:hi>
            </tei:p>
```

Abbildung 2: Ausschnitt aus einer der XML-Datei

Man erkennt, dass der zu extrahierende Text in *p*-Elementen des Namespaces *tei* steht. Weitere Verschachtelungen wie beispielsweise die Überschriften sind zustätzlich zu beachten. Für die Extrahierung wurde das Skript *extract_from_tei.py*[2] verwendet. Ausserdem werden binäre Objekte ausgelassen. Die Texte werden neu in Textdateien mit der selben Bezeichnung gespeichert.

Aus den 188 Dateipaaren konnten im deutschen 205'274 Zeilen und im englischen 240'594 Zeilen extrahiert werden. Der Unterschied liegt unter anderem daran, dass mehrere Sätze auf einer Zeile sein können. Im folgenden Schritt wird dies korrigiert.

2.1.3 Normalisieren und Zerlegen

Als Nächstes muss dafür gesorgt werden, dass jeder Satz auf einer eigenen Zeile steht (Normalisieren). Dies ist der erste Schritt zur Verknüpfung von deutschen und den zu-

[2] Erhalten zusammen mit den XML-Dateien.

gehörigen englischen Sätzen, auf dessen Basis das SMÜ-System lernen wird. Hierfür wird das Modul „Sentence Tokenizer“[3] (deut.: Satzzerleger) von NLTK (Bird, 2006) verwendet. Dieser basiert im Englischen auf dem *Wall Street Journal* mit rund 469'000 Token, im Deutschen auf der *Neue Zürcher Zeitung* mit etwa 847'000 Token.

Da der Tokenizer auf Zeitungsartikel beruht, musste eine Optimierung hinzugefügt werden, welche die gängigsten Abkürzungen der Dokumente (vgl. Tabelle 1) abdeckt, an welchen nicht getrennt werden soll.

Sprache	**Abkürzungen**
Englisch	e.g. , E.g. , i.e. , incl. , Art.
Deutsch	resp. , bzw. , ca. , ev. , o.ä. , evtl. , Evtl. , Bsp. , bspw. , Kap. , Art.

Tabelle 1: Abkürzungslisten als Erweiterung des „Sentence Tokenizer“

Jetzt ist jeder Satz auf einer Zeile. Es fehlt noch, dass jedes Token getrennt, mit einem Leerzeichen, zu seinen Nachbarn steht.

Beispielsweise würde ‚*Wie geht es dir?*’ anschliessend als ‚*Wie geht es dir ?*’ geschrieben werden. Dies ist wichtig, um eine Tokenalinierung vornehmen zu können. Würden die Satzzeichen beispielsweise nicht von den vorgängigen Wörtern getrennt, würde der Eintrag in der Übersetzungstabelle die Kombination beinhalten. Bei diesem Beispiel würde also in der Übersetzung nur ‚*dir?*`→‚*you?*’ stehen. Nehmen wir einen zweiten Beispielssatz ‚*Ich gab es dir*.’. Da wir vorher keine Trennung vorgenommen haben, ist in unserer Übersetzungstabelle kein Eintrag für ‚*dir*.’ vorhanden, daher kann es nicht übersetzt werden. Mit einer Trennung der Token, ist es also möglich, mehr Fälle abzudecken.

Für die Tokenizierung wird das Skript *ctok.pl*[4] verwendet. Dabei wurde der Aufrufparameter *-c* verwendet, um zu verhindern das alle Token klein geschrieben werden. Im-Schritt „2.2.2 Truecase“ wird hierfür eine Optimierung vorgenommen.

Am Ende sind 1’768’793 deutsche Token und 2’215’185 englische Token vorhanden. Zu diesem Zeitpunkt liegt ein Unterschied in den Anzahl Sätzen vor, dies liegt einerseits an den sinngemässe Übersetzungen und andereseits an dem nicht perfekt arbeitenden Tokenizer.

[3] Sentence Tokenizer (Stand: 5.7.2014): http://www.nltk.org/_modules/nltk/tokenize/punkt.html

[4] Erhalten zusammen mit den XML-Dateien.

2.1.4 Satzalinierung

Die Dateipaare sind jetzt soweit vorbereitet, dass nun eine Alinierung der Sätze vorgenommen werden kann. Das Alinierungsprogramm muss herausfinden, welcher deutsche Satz in welchen englischen Satz übersetzt wurde. Hierfür wird die Satzlänge in beiden Sprachen zu Hilfe genommen und übereinstimmende Token wie Eigennamen und Zahlen als Ankerpunkte verwendet.

Für die Alinierung der Sätze wird das Programm HunAlign (Varga et al., 2005) verwendet. Ausserdem wurden folgende Aufrufsparameter übergeben, um ein möglichst gutes Ergebnis zu erzielen:

- `-realign`: Die Alinierung wird in drei statt einer Phase durchgeführt, um ein besseres Ergebnis zu erzielen. Nach der initialen Alinierung fügt der Algorithmus heuristisch Token zum Wörterbuch hinzu, basierend auf dem gemeinsamen Auftreten von identifizierten Nachbarsätzen. Anschliessend wird der Alinierungsprozess noch einmal durchgeführt, basierend auf dem grösseren Wörterbuch.
- `-utf`: Für die richtige Interpretation der Zeichen (beispielsweise Umlaute)
- `-text`: Rückgabe als Text

Da die Alinierung anfänglich kein Wörterbuch zur Verfügung hat, werden die Ergebnisse besser je mehr Daten dem Algorithmus zur Verfügung gestellt werden. Das heisst es werden möglichst viele der einzelnen, relativ kleinen Dateien zu möglichst grossen zusammengefasst. HunAlign kann maximal mit Dateien der Grössenordnung 49'000 Zeilen umgehen (Fehlercode 11: Segmentation fault), deshalb werden die 188 Dateien in sieben zufällige Päckchen zusammengefasst.

Nach der Alinierung werden jene Zeilen entfernt, welche mindestens eine der folgenden Kriterien erfüllen:

1. auf beiden Seiten gleich sind → keine Übersetzung vorhanden
2. auf einer der Seiten leer ist → keine Alinierung gefunden
3. mehrere Sätze wurden von HunAlign zusammen gezogen, gekennzeichnet mit „~~~“

Die Texte liegen jetzt als paralleler Korpus mit 1’647’513 deutschen Token und 1'917’790 englischen Token vor.

2.2 Training

Für das Trainieren eines Übersetzungsmodells wird Moses (Koehn et al., 2007) verwendet. Moses ist ein statistisches maschinelles Übersetzungssystem, welches ein Über-

setzungsmodell für jedes beliebige Sprachpaar mit Hilfe eines parallelen Korpus zu trainieren.

Bevor mit dem Training angefangen werden kann, müssen noch weitere Verarbeitungsschritte vorgenommen werden.

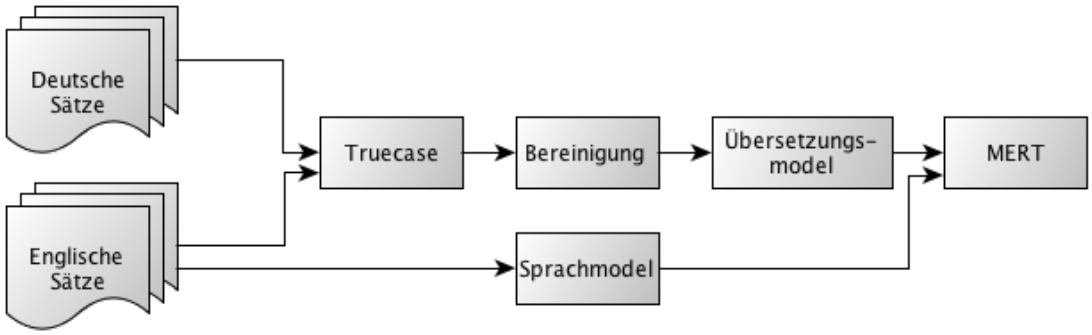


Abbildung 3: Systematische Darstellung zum Training des statistischen maschinellen Übersetzungssystems

Alle verwendeten Moses Aufrufe sind im Anhang A: Moses Befehle aufgeführt.

2.2.1 Aufteilen des Korpus

Als Erstes wird der Korpus in drei sogenannte Sets aufgeteilt:

- Test-Set: 3'988 Zeilen
- Dev-Set: 1'991 Zeilen
- Train-Set: 171'781 Zeilen

Das Test-Set wird zur Seite genommen, um am Ende das Übersetzungsmodell daran zu testen und zu evaluieren wie gut die Übersetzungen sind.

Das Dev-Set wird für Optimierungszwecke benötigt. Es wird beispielsweise für Minimum Error Rate Training (S.22) benutzt, dazu später mehr.

Das Train-Set ist der grösste Anteil und wird, wie der Name sagt, für das Training des SMÜ-Systems verwendet.

2.2.2 Truecase

Beim Truecasen geht es darum, dass alle Token in ihrer richtigen Gross-/Kleinschreibung geschrieben werden. Der verwendete Algorithmus ist simpel: für jedes Wort wird die am häufigsten gefundene Schreibweise verwendet.

Der Truecaser wird als Erstes mit Hilfe des Train-Sets trainiert. Anschliessend wird sowohl das Train-Set als auch das Dev-Set damit überarbeitet.

```
$ ./moses/scripts/recaser/train-truecaser.perl --model \
./temp/truecase-model.de --corpus ./finnova/train.de

$ ./moses/scripts/recaser/truecase.perl --model \
./temp/truecase-model.de < ./finnova/train.de > \
./finnova/train.truecase.de

$ ./moses/scripts/recaser/truecase.perl --model \
./temp/truecase-model.de < ./finnova/dev.de > \
./finnova/dev.truecase.de
```

Programm 1: Truecasen des deutschen Korpus

2.2.3 Bereinigung

Der Korpus sollte an dieser Stelle bereinigt (engl.: cleaning) werden. Der Grund hierfür ist, dass die Wortalinierung mit langen Sätzen rechenintensiv ist, aber die Präzision nicht zunimmt (Sennrich, 2009). Daher werden Zeilen mit mehr als 80 Token entfernt.

Zusätzlich werden Zeilen mit weniger als einem Token entfernt, hiermit wird noch einmal sicher gestellt, dass keine Leerzeilen vorkommen. Das Bereinigungsskript von Moses entfernt zusätzlich auch alle überflüssigen Zwischenraumzeichen (engl.: white space). Dadurch werden beispielsweise doppelte Leerzeichen auf nur ein Leerzeichen reduziert.

```
$ ./moses/scripts/training/clean-corpus-n.perl \
./finnova/train.truecase de en ./finnova/train.clean 1 80
```

Programm 2: Bereinigung des parallelen Korpus

Am Ende sind noch 171’593 Zeilen vorhanden. Das endgültige Train-Set enthält:

- 1'583’682 Token im Deutschen (~9 Token pro Satz)
- 1’846’677 Token im Englischen (~11 Token pro Satz)

2.2.4 Sprachmodell

Ein Sprachmodell (engl.: language model) misst wie wahrscheinlich es ist, dass eine Sequenz von Wörtern von einem Redner, mit Muttersprache Englisch, gesagt wird. Dies hilft dabei, dass die Übersetzung nicht nur in ihrer Bedeutung korrekt ist (Übersetzung von Wort zu Wort), sondern auch in der Satzstellung und damit einen fliessenden englischen Satz ergeben. Man kann also sagen, ein Sprachmodell nimmt einen englischen Satz und gibt die Wahrscheinlichkeit zurück, dass dieser von einem englischen Muttersprachler stammt.

Ein wahrscheinlichkeitsbasiertes Sprachmodell p_{LM} sollte deshalb korrekte Satzstellung bevorzugen:

$$p_{\mathrm{LM}}(\text{the cat is black}) > p_{LM}(\text{black the is cat}) \tag{1}$$

Zusätzlich zur Satzstellung kann das Sprachmodell auch helfen die korrekte Übersetzung aus verschiedenen Möglichkeiten zu wählen. Zum Beispiel kann das deutsche Wort ‚*Haus*' im Englischen sowohl mit ‚*home*' als auch mit ‚*house*' übersetzt werden:

$$p_{\mathrm{LM}}(\text{I am home}) > p_{LM}(\text{I am house}) \tag{2}$$

Die Wahrscheinlichkeit für eine Sequenz von n Token wird, anhand der Wahrscheinlichkeit das Token n nach den 1 bis $n-1$ Token vorkommt, berechnet. Für eine Sequenz von 3 Token würde die Formel wie folgt aussehen:

$$p(n_3|n_1,n_2) = \frac{\mathrm{anz}(n_1,n_2,n_3)}{\sum_t \mathrm{anz}(n_1,n_2,n)} \tag{3}$$

In einem Korpus, mit Texten aus dem Europäischen Parlament, beispielsweise ist die Wahrscheinlichkeit für $p(\text{cross}|\text{the,red}) = 0.547$, die Wahrscheinlichkeit für $p(\text{tape}|\text{the,red}) = 0.138$. Daraus kann man schliessen, das ‚*cross*' viermal häufiger nach ‚*the red*' vorkommt als ‚*tape*'.

In dieser Arbeit wurde das IRSTLM toolkit (Federico et al., 2008) verwendet, um das Sprachmodell zu erzeugen. Der erste Schritt mit IRSTLM ist es, Satzgrenzensymbole zu setzen.

```
$ ./irstlm/bin/add-start-end.sh < ./finnova/train.clean.de > \
./finnova/train.sb.de
```

Programm 3: Satzgrenzensymbole setzen

Als Nächstes wird das Sprachmodell erzeugt. Als Konfigurationseinstellung wird *-n 5* gesetzt um ein Sprachmodell mit 1 bis 5-Grammen[5] zu erzeugen. Ausserdem *-p* um Singletons zu aktivieren um eine kleinere Sprachmodelldatei zu erzeugen.

```
$ export IRSTLM=./irstlm; ./irstlm/bin/build-lm.sh \
-i ./finnova/train.sb.de -t ./tmp -p -s improved-kneser-ney -n 5 \
-o ./finnova/train.lm.de -k 10
```

Programm 4: Sprachmodell erzeugen

[5] Gruppierung von einem bis zu fünf nacheinander folgenden Token.

Als Nächstes wird die Sprachmodelldatei in das nötige ARPA[6]-Format verwandelt.

```
$ ./irstlm/bin/compile-lm --text=true ./finnova/train.lm.de.gz
./finnova/train.arpa.de
```

Programm 5: Sprachmodelldatei in ARPA Format speichern

2.2.5 Übersetzungsmodell

Das Übersetzungsmodell ist der Kern jedes SMÜ-Systems: es beinhaltet die Wahrscheinlichkeiten dass ein n-Gramm der Ausgangssprache, in ein n-Gramm der Zielsprache übersetzt werden sollte.

```
$ ./moses/scripts/training/train-model.perl -root-dir . \
-corpus ./finnova/train_clean -f de -e en \
-alignment grow-diag-final-and -reordering msd-bidirectional-fe \
-lm 0:5:$HOME/finnova/train_blm.de:8 --parallel -cores 5 \
-external-bin-dir ./mgizapp/bin/ -mgiza -mgiza-cpus 5 \
--write-lexical-counts
```

Programm 6: Vollständiger Aufruf für das Übersetzungsmodell

Der Trainingsprozess wird von Moses in neun Schritten durchgeführt:

1. **Vorbereiten der Daten**
   Als Erstes wird für jede Sprache eine Vokabulardatei erstellt in der jedem Wort eine einzigartige Zahl (ID) zugeordnet wird. Ausserdem wird bestimmt wie häufig dieses Wort im Korpus erscheint. Anschliessend wird der parallele Korpus auf dieses Format angepasst. Jeder Satz wird durch drei Zeilen dargestellt: Die erste Zeile ist immer 1, diese Zahl kann angepasst werden, sie dient als Gewichtung. Die zweite Zeile spiegelt den deutschen Satz wider, wobei jedes Wort durch seine ID dargestellt wird. Die dritte Zeile ist der englische Satz dargestellt durch die IDs der englischen Wörter.

2. **GIZA++ ausführen**
   GIZA++ toolkit (Och & Ney, 2003) ist eine frei verfügbare Implementation der IBM (Och & Ney, 2003) und Hidden Markov Modelle (Och & Ney, 2003). Der Algorithmus ermöglicht, dass ein Wort in der Ausgangssprache zu mehreren in der Zielsprache aliniert werden kann, aber nicht umgekehrt (vgl. Abbildung 4, S. 20). Daher wird die Alinierung zweimal durchgeführt, einmal mit Deutsch als Ausgangssprache, das zweite Mal mit Englisch.
   Zu diesem Zeitpunkt sind keinerlei Informationen über einen Zusammenhang zwischen den deutschen und englischen Wörtern bekannt. Daher ist es notwendig einen iterativen Prozess zu starten.

[6] ARPA ist ein Textformat welches weniger Speicherplatz benötigt.

Der Algorithmus nimmt als Startwert für jede mögliche Alinierung die gleiche Gewichtung an. Anschliessend werden alle möglichen Alinierungen eines Wortes der Ausgangssprache anhand der grössten Wahrscheinlichkeit neu abgeschätzt. Als Berechnungsgrundlage für die Abschätzung dienen die alternativen Alinierungen.
Dieser Schritt wird so oft wiederholt bis eine Annäherung stattfindet, das heisst die alte und neue Gewichtung der Alinierungen sind einander annährend gleich.

Folgendes Beispiel zur Veranschaulichung:

das Haus das Buch ein Buch

the house the book a book

Initialisierung: alle Alinierungen haben die gleiche Wahrscheinlichkeit

das Haus das Buch ein Buch

the house the book a book

Nach der ersten Iteration sind die Alinierungen für ‚*das→the*' und ‚*Buch→book*' wahrscheinlicher als die Alternativen, da sie häufiger (zweimal) vorkommen. Für ‚*ein→?*' und ‚*Haus→?*' sind alle möglichen Alinierungen immer noch gleich wahrscheinlich.

das Haus das Buch ein Buch

the house the book a book

Nach erreichen der Konvergenz ist die Alinierung eindeutig.

3. Wortalinierung

Die Wortalinierung basiert sowohl auf den zwei GIZA++-Alinierungen als auch Heuristika. Als Alinierungsmethode wird `grow-diag-final-and` verwendet.
Als Erstes werden die zwei Alinierungen *DE→EN* und *EN→DE* aus Schritt 2 zu einer Alinierung zusammen geführt, in dem nur die Übereinstimmungen behalten werden (vgl. Abbildung 4, S. 20).
Als Nächstes werden jene Alinierungspunkte hinzugenommen, welche benachbart zu einer Übereinstimmung sind.
Als Letztes werden jene Alinierungspunkte hinzugefügt, bei welchen in beiden Sprachen das Wort noch keine Alinierung aufweisen.

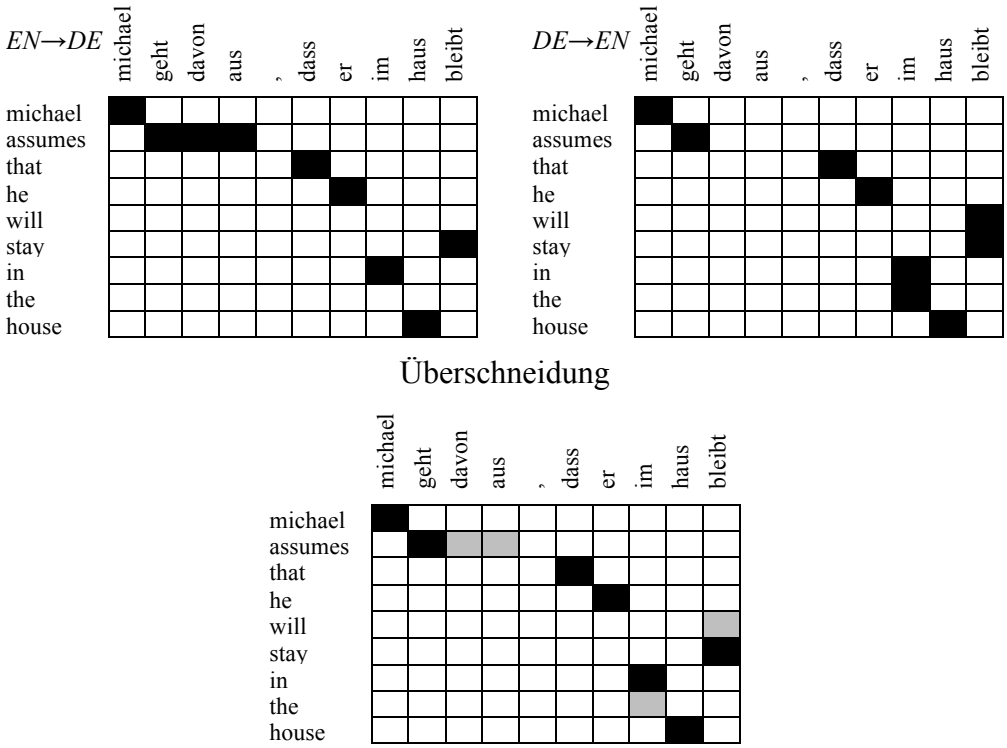


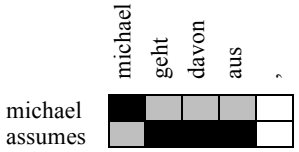
Abbildung 4: Schwarze Felder sind die Überschneidungen aus beiden Alinierungen. Graue Felder sind jene, welche nur in einer Alinierung vorgekommen sind. (Koehn, 2010)

4. Lexikalische Übersetzungstabelle

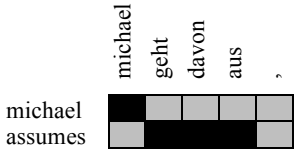
Für jede so ermittelte Alinierung wird ihre maximale Wahrscheinlichkeit berechnet. Diese Berechnung findet wieder für beide Richtungen statt.

5. Phrasen extrahieren

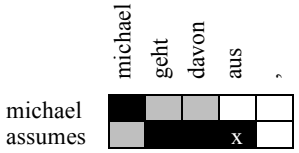
Anhand der Alinierungen können jetzt auch Phrasen extrahiert werden. Diese werden zusammen mit ihren Alinierungskoordinaten in eine Datei geschrieben.
Eine Phrase wird definiert als eine Alinierung in der alle deutschen Wörter Alinierungspunkte zu den englischen Wörtern innerhalb der Phrase haben und umgekehrt.
Beispielsweise eine Phrase aus Abbildung 4:



Korrekt: alle Alinierungspunkte befinden sich innerhalb der Phrase



Korrekt: das Komma hat keine Alinierung, daher verletzt es die Regel nicht.



Falsch: ‚*assumes'* hat zusätzlich ein Alinierungspunkt zu ‚*aus'* welcher ausserhalb der Phrase liegt.

Gespeichert werden alle möglichen Phrasenalinierungen. Es würde für dieses Beispiel daher sowohl ‚*michael geht davon aus→michael assumes*' als auch ‚*michael geht davon aus ,→michael assumes*' gespeichert werden.

6. **Phrasen bewerten**

   Für die extrahierten Phrasen werden vier Werte berechnet:

   - Inverse Phrasenübersetzungswahrscheinlichkeit $\varphi(\mathrm{DE}|\mathrm{EN})$
   - Direkte Phrasenübersetzungswahrscheinlichkeit $\varphi(\mathrm{EN}|\mathrm{DE})$
   - Inverse lexikalische Gewichtung $lex(\mathrm{DE}|\mathrm{EN})$
   - Direkte lexikalische Gewichtung $lex(\mathrm{EN}|\mathrm{DE})$

7. **Neuanordnungsmodell** (engl.: reordering model)

   Es werden die Kosten für die Neuanordnung eines Satzes in der Zielsprache berechnet. Als Modelltyp wird eine wortbasierende Extraktion verwendet.

   Die Neuanordnung darf sowohl monoton (gleich wie in der Ausgangssprache), vertauschend (beispielsweise ‚*Gestern kaufte John Obst*' wird zu ‚*Yesterday, John bought apples*') und unterbrechend (beispielsweise ‚*Ich gab dir das Buch*' wird zu ‚*I gave ~~to you~~ the book to you*') sein. Dies wird mit dem Konfigurationsparameter `--reordering msd` (engl.: monotone, swap and discontinuous) gesteuert.

   Zusätzlich wird mit `bidirectional` festgelegt, dass das Modell sowohl die Phrase vorher als auch nachher beachten soll. Mit `fe` wird festgelegt, dass sowohl die Ziel- als auch die Ausgangssprache in die Entscheidungen für das Modell eingebunden werden soll.

   Die vorgenommenen Konfigurationen sind die am häufigsten verwendeten.

8. **Generationmodell**

   Das Generationmodell wird aus der Zielsprache des parallelen Korpus erstellt.

   In meiner Arbeit wird dieser Schritt nicht benötigt.

9. **Konfigurationsdatei**

   Als Letztes wird eine Konfigurationsdatei moses.ini für den Decoder erstellt, welche alle Pfade für das Übersetzungsmodell sowie die Einstellungen der default Gewichtung enthält.

2.2.6 Minimum Error Rate Training

Als Letztes wird eine Optimierung mit MERT (Och, 2003) vorgenommen. Alle bis jetzt berechneten Modelle werden für die Übersetzung benötigt. Die darin enthaltenen Wahrscheinlichkeiten werden vom Decoder verrechnet um eine Übersetzung zu generieren. Dabei werden die verschiedenen Modelle unterschiedlich gewichtet.

Diese Gewichtungen werden mit Hilfe von MERT optimiert. Hierfür wird das deutsche Dev-Set übersetzt, dabei wird aber nicht nur die beste Übersetzung angesehen, sondern die n-Besten. Anschliessend werden diese n-besten Übersetzungen mit dem englischen Dev-Set mit Hilfe der Evaluationsmetrik BLEU-Score verglichen (siehe S. 24).

Es werden diejenigen Gewichtungen ermittelt, welche die besten Übersetzungen über das gesamte Dev-Set ergeben. Mit diesen neuen Gewichten wird wieder das deutsche Dev-Set übersetzt und die n-besten Übersetzungen betrachtet.

Dies ist ein iterativer Prozess der entweder nach einer fixen Anzahl an Durchgängen abgebrochen wird, oder wenn sich die neue und die alte Gewichtung annähern (vgl. Abbildung 5).

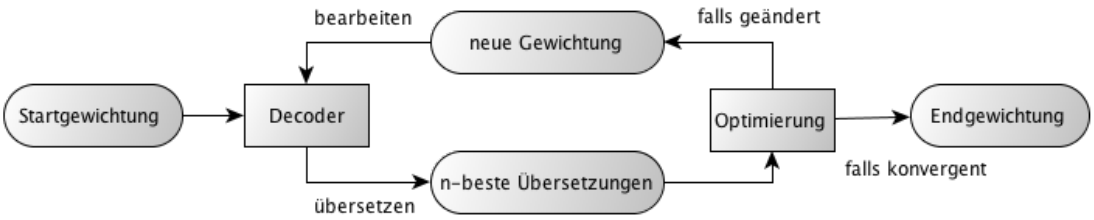


Abbildung 5: Systematische Darstellung des iterativer Prozess von MERT (Koehn, 2010)

Da die Optimierung von der Auswahl der Startwerte abhängig ist, wird nicht nur einmal mit MERT optimiert sondern mindestens dreimal.

```
$ ./moses/scripts/training/mert-moses.pl ./finnova/dev.truecase.de \
./finnova/dev.truecase.en ./moses/bin/moses ./model/moses.ini \
--mertdir ./bin/ --decoder-flags="-threads 10"
```

Programm 7: Aufruf für einen MERT-Durchlauf

2.3 Evaluation

Bevor zusätzliches Material hinzugefügt werden kann, sollte das aktuelle SMÜ-System getestet werden, um an Ende einen Vergleichswert zu haben. Hierfür wird das Test-Set vom Deutschen mit Hilfe des trainierten SMÜ-Systems ins Englisch übersetzt.

Für die Evaluation wird die englische Version des Test-Sets als Referenz verwendet und mit dem übersetzten Text verglichen (vgl. Abbildung 6).

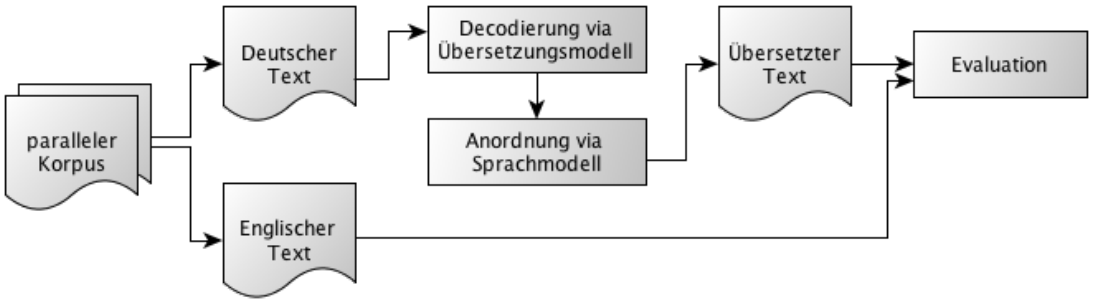


Abbildung 6: Systematische Darstellung der Übersetzung und Evaluation des SMÜ-System

2.3.1 Übersetzung

Dem Decoder wird ein Satz in der Ausgangssprache s übergeben, welcher diesen in einen Satz der Zielsprache t übersetzen soll. Das Ziel des Decoders ist, jene Übersetzung $\hat{t}$ zu finden, für die gilt:

$$\hat{t} = \arg\max_t p(t|s) \tag{4}$$

Das heisst, es soll jene Übersetzungen $\hat{t}$ aus allen möglichen Übersetzungen t genommen werden, so dass die Wahrscheinlichkeit am grössten ist.

Dabei wird nicht nur die Wahrscheinlichkeit für die Übersetzung einzelner Phrasen verwendet, sondern alle berechneten Modelle wie Sprachmodell, Neuanordnungsmodell, „Nachteil“ (engl.: penalty) von Wörtern und Phrasen. Jedes dieser Features hat zusätzlich eine Gewichtung, die berücksichtigt wird (vgl. Abschnitt 2.2.6).

Ausserdem werden XML-Tags zugelassen, welche zusätzlich vom Decoder beachtet werden müssen. Dies kommt vom Schritt 2.1.3 in dem ein Tokenizer verwendet wurde, welcher XML-Tags benutzt. Einer der zu übersetzenden Sätze sieht beispielsweise wie folgt aus:

> *Beispiele für Testobjekte sind Anlagevorschlag ,* <mask type="generic_mixcase" translation="GlobalOrder">*Globalorder*</mask> *, Kontrollliste ,* <mask type="ucase_id" translation="B2_AUF">*B2_AUF*</mask> *oder* <mask type="ucase_id" translation="B2_ABR">*B2_ABR* </mask> .

Jene Elemente, welche von einem XML-Tag umgeben sind, enthalten einen Übersetzungshinweis. Dieser soll vom Decoder berücksichtigt werden. Standardmässig ist der Umgang mit XML von Moses deaktiviert. Mit dem Konfigurationsparameter `-xml-input` lässt sich die Integration steuern. Es gibt drei verschiedene Möglichkeiten:

- `exclusive`: nur die Übersetzung im XML-Tag wird verwendet. Überlappungen von Phrasen in der Übersetzungstabelle werden ignoriert.

- `inclusive`: die im XML-Tag spezifizierte Übersetzung wird mit denen der Phrasentabelle verglichen.
- `constraint`: es werden nur Phrasen verwendet, welche die Übersetzung enthalten welche im XML-Tag stehen.

Für die Übersetzung wurde dem SMÜ-System das deutsche Test-Set übermittelt. Dies wurde für jede der drei möglichen XML-Konfigurationsparameter und jeweils mit den drei verschiedenen MERT Gewichtungen durchgeführt. Dies ist nötig um die beste Konfiguration herauszufinden (vgl. Tabelle 2).

2.3.2 Evaluationsmetriken

Zum Evaluieren der Übersetzungsqualität wurde das Script MultEval (Clark et al.,) verwendet. Dabei werden drei verschiedene Metriken untersucht: BLEU, METEOR und TER.

2.3.2.1 BiLingual Evaluation Understudy

BLEU ist ein Algorithmus der Texte mit einander vergleicht und angibt, wie ähnlich der zu vergleichende Text zu den Referenz-Texten ist. Je mehr Referenztexte vorhanden sind, desto aussagekräftiger wird der BLEU-Score.

Ab einem BLEU-Score über 30 kann man davon ausgehen, dass die Übersetzung verständlich ist. Über 50 ist es eine gute, fliessende Übersetzung (Lavie, 2010).

Typischerweise wird der BLEU-Score mit maximal 4-Grammen berechnet. Ausserdem wird die unterschiedliche Gewichtung für n-Gramm vernachlässigt (auf 1 gesetzt). Somit lautet die Formel:

$$\text{BLEU-4} = \min\left(1, \frac{\text{Übersetzung-Länge}}{\text{Referenz-Länge}}\right) * \prod_{n=1}^{4} \text{Präzision}_n \tag{5}$$

Präzision_n beschreibt das Verhältnis von richtigen n-Grammen im Verhältnis zu der totalen Anzahl an generierten n-Grammen in der Übersetzung.

Die Multiplizierung des Minimums ersetzt die Abdeckung (engl.: recall) und ist eine Bestrafung für zu kurze Übersetzungen.

Folgender Beispielssatz soll zum besseren Verständnis dienen. Der Referenzsatz wurde von einem Humanübersetzer erstellt, die Hypothese von einem maschinellen Übersetzungssystem.

Referenz: Isreali officials are responsible for airport safety

Hypothese: airport security Israeli officials are responsible

$\text{Präzision}_1 = 6/6$; $\text{Präzision}_2 = 3/5$; $\text{Präzision}_3 = 2/4$; $\text{Präzision}_4 = 1/3$

$$\text{BLEU-4} = 6/7 * \prod_{i=1}^{4} \text{Präzision}_i \cong 0.0857 = 8.57\%$$

An diesem Beispiel erkennt man auch die Vor- und Nachteile von BLEU. Einerseits erkennt es, bis zu einem gewissen Grad, Grammatik durch die n-Gramme. Andererseits werden nur exakt gleiche n-Gramme betrachtet. Synonyme einzelner Wörter können nicht erkannt werden und werden als falsch bewertet.

2.3.2.2 Metric for Evaluation of Translation with Explicit ORdering

METEOR verwendet weitere Funktionen, um eine genauere Aussage über die Qualität der Übersetzung zu ermitteln. Der Hauptunterschied zu BLEU liegt darin, dass METEOR zusätzlich Stammformreduktion, Bedeutungsklassen und Synonyme in die Bewertung einfliessen lässt.

Deswegen ist der Einfluss von mehreren menschlichen Übersetzungen als Referenz nicht so stark wie bei BLEU, dennoch gilt je mehr desto genauer wird die Evaluation.

Im Unterschied zum BLEU-Score kann man bei METEOR erst ab einem Wert von über 50 von verständlichen Übersetzungen sprechen und ab 70 von guter, fliessender Übersetzung (Lavie, 2010).

Abbildung 7: Alinierungsbeispiel für METEOR (Wikipedia, 2014)

Für die Berechnung wird als Erstes eine Alinierung auf Unigrambasis zwischen Referenzsatz und Übersetzungssatz vorgenommen.

Daraus wird die Unigrampräzision $P = \frac{m}{w_t}$, wobei m die Anzahl an Unigramme, die sowohl im Referenz als auch im Übersetzungssatz vorhanden sind. w_t ist die Anzahl von Unigrammen im Übersetzungssatz. Abdeckung $R = \frac{m}{w_r}$ wird berechnet mit der Anzahl von Unigramen in dem Referenzsatz w_r. Daraus lässt sich das harmonische Mittel mit einer Gewichtung berechnen: $F_{mean} = \frac{10*P*R}{R+9*P}$.

Als Nächstes werden möglichst viele der Unigrams zusammengefasst in möglichst wenige sogenannte Chunks (beispielsweise ‚*the cat*'). Wobei ein Chunk nur dann

erlaubt ist, wenn er in beiden Sätzen gleich ist. Mit deren Hilfe lässt sich der Nachteil (engl.: penalty) $p = 0.5 * \left(\frac{c}{u_m}\right)^3$ bestimmen, wobei c die Anzahl der Chunks ist und u_m die Anzahl von in Chunks zusammengefassten Unigrammen.

Das Endergebnis für einen Satz ist $M = F_{mean} * (1 - p)$. Der Nachteil p hat den Effekt, das Ergebnis zu reduzieren falls nicht mindestens ein Bigramm als Chunk gefunden wurde.

Für das Beispiel in Abbildung 7 würde die Berechnung wie folgt aussehen:

Ref: the cat sat on the mat $\quad P = \frac{6}{6} = 1 \quad R = \frac{6}{6} = 1$

Hyp: on the mat sat the cat

$F_{mean} = \frac{10*1*1}{1+9*1} = \frac{1}{1} = 1 \quad p = 0.5 * \left(\frac{4}{4}\right)^3 = 0.5 \quad \boldsymbol{M = 1 * (1 - 0.5) = 0.5}$

2.3.2.3 Translation Edit/Error Rate

Der Ansatz von TER ist ähnlich zur Levenshtein Distanz. Es wird berechnet wie viele Schritte notwendig sind, um die Übersetzung in die Referenz zu verwandeln. Dabei werden die Operationen Einfügen, Löschen und Vertauschen von Wörtern verwendet, sowie das Vertauschen von Chunks.

$$\text{TER} = \frac{\text{\# von Editieroperationen}}{\text{Durchschnittliche \# von Referenz Wörtern}} \tag{6}$$

Je höher der TER-Score ist, desto mehr Operationen sind für die Übersetzung notwendig. Ein TER-Score von 0 heisst daher, dass keine Veränderungen vorgenommen werden müssen, um den übersetzten Text in den Referenztext zu verwandeln.

Ein mögliches Beispiel wäre:

	apply	excluding	the	history
transfer	1	2	3	4
without	2	2	3	4
history	3	3	3	3

Dem grauen Pfad folgend würde dies bedeuten:

1. Ersetze ‚*transfer*' durch ‚*apply*'
2. Ersetze ‚*without*' durch ‚*excluding*'
3. Füge ‚*the*' hinzu
4. Behalte ‚*history*'

Daraus folgt TER $= \frac{3}{4} = 0.75$, oder anderst ausgedrückt: 75% des Satzes muss angepasst werden um den Satz ‚*transfer without history*' in die Referenz ‚*apply excluding the history*' zu verwandeln.

2.3.3 Auswertung der Ergebnisse

In der Tabelle 2 sind die Resultate von MultEval zusammengefasst. In der ersten Zeile ist das „rohe“ Übersetzungssystem ohne Optimierung durch MERT. Die anderen drei Zeilen enthalten die drei gleichen Durchläufe von MERT, aber unterschiedliche Angaben beim Umgang mit den XML-Tags.

Die Anwendung von MERT führt dazu, dass der BLEU-Score um 21% erhöht wird, dies entspricht fast einer Verdreifachung. METEOR wird auch um 13% verbessert. Der TER-Score sinkt um 20%, dies ist auch eine deutliche Verbesserung.

Das beste Ergebnis hat der Parameter `inclusive` geliefert, daher habe ich mich für diese Konfiguration in den weiteren Schritten entschieden.

	BLEU (s_sel[7]/s_opt[8]/p[9])	**METEOR** (s_sel/s_opt)	**TER** (s_sel/s_opt)
Ohne MERT	13.4 (0.3/0.0/-)	21.4 (0.2/0.0/-)	64.0 (0.4/0.0/-)
exclusive (3x MERT)	34.2 (0.4/0.7/0.0)	34.2 (0.2/0.1/0.0)	47.0 (0.4/1.0/0.0)
inclusive (3x MERT)	34.5 (0.4/0.6/0.0)	34.5 (0.2/0.1/0.0)	46.7 (0.4/0.9/0.0)
constraint (3x MERT)	34.3 (0.4/0.6/0.0)	34.3 (0.2/0.1/0.0)	46.9 (0.4/0.9/0.0)

Tabelle 2: MultEval für das domänenspezifische SMÜ-System

[7] Varianz aufgrund der Test-Set Auswahl

[8] Varianz aufgrund der Optimierer Instabilität

[9] Wahrscheinlichkeit der absoluten Differenz zwischen dem ersten (baseline) System und dem *n*-ten System.

3 Zusätzliche Korpora

Um das SMÜ-System zu verbessern, werden weitere parallele Korpora verarbeitet. Diese werden anschliessend mit dem domänenspezifischen Korpus verrechnet.

3.1 Auswahl

Als Erstes wurden zwei weitere Korpora von Finnova zur Verfügung gestellt, da diese firmenspezifische Daten enthalten. Diese zwei Korpora sind aber nicht vom selben Themengebiet wie jenes auf dem das SMÜ-System basiert.

- Korpus 1, BHB Dateien: 55'617 Sätze
- Korpus 2, HD Dateien: 173’718 Sätze

Ausserdem wurden drei öffentlich zugängliche Kopora verwendet. Diese sind umfangreicher. Ausserdem sind die Alinierungen und Übersetzungen korrekt.

- **Euro**pean **Parl**ament Proceeding Parallel Corpus (Koehn, 2005)
  Dieser Korpus besteht aus den Sitzungen des Europäischen Parlaments und liegt momentan in 21 Sprachen vor. Für das Sprachpaar Deutsch-Englisch sind 1'920'209 Sätze vorhanden. Im Deutschen sind es durchschnittlich 23 Wörter pro Satz, im Englischen 25 Wörter pro Satz.

- **Open Sub**title**s** (Tiedemann, 2009)
  Dieser Korpus ist die grösste vielsprachige Untertiteldatenbank. Es sind Untertitel in 30 Sprachen vorhanden. Für das Sprachpaar Deutsch-Englisch sind 5'269'507 Sätze vorhanden. Im Deutschen sowie im Englischen sind es durchschnittlich 6 Wörter pro Satz.

- **JRC-Acquis** (Steinberger et al., 2006)
  Dieser Korpus besteht aus Gesetztestexten in 22 Sprachen. Das Sprachpaar Deutsch-Englisch hat 1’285’187 Sätze. Im Deutschen sind es durchschnittlich 23 Wörter, im Englischen 25 Wörter pro Satz.

Übersetzt man das Test-Set nur mit einem dieser Korpora, so erhält man einen sehr niedrigen BLEU- und METEOR Score. Selbst die Ergebnisse mit MERT sind im Vergleich zum kleinen domänenspezifischen Korpus extrem schlecht. Dies zeigt, wie wichtig die richtige Domäne ist. Im Abschnitt 3.5 wird aufgezeigt, dass die zusätzliche Menge an Texten eine weitere Verbesserung erzeugt.

	BLEU (mit MERT)	**METEOR** (mit MERT)	**TER** (mit MERT)
domäne	13.4 (34.2)	21.4 (34.2)	64.0 (47.0)
Europarl	6.2 (20.0)	15.1 (24.0)	74.7 (60.2)
OpenSubs	3.9 (10.1)	8.7 (14.8)	84.0 (73.9)
JRC-Acquis	9.4 (17.3)	15.4 (22.3)	72.3 (64.8)

Tabelle 3: MultEval für den Vergleich von domänenspezifische versus domänenfremden SMÜ-Systemen

3.2 Aufarbeitung

Da die Texte schon als parallele Korpora vorliegen, müssen keine Aufarbeitungsschritte vorgenommen werden. Einzig beim JRC-Acquis Korpus musste das mitgelieferte Extraktionsskript ausgeführt werden.

3.3 Training

Jeder Korpus wird separat trainiert, gleich wie der domänenspezifische Korpus. Der Schritt „2.2.1 Aufteilen des Korpus“ wird ausgelassen, da nur ein Train-Set nötig ist.

Ausserdem werden beim Abschnitt „2.2.5 Übersetzungsmodell“ nur die Schritte bis und mit 6 ausgeführt. Alle nachfolgenden Schritte sind nicht nötig, da kein eigenes Übersetzungsmodell berechnet werden muss.

Der Output des Trainings wird im nächsten Schritt mit dem Übersetzungsmodell des domänenspezifischen Korpus verrechnet.

3.4 Kombinieren

TMCombine (Sennrich, 2012) ist ein Programm zur Kombination von Phrasentabellen, welche mit Moses generiert wurden. Der Konfigurationsparameter `combine_given_tuning_set` wird gesetzt um eine neue Phrasentabelle zu schreiben, in welcher die Gewichte so berechnet werden, dass die Kreuzentropie möglichst klein gegenüber dem Dev-Set is. Um dies zu ermöglichen muss ein Übersetzungsmodell (S. 18) bis und mit Schritt 5 durchgeführt werden. Die dabei generierte Datei `extract.sorted.gz` ist für die Neugewichtung relevant.

3.4.1 Interpolation

Bei diesem Verfahren möchte man verhindern, dass ein domänenspezifische Modell mit kleinem Vokabular bestraft wird. Es sollen nur fehlerhafte Übersetzungen bestraft werden.

```
$ ./tmcombine.py combine_given_tuning_set \
./model-finnova/ ./model-jrc/ ./model-europarl/ ./model-opensub/ \
./model-bhb/ ./model-hd/ -o ./phrase_table_interpolation \
-r ./model-finnova-dev/extract.sorted.gz
```

Programm 8: Aufruf von TMCombine mit Interpolation

3.4.2 Gewichtung

Bei der Gewichtung wird zusätzlich berücksichtigt, wie gut ein Phrasenpaar belegt ist. Das heisst es wird unterschieden zwischen negativen und mangelnden Beweisen.

Zuerst müssen die Phrasentabellen bereinigt werden:

```
$ zcat phrase-table.gz | sed -e "s/ |||$//g" | gzip > \
fixed-phrase-table.gz
```

Programm 9: Phrasentabellen bereinigen

Anschliessend kann TMCombine aufgeruft werden, mit dem Konfigurationsparameter `-m counts` um die Gewichtung zu aktivieren:

```
$ ./tmcombine.py combine_given_tuning_set \
./model-finnova/ ./model-jrc/ ./model-europarl/ ./model-opensub/ \
./model-bhb/ ./model-hd/ -o ./phrase-table-counts \
-r ./model-finnova-dev/extract.sorted.gz -m counts
```

Programm 10: TMCombine mit Konfiguration für die Gewichtung

Anschliessend muss die neue Phrasentabelle wieder für MERT korrigiert werden:

```
$ zcat phrase-table-counts.gz | sed s/\ nan\ /\ 0\ /g | gzip > \
fixed-phrase-table-counts.gz
```

Programm 11: Phrasentabellen korrigieren für MERT

3.5 Evaluation

Für die Evaluation werden beide Versionen mit MERT optimiert. Anschliessend werden die Ergebnisse mit der besten Version des domänenspezifischen SMÜ-Systems verglichen.

Die Ergebnisse zeigen, dass mit der Kombinationsvariante „Gewichtung“ eine stärkere Verbesserung erreicht wird. Es ergab sich dadurch eine Verbesserung des BLEU-Score und des METEOR-Score um +0.6 Prozent.

	BLEU (s_sel/s_opt/p)	**METEOR** (s_sel/s_opt)	**TER** (s_sel/s_opt)
Domäne (3x MERT)	34.5 (0.4/0.6/-)	34.5 (0.2/0.1/-)	46.7 (0.4/0.9/-)
Interpolation (3x MERT)	34.9 (0.5/0.3/0.0)	35.0 (0.2/0.1/0.0)	46.4 (0.4/0.4/0.0)
Gewichtung (3x MERT)	35.1 (0.5/0.4/0.0)	35.1 (0.2/0.1/0.0)	45.7 (0.5/0.5/0.0)

Tabelle 4: MultEval für die kombinierten SMÜ-Systeme

4 Analyse der häufigsten Fehler

Zur Analyse wurden 100 zufällige Sätze aus dem Test-Set extrahiert. Anschliessend wurden die Sätze manuell auf ihre Fehler untersucht. Von diesen 100 Sätzen waren neun perfekt übersetzt, dass heisst sie haben einen BLEU-Score von 100% erreicht.

Es wurden insgesamt 338 Fehler in 13 verschiedenen Fehlerkategorien gefunden:

- **Synonym**: fehlerhafte Auswahl von Synonymen.
- **Fehlerhafte Wortwahl**: im Gegensatz zu Synonymfehlern wurde hier definitiv das falsche Wort gewählt.
- **Komposita**: fehlerhaftes Übersetzen oder nicht Übersetzen eines zusammengesetzten Wortes
- **Unbekanntes Wort**: nicht Übersetzen eines Wortes (exkl. Komposita)
- **Satzstellung**: fehlerhafte Neuanordnung des englischen Satzes
- **Satzzeichen**: ein Unterschied, der entweder durch die Übersetzung entstanden ist, oder durch einen Unterschied zwischen dem deutschen und dem englischen Satz.
- **Numerus & Truecase**: Fehlerhafter Numerus oder ein Fehler durch das Truecasen.
- **Zeit**: die falsche Zeitform wurde verwendet
- **Lücken**: Wörter wurden weggelassen
- **Zusätzliche Wörter**: zusätzliche Wörter werden eingefügt. Am häufigsten ist der bestimmte Artikel ‚*the*' eingefügt worden.
- **Src-Ref-Differenz**: Differenzen im deutschen und englischen Satz, die durch sinngemässe Übersetzungen entstehen können.
- **Alinierungsfehler**: Fehler, welche während der Trennung der Sätze oder anschliessend bei der Alinierung entstanden sind (Abschnitt 2.1.3 & 2.1.4).
- **Worttokenizer**: Fehler, die durch fehlerhafte Worttokenizierung entstanden sind.

Die weiteren Analysen beschränkten sich auf die am häufigsten auftretenden Fehlerkategorien. Die häufigsten Fehlerkategorien sind all diejenigen, welche häufiger als die durchschnittliche Anzahl an Fehler aufgetreten sind. Der Durchschnitt lässt sich berechnen als $\emptyset = \frac{\#\text{ Fehler}}{\#\text{ Fehlerkategorien}} = \frac{338}{13} = 26$.

4.1 Synonym

Synonyme sind mit 60 Fehlern die häufigste Fehlerquelle. Der übersetzte Text bleibt zwar verständlich, aber man kann teilweise erkennen, dass es nicht von einem englischen Muttersprachler übersetzt wurde.

Folgendes Beispiel zeigt einen Ausschnitt aus einem solchen Satz:

Src: *[...] das vorhanden System [...]*

Ref: *[...] the available system [...]*

Hyp: *[...] the existing system [...]*

Abbildung 8: Toptreffer bei LEO Wörterbuch[10]

Abbildung 8 zeigt einen Ausschnitt der Topptreffer des online Wörterbuches LEO für die Übersetzung des deutschen Wortes ‚*vorhanden*' ins Englische. Man sieht dass ‚*available*' der Toptreffer ist, gefolgt von ‚*existing*'. Das SMÜ-System hat also nicht ein extrem seltenes Synonym verwendet, sondern eine durchaus plausible.

Ein zweiter Beispielssatz sieht folgendermassen aus:

Src: *Übernahme ohne Historie*

Ref: *apply excluding the history*

Hyp: *transfer without history*

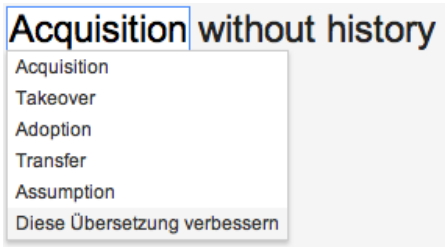


Abbildung 9: Übersetzung mit Google Translate[11]

Die Vergleichsübersetzung mit Google Translate (Abbildung 9) zeigt auch hier, dass die Wahl des SMÜ-Systems nicht ungewöhnlich ist. Der Kontext ist verständlich auch wenn der BLEU-Score[12] hier bei 0 liegen würde, da nur ⅓ der Übersetzung mit der Referenz übereinstimmt. Dieser Fehler wird auch durch eine eher sinngemässe Referenzübersetzung verstärkt.

Diese Fehlerkategorie ist sehr typisch für SMÜ-Systeme und wurde schon im Abschnitt „1.1 Motivation" mit dem Beispiel ‚*pass*' versus ‚*passport*'[13] angeschnitten. Der domänenspezifische Korpus kann dazu beitragen, fehlerhafte Synonyme zu verhindern, aber um alle Finessen zwischen Synonymen zu erkennen, braucht es einen Muttersprachler.

---

[10] Vollständigen Suchergebnis (Stand: 11.8.2014): http://dict.leo.org/#/search=vorhanden

[11] Übersetzung (Stand: 11.8.2014): https://translate.google.de/#de/en/Übernahme%20ohne%20Historie

[12] Der BLEU-1-Score mit nur Unigrammen wäre 25%, sobald mehr als nur Unigramme betrachtet werden, liegt er bei 0.

[13] Dies ist ein Beispiel für ein fehlerhaftes Synoym und gehört zur Kategorie „fehlerhafte Wortwahl".

Mit steigender Korpusgrösse kann man davon ausgehen, dass auch die Synonyme besser gewählt werden können. Hierfür wurden auch schon zu dem eigentlichen domänespezifischen Korpus insgesamt fünf weitere Korpora hinzugefügt, so dass der endgültige Korpus über 8’875'831 Zeilen verfügt.

Eine Optimierung für dieses ist bis heute nicht möglich.

4.2 Zusätzliche Wörter

Insgesamt wurden 46 Mal zusätzliche Wörter eingefügt. Es gibt grob gesagt drei Unterkategorien von zusätzlichen Wörtern.

Das am häufigsten zusätzlich eingefügte Wort ist der bestimmte Artikel ‚*the*’:

Src: *Dateiname der Kommando-Datei ohne Erweiterung (Reihenfolge : 1 .*

Ref: *file name of command file without extension (Sequence : 1 .*

Hyp: *file name of the command file without extension (sequence : 1.*

Der Kontext des Satzes und die Leserlichkeit wird dabei nicht vermindert. Beachten wir nur den ersten Teil des Satzes bis hin zur Klammer[14]. Der BLEU-Score beträgt für dieses Beispiel nur 6.25%. Mit anderen Worten würde das heissen, dass der Satz sehr schlecht und unverständlich ist. Sätze wie diese ziehen den BLEU-Score über das gesamte Test-Set herunter.

Einerseits könnte der Fehler von einer wörtlichen Übersetzung von ‚*der*’ stammen, oder ein Blick in die Phrasentabelle zeigt, dass der Eintrag ‚*Dateiname der Kommando-Datei*’ zu ‚*file name of the command file*’ vorhanden ist, genauso auch ein Eintrag zur Übersetzung mit ‚*file name of command file*’. Das heisst auch, dass eine Übersetzung mit dem bestimmten Artikel im Korpus vorhanden ist und daher das SMÜ-System diese Phrase als Ganzes übersetzt. Es ist also in diesem spezifischen Fall keine wörtliche Übersetzung, sondern es wurde die gesamte Phrase übersetzt.

Eine Verbesserung für diese Unterkategorie wäre daher nur möglich, wenn der Korpus überarbeitet und vereinheitlicht werden würde.

Die zweite Unterkategorie entsteht durch das „wörtliche“ Übersetzen aus dem Deutschen ins Englische. Es gibt Satzkonstruktionen die in beiden Sprachen verschieden sind, und deshalb zusätzlichen Wörter im Vergleich zum Referenzsatz vorhanden sind:

[14] Es wird nur bis zu Klammer betrachtet, da anschliessend ein Gross-/Kleinschreiben Fehler vorhanden ist, der nicht in die Berechnung mit einfliessen soll.

Src: *es werden alle Portfolios vom definierten Kunden geprüft .*

Ref: *all portfolios of the defined client are checked*

Hyp: *it all protfolios are of the definied client checked .*

Der Satz ist zwar noch knapp verständlich, besonders wenn man die deutsche Sprache beherrscht. Grammatikalisch ist er im Englischen dennoch falsch. Genau wie die zusätzlichen Artikel lässt sich diese Fehlerart nur schwer beheben mit der statistischen maschinellen Übersetzung. Regelbasierte maschinelle Übersetzungen wie beispielsweise Linguatec (LinguaTec Sprachtechnologien, 2014) hingegen haben hier einen Vorteil und übersetzten diesen Aspekt richtig:

Linguatec: *all portfolios are checked by the defined customer.*

Der Satz verfügt über eine Mehrdeutigkeit: die Daten vom Kunden bearbeiten, oder muss der Kunde diese bearbeiten. Linguatec verwendet die zweite Interpretation, der Referenzsatz die erste. Es geht bei dem Beispiel aber um den Umgang mit der Grammatikalischen Änderung des Satzes, welches von Linguatec korrekt umgesetzt wird.

Die dritte und letzte Unterkategorie entsteht durch fehlerhafte Alinierungen und führt dazu, dass wie im folgenden Beispiel ein unleserlicher Satz entsteht:

Src: *Anlagevorschlag Sparkonto aggregieren :*

Ref: *aggregate savings account in investment proposal .*

Hyp: *money (investment proposal pseudo security savings account aggregate :*

Die Elemente für eine korrekte Übersetzung sind vorhanden, aber es wurden zusätzlich vier weitere Token eingefügt. Ein Blick in die Phrasentabelle zeigt, dass die Übersetzung vom deutschen Wort ‚*Anlagevorschlag*' mit ‚*money (investment proposal pseudo security*' ins Englische übersetzt wurde.

Diese Fehler stammt aus der originalen Phrasentabelle des domänenspezifischen Korpus. Sucht man in dem Train-Set nach dem Vorkommen von ‚*Anlagevorschlag*' so findet man vier Einträge. Alle vier Sätze bestehen im Deutschen nur aus ‚*f. Anlagevorschlag)*' welches zweimal zu ‚*Call money (investment proposal pseudo security)*' und zweimal zu ‚*Trust term deposit*' aliniert wurde.

An diesem Beispiel sieht man sehr schön, dass die letzte Fehlerkategorie durch eine schlechte Satzalinierung entstanden ist. Um diese Fehler zu vermeiden, wäre es daher nötig den Alinierungsschritt (Abschnitt 2.1.3) zu verbessern.

4.3 Satzzeichen

Es wurden 33 Mal das falsche Satzzeichen in der Übersetzung im Vergleich zur Referenz festgestellt. Die meisten dieser Fehler fallen schon als Differenz zwischen dem deutschen Satz und dem englischen Referenzsatz auf.

Src: *Printing System , Design*

Ref: *printing system / Design*

Hyp: *Printing system , design*

Fehler wie in diesem Beispiel sind durch die Übersetzung gegeben und lassen sich durch das SMÜ-System nicht korrigieren.

Anders sieht es mit den Anführungszeichen aus, hier gibt es zwei Probleme, welche beide behoben werden können:

Src: *[...] & quot; Segmentdetail öffnen & quot; [...]*

Ref: *[...] " Open segment detail " [...]*

Hyp: *[...] & quot; open & quot; Segmentdetail [...]*

Im deutschen Satz wurden die Anführungszeichen durch benannte Zeichen (engl.: named entities) ersetzt. Im Englischen hingegen direkt geschrieben. Daher kommt es zu einer Differenz. Dies würde sich durch eine Regel bei der Korpus Aufarbeitung (Abschnitt 2.1) beheben lassen. Vor dem Training des SMÜ-Systems hilft es zusätzlich allen Alinierungsschritten als weiterer Ankerpunkt.

Zusätzlich zu den benannten Zeichen gibt es noch das Problem, dass im Englischen teilweise "" verwendet wird und teilweise ''. Auch hier wäre eine Verallgemeinerung eine einfache Möglichkeit, Fehler zu reduzieren.

Die Fehler mit den falschen Anführungszeichen sind für das Verständnis irrelevant. Es sind sozusagen Schönheitsfehler. Dennoch haben sie einen teilweise grossen Einfluss auf die Evaluation.

Es gibt selten auch bei den Satzzeichen „fehlerhafte Übersetzungen":

Src: *Laufwerk und Pfad (ohne Dateinamen) für das Erstellen der CVS-Daten-Datei .*

Ref: *drive and path (without file name) for the creation of the CSV data file .*

Hyp: *drive and path (without file name) for the creation of the CSV datafile -*

Wie auch schon im Abschnitt „4.2 Zusätzliche Wörter" liegt die Phrase ‚*CVS-Daten-Datei .*' nach ‚*CVS datafile-*' in der Phrasentabelle vor. Das heisst aber nicht, dass der

Punkt durch ein Bindestrich ersetzt worden ist, sondern viel mehr wurde ‚*Daten-Datei*' ersetzt durch ‚*datafile-*' und der Punkt weggelassen. Dieser Fehler stammt aus dem domänenspezifischen Korpus und lässt sich, ausser durch eine Überarbeitung dessen, nicht beheben.

4.4 Satzstellung

Fehlerhafte Neuanordnung des Satzes waren in 41 der 100 Sätze ein Problem. Die Übersetzung ist zwar meistens richtig, aber durch die fehlerhafte Satzstellung, ist das entziffern schwierig.

Folgender sehr kurzer Satz zeigt in aller Kürze den Unterschied in der Satzstellung zwischen Deutsch und Englisch. Eine „Wort-für-Wort"-Übersetzung ist daher oft falsch.

Src: *für Pseudovaloren mandatory*

Ref: *mandatory for pseudo-securities*

Hyp: *for pseudo securities mandatory*

Regelbasierte maschinelle Übersetzungssysteme haben bezüglich dieser Fehlerkategorie einen Vorteil:

Src: *die Aktivitäten können auch via " Start Job " gestartet werden .*

Ref: *the activitiy can also be started via , Start Job ' .*

Hyp: *the activities can also via , start job ' be started .*

Linguatec: *the activities can be started also via " start job " .*

Das RBMÜ-System hat als einzige Differenz die Position von ‚*also*', die restliche Anordnung stimmt mit dem Test-Set überein. Der übersetzte Satz hingegen hat ‚*be started*' und ‚*via Start Job*' vertauscht, womit der Satz nicht mehr korrekt ist. Um es genau zu nehmen, hat das SMÜ-System die Elemente nicht vertauscht, sondern wieder eine wörtliche Übersetzung vorgenommen.

Das Problem liegt also darin, dass der deutsche und englische Satz verschieden aufgebaut wird. Englische Sätze werden nach der „Subjekt-Verb-Objekt" (SVO) Regel gebaut. Deutsch Sätze hingegen nach der „Verbzweitstellung"-Regel. Hierbei steht das finite Verb an zweiter Position im Satz, wobei die Position davor frei wählbar ist. Weitere Unterschiede wie beispielsweise die Reihenfolge innerhalb eines Satzteiles ist im Deutschen „Zeit vor Ort"[15], im Englischen hingegen „Ort vor Zeit"[16].

[15] Deutsche Grammatik (Stand: 11.8.2014): http://de.wikipedia.org/wiki/Deutsche_Grammatik

[16] Englische Grammatik (Stand: 11.8.2014): http://de.wikipedia.org/wiki/Englische_Grammatik

Der Satz ‚*Gestern kaufte John Obst.*' wird ins Englische mit ‚*Yesterday, John bought apples.*' übersetzt. Die wörtliche Übersetzung ‚*Yesterday bought John apples.*' stimmt im Englischen wegen der SVO-Regel nicht.

Mit Hilfe der Wortartenbestimmung (engl.: Part-of-Speech Tagging) könnte überprüft werden, ob der englische Satz der SVO-Regel Genüge tut. Dafür müsste für jedes Wort dessen Wortart bestimmt werden. Für die Neuanordung müsste zuerst ein System auf einem korrekten und getaggten Text trainiert werden.

4.5 Komposita

Insgesamt gab es bei 24 Kompositafehler. Ein Viertel dieser entstehen durch fehlerhafte Zusammensetzungen. Der harmloseste Fehler ist, wenn das Verbindungszeichen fehlt oder falsch ist. Dadurch verliert die Übersetzung nichts an ihrer Verständlichkeit.

Src: *nein (Checkbox nicht selektiert)*

Ref: *no (checkbox not selected)*

Hyp: *no (check box not selected)*

In diesem Beispiel wurde ein Leerzeichen eingefügt, welches im Referenzsatz nicht vorhanden ist. Ein Blick in den domänenspezifischen Korpus zeigt, es gibt sowohl die Übersetzung mit als auch ohne Leerzeichen.

Fehlt aber das Verbindungswort und/oder ist die Wortanordnung falsch, kann es die Qualität beeinträchtigen:

Src: *Kontoart, Titelart und Risikoland für [...]*

Ref: *account type, securities type, and country **of** risk for [...]*

Hyp: *account type, security type and Risk country for [...]*

Schlägt man das Wort im online Wörterbuch LEO nach, wird dort nur die Übersetzung ‚*country of risk*' angezeigt. In der Phrasentabelle wird ‚*country of risk*' nur im Zusammenhang mit ‚*Risikostaat*' erwähnt.[17] Die Übersetzung von ‚*Risikoland*' ist aber grösstenTeils mit ‚*risk country*' eingetragen, selten mit ‚*nationalty*' oder ‚*land*' statt ‚*country*'.

Im Englischen wird die Bedeutung von ‚*risk country*' verstanden, würde aber von einem Muttersprachler nicht verwendet werden. Es ist gleichbedeutend, als würde jemand ‚*Land des Risikos*' sagen – wir verstehen es, aber es ist befremdlich.

[17] Es gibt noch zwei weitere Einträge, diese sind aber Fehlalinierungen zu ‚*als neue Auswahl*' und somit nicht relevant.

Bei drei Viertel der Fälle ist das vollständige Fehlen einer Übersetzung des Kompositum der Fehler. Dies liegt daran, dass im Deutschen die Kompositabildung sehr oft vorgenommen wird. Um ein Kompositum übersetzen zu können, muss es im Korpus mindestens einmal vorkommen. Dies ist aber wegen der immensen Anzahl an möglichen Kombinationen nicht erreichbar.

Es existiert eine empirische Methode zur Lösung dieses Problems (Koehn & Knight, 2003). Als Erstes muss ein System mit möglichen Einzelwörtern für die Erkennung der Komposita trainiert werden. Anschliessend können die Komposita in ihre Einzelwörter aufgeteilt werden. Das Kompositum ‚*Aktionsplan*' kann zerteilt werden in ‚*Aktions•plan*', ‚*Aktion•s•plan*' und ‚*Akt•ion•s•plan*'.

Um zu verhindern das Wörter wie ‚*folgenden*' in ‚*folgen•den*' geteilt werden, wird zusätzlich die Wortarten bestimmt und nur Trennungen zu Nomen, Adjektiven, Adverbien und Verben erlaubt.

Für ein Kompositum kann mit Hilfe der Häufigkeit der Komponenten p_i der Aufteilung S bestimmt werden, welche rein statistisch gesehen am wahrscheinlichsten ist. Wobei n die Anzahl von Komponenten. Das gesamte Kompositum ist genauso eine Aufteilung in $n = 1$ Stück.

$$\text{argmax}_S\left(\left(\prod_{p_i \in S} \text{count}(p_i)\right)^{\frac{1}{n}}\right) \tag{7}$$

Anschliessend wird mit Hilfe eines Übersetzungslexikons überprüft, ob für die Aufteilungen S eine Übereinstimmung in der englischen Übersetzung vorhanden ist. Unter zur Hilfenahme dieses Schrittes kann die Wahrscheinlichkeit berechnet werden, welche Aufteilung S am wahrscheinlichsten ist.

Anschliessend wird ein neues Übersetzungslexikon erstellt, in welchem die Komponenten aufgeteilt sind und die Englischen bleiben gleich. Das dadurch entstandene Übersetzungslexikon enthält die möglichen Übersetzungen.

5 Schlusswort

Für diese Arbeit habe ich ein domänenspezifisches SMÜ-System trainiert. Dafür habe ich von der Firma Finnova einen domänenspezifischen Korpus erhalten. Dieser musste in einem ersten Schritt von den einzelnen XML-Dateien zu einem parallelen Korpus aufgearbeitet werden. Hierfür wurden zuerst die Dokumente anhand ihrer Dateinamen aliniert, so dass man die Verknüpfung zwischen englischer und deutscher Datei vorliegt. Danach konnte der Text extrahiert werden, um anschliessend sowohl auf Satz- als auch auf Wortebene tokenisiert zu werden. Als letzter Schritt wurde mit dem Programm HunAlign die Satzalinierung vorgenommen. Die Sätze lagen nun in einem parallelen Korpus vor.

Mit diesem parallelen Korpus konnte ich anschliessend mit Hilfe von Moses ein domänespezifisches SMÜ-System trainieren. Hierfür wurde als Erstes der Korpus in drei Sets aufgeteilt (Test-, Train- und Dev-Set). Anschliessend wurden alle Wörter in ihre natürliche Gross-/Kleinschreibung gebracht. Da es für die Programme sehr wichtig ist, dass keine Leerzeilen vorhanden sind und zu lange Sätze nicht sinnvoll sind, wurde hier eine Bereinigung des Korpus vorgenommen. Anschliessend wurde das Sprachmodell mit bis zu 5-Grammen trainiert. Nun lagen alle Dokumente für das effektive Training des SMÜ-Systems vor und es konnte trainiert werden. Um das SMÜ-System zu verbessern, wurde anschliessend das Minimum Error Rate Training dreimal durchgeführt.

Es wurde eine erste Evaluation durchgeführt, um die besten Konfigurationen herauszufinden und um einen Vergleichswert für später zu erhalten. Dabei wurde festgehalten, dass der Konfigurationsparameter `inclusive` für die Verarbeitung der XML-Tags das beste Ergebnis erzielt.

Eine weitere Verbesserung des SMÜ-Systems wurde mit zusätzlichen domänenfremden Korpora erzieht. Hierfür wurden zwei weitere von der Firma und drei öffentlich zugängliche (Europarl, OpenSubs und JRC-Acquis) verwendet.

Die Korpora lagen parallel vor, daher waren hier keine Aufarbeitungsschritte notwendig. Das Training lief bis zum Trainingsschritt gleich ab wie beim domänenspezifischen Korpus. In diesem musste das Programm nur bis zum Schritt „Phrasen bewerten“ angewendet werden, da der Rest für das Kombinieren nicht relevant ist.

Die gewonnenen Phrasentabellen wurden mit der domänenspezifischen Phrasentabelle verrechnet: einmal mit Interpolation das anderemal mit Gewichtung.

Anschliessend wurde wieder eine Evaluation durchgeführt. Hierbei wurde festgestellt, dass eine Verbesserung durch die Kombination erzielt werden kann, und dass die Verrechnung mit der Gewichtung zu einer stärkeren Verbesserung führt.

Für die Analyse der häufigsten Fehler wurden zufällig 100 Sätze aus dem Test-Set extrahiert und von Hand verglichen. Dabei stellte ich insgesamte 13 Fehlerkategorien fest. Fünf davon hatten mehr als die durchschnittliche Fehleranzahl und wurden genauer analysiert.

Am häufigsten treten Synonym Fehler auf. Diese beeinträchtigen das Verständnis nicht, können aber je nach Evaluationsmetrik einen stark negativen Einfluss haben.

Im Kontrast dazu steht die falsche Satzstellung, diese beeinflusst das Verständnis der Übersetzung sehr stark. Ein möglicher Ansatz für eine Verbesserung wäre die Verwendung der Wortartenbestimmung. Anhand dieser könnte überprüft werden, ob der englische Satz den grammatikalischen Regeln wie beispielsweise der „Subjekt-Verb-Objekt“-Regel genügt.

Die dritte Fehlerkategorie sind zusätzlich eingefügte Wörter. Diese besteht aus drei Unterkategorien: die erste ist das Einfügen von zusätzlichen bestimmten Artikeln. Dies kann starke Auswirkungen auf die Evaluationsmetrik haben, aber der englische Satz bleibt korrekt. Die zweite Unterkategorie entsteht durch die unterschiedlichen Satzaufbauten. Diese Fehler können von regelbasierten MÜ-System erkannt werden. Um diese zu verbessern, müssten Grammatik Regeln eingebunden werden. Die letzte Unterkategorie entsteht durch fehlerhafte Alinierungen während des Trainingschrittes. Die Behebung liegt in der Verbesserung des Alinierungschrittes, welcher auf der Tokenisierung aufbaut. Besonders der Satztokenisierer könnte besser an den domänenspezifischen Korpus angepasst werden.

Die Fehler im Umgang mit Satzzeichen haben so gut wie keinen Einfluss auf das Verständnis des Textes, können aber auch wieder einen Einfluss auf die Evaluation haben. Eine Vereinheitlichung von Anführungszeichen würden schon ein Viertel dieser Fehler beheben. Ein Grossteil der restlichen Fehler liesse sich durch eine Überarbeitung des Korpus beheben, da es Diskrepanzen zwischen Quell- und Referenzsatz hat.

Der letzte Fehler entsteht besonders durch die Vorliebe in der deutschen Sprache, Wörter zusammen zu setzen und Komposita zu bilden. Gleich wie bei der Satzstellung unterscheiden sich hier die deutsche und englische Sprache stark. Im Englischen wird dieses Vorgehen nicht verwendet, es ist zwar verständlich, aber es ist kein korrektes Englisch. Um dieses Problem zu lösen, gibt es schon Ansätze zur Teilung von Komposita in ihre Komponenten.

Als nächster Schritt sollte sicher der schon vorhandene Ansatz für die Komposita Fehler angewendet werden. Die Verbesserung des Satztokenizer kann an vielen Orten zu einer Verbesserung führen. Da der Tokenizer auf Basis von Zeitungen trainiert wurde, hat er mit Abkürzungen und unvollständigen Sätzen Schwierigkeiten. Ihn mit einem (oder mehreren) anderen Korpora zu trainieren, könnte Verbesserungen hervorrufen. Da die Satzstellung bei 36% der Sätze fehlerhaft war, sollte dieser Punkt nicht vernachlässigt werden.

Anhang A: Moses Befehle

```
01 $ ./moses/scripts/recaser/train-truecaser.perl --model \
   ./temp/truecase-model.de --corpus ./finnova/train.de

02 $ ./moses/scripts/recaser/train-truecaser.perl --model \
   ./temp/truecase-model.en --corpus ./finnova/train.en

03 $ ./moses/scripts/recaser/truecase.perl --model \
   ./temp/truecase-model.de < ./finnova/train.de > \
   ./finnova/train.truecase.de

04 $ ./moses/scripts/recaser/truecase.perl --model \
   ./temp/truecase-model.en < ./finnova/train.en > \
   ./finnova/train.truecase.en

05 $ ./moses/scripts/recaser/truecase.perl --model \
   ./temp/truecase-model.de < ./finnova/dev.de > \
   ./finnova/dev.truecase.de

06 $ ./moses/scripts/recaser/truecase.perl --model \
   ./temp/truecase-model.en < ./finnova/dev.en > \
   ./finnova/dev.truecase.en

07 $ ./moses/scripts/training/clean-corpus-n.perl \
   ./finnova/train.truecase de en ./finnova/train.clean 1 80

08 $ ./irstlm/bin/add-start-end.sh < ./finnova/train.clean.de \
   > ./finnova/train.sb.de

09 $ export IRSTLM=./irstlm; ./irstlm/bin/build-lm.sh \
   -i ./finnova/train.sb.de -t ./tmp -p -s improved-kneser-ney -n 5 \
   -o ./finnova/train.lm.de -k 10

10 $ ./irstlm/bin/compile-lm --text=true ./finnova/train.lm.de.gz \
   ./finnova/train.arpa.de

11 $ ./moses/bin/build_binary -i ./finnova/train.arpa.de \
   ./finnova/train.blm.de

12 $ ./moses/scripts/training/train-model.perl -root-dir . \
   -corpus ./finnova/train_clean -f de -e en \
   -alignment grow-diag-final-and -reordering msd-bidirectional-fe \
   -lm 0:5:$HOME/finnova/train_blm.de:8 --parallel -cores 5 \
   -external-bin-dir ./mgizapp/bin/ -mgiza -mgiza-cpus 5 \
   --write-lexical-counts

13 $ ./moses/scripts/training/mert-moses.pl ./finnova/dev.truecase.de \
   ./finnova/dev.truecase.en ./moses/bin/moses ./model/moses.ini \
   --mertdir ./bin/ --decoder-flags="-threads 10"
```

Anhang B: Fehleranalyse

Legende			
TIM	= time (Zeit)	UNK	= unknown word (unbekanntes Wort)
PUN	= punctuation mark (Satzzeichen)	SYN	= synonym (Synonym)
WRO	= wrong word (fehlerhaftes Wort)	COM	= compounds (Komposita)
EXW	= extra word (zusätzliches Wort)	ORD	= sentence order (Satzstellung)
EXT	= extra ‚*the*' (zusätzlicher Artikel)	GAP	= gaps (Lücken)
TRU	= truecase (Gross-/Kleinschreibung)	DIF	= Src-Ref-Difference
NUM	= numerus (Numerus)	ALI	= Alingnment error
		TOK	= tokenization error

Satz 3266: exportiert Output (manuelle Änderungen) in ein csv-File
Ref: exports output (manual changes) to a csv file .
Hyp: [exported output (manual changes) [in]WRO a csv file []PUN]TIM

Satz 2391: Laufnummer des Composite Stamms
Ref: sequence number of Composite master
Hyp: sequence number of [the]EXT [composite]TRU [Stamms]UNK

Satz 529: dieser Entscheidungsbaum bestimmt die Frist anhand von :
Ref: this decision tree determines deadline by means of the following criteria :
Hyp: this decision tree determines [the period with of :]DIF

Satz 245: das Kriterienpanel wird nicht angezeigt , somit können keine Kriterien erfasst werden
Ref: the criteria panel is not displayed ; therefore no criteria can be entered .
Hyp: the [Kriterienpanel]COM is not displayed [,]PUN [can therefore no criteria be [recorded]SYN []PUN]ORD

Satz 3413: Umlage der Erfolgskennzahlen auf die Produkte und die Kostenverursacher .
Ref: allocation of P & amp ; L key figures to products and cost driver
Hyp: allocation of [the profit and loss reference numbers on]DIFF [the]EXT products and [the]EXT cost [causer]SYN .

Satz 3916: idealerweise eine existierende dem Empfänger unterschiedliche Adresse .
Ref: we recommend defining an existing address that is different form the recipient address .
Hyp: []ALI should ideally do an existing the recipient different address .

Satz 2853: nein (Checkbox nicht selektiert)
Ref: no (checkbox not selected)
Hyp: no ([check box]COM not selected)

Satz 3720: Dateiname der Kommando-Datei ohne Erweiterung (Reihenfolge : 1 .
Ref: file name of command file without extension (Sequence : 1 .
Hyp: file name of [the]EXT command file without extension ([sequence]TRU : 1 .

Satz 2462: @ Uebrige Hedge Funds
Ref: @ Remaining hedge funds
Hyp: @ [other]SYN hedge [fund]NUM

Satz 619: sämtliche Funktionen können / müssen über den Formschutz freigegeben oder gesperrt werden .
Ref: you can / need to clear or block all functions via form protection :
Hyp: [all functions can / [must]SYN via [the]EXT form protection [released]SYN or be blocked [.]PUN]ORD

Satz 1366: dieser Parameter steuert die Übernahme des BEG_DAT aus der Zuordnung des Kundenportfolio (PMS +) .
Ref: this parameter controls the takeover of the BEG_DAT from the allocation of the client portfolio (PMS +) .
Hyp: this parameter controls the [transfer]SYN of the BEG_DAT from the allocation of the [Client]TRU portfolio (PMS +) .

Satz 3638: GR00 – Grundstück Daten
Ref: GR00 - Land register data
Hyp: GR00 – [Property]SYN data

Satz 633: Mahnungsstatus anpassen = = & gt ; Mit dieser Funktion können Massenupdates (Multiselect) auf dem Mahnungsstatus gemacht werden
Ref: change reminder status==> This function allows you to make a (multiselect) bulk update of the reminder status .
Hyp: [[[Mahnungsstatus]COM [necessary]WRO [= = & gt ;]TOK With this function can [Massenupdates]COM (,) on the [Mahnungsstatus]COM be made]TIM]ORD

Satz 2485: Darstellung der Hierarchie
Ref: representation of the
Hyp: representation of the [hierarchy]ALI

Satz 1012: der JSC-Job " ZVE Abstimmung finnova und CS " erstellt die Abstimmung zwischen Finnova und dem CS für Eingänge .
Ref: the JCS job ' ZVE Abstimmung finnova und CS ' creates the coordination between Finnova and the CS for incoming payments .
Hyp: the JCS job [']PUN ZVE Abstimmung finnova and CS ' [created]TIM the coordination between Finnova and the CS for incoming payments .

Satz 3187: Laufwerk und Pfad (ohne Dateiname) für das Erstellen der CVS-Daten-Datei .
Ref: drive and path (without file name) for the creation of the CSV data file .
Hyp: drive and path (without file name) for the creation of the CVS [datafile-]COM []PUN

Satz 923: die Aktivität muss in Abstimmung auf das vorhandene System (Anzahl laufenden Execution Instanzen) optimal parametriert werden .
Ref: the activity needs to be optimally parameterised in conformity with the available system (number of running execution instances) .
Hyp: [the activity [must]SYN in [coordination]SYN to the [existing]SYN system (number of [current]SYN execution instances) way be parameterised .]ORD

Satz 3201: wenn bei der Kundeneröffnung kein Berater zugeordnet wird , dann kann mit diesem PPAR definiert werden , wie sich das System bei der Speicherung des Kunden verhalten soll .
Ref: if no advisor is assigned with the initiation of a customer , this PPAR may be used to define how the system will behave when saving the customer ..
Hyp: [if in the [client]SYN [opening]SYN no advisor is assigned to , then can with this [program parameter]SYN be defined [,]PUN how the system with the saving of the [client]SYN [behaviour]SYN should [.]PUN]ORD

Satz 665: Ablösung des letzten Verarbeitungsschritts in JCS-Job zv0017 ZVE Process pending CS-files (Kapitel 2.2)
Ref: replacement of last processing step in JCS job zv0017 ZVE Process pending CS files (Chapter 2.2)
Hyp: replacement of [of]EXW [the]EXT last [Verarbeitungsschritts]COM in JCS job zv0017 ZVE Process pending [CS-files]COM ([chapter]TRU 2.2)

Satz 44: es werden alle Portfolios vom definierten Kunden geprüft .
Ref: all portfolios of the defined client are checked
Hyp: [[it]EXW all portfolios are of the defined client checked [.]PUN]ORD

Satz 3388: Steuercodes gemäss = & gt ; 3.5.1
Ref: controls corresponding with 3.5.1
Hyp: [control codes [according]SYN to = & gt ; 3.5.1]DIF

Satz 2347: Konfiguration DenoTask Renderer AL1WRG
Ref: configuration of DenoTask Renderer AL1WRG
Hyp: configuration []GAP DenoTask [renderer]TRU AL1WRG

Satz 1797: (für Pseudovaloren mandatory)
Ref: (mandatory for pseudo-securities)
Hyp: [(for pseudo [securities mandatory]COM)]ORD

Satz 1877: für alle Titelarten , die mit Pseudovaloren verwendet werden , muss in der Steuercodegruppe „ StCd “ die Option = H (VL_REF bilden) den Wert = E (Bei Bank-VL-Stamm-Eröffnung) haben .
Ref: all securities types that are used with pseudo-securities In the control code group ‘ StCd ’ , option = H (form VL_REF) must have the value = E (at bank securities master opening) for all security types that are used with pseudo-securities .
Hyp: [[[for]EXW all securities types [,]PUN the with [pseudo securities]COM be used [,]PUN must in the [Steuercodegruppe]COM ‘ StCd ’[]PUN [the]EXT option = H (VL_REF [groups]WRO) the value = E ([For]WRO [Bank-VL-Stamm-Eröffnung]COM) .]ORD]TIM

Satz 3746: Laufwerk und Pfad für das Erstellen der XML-Daten-Datei . am Ende muss zwingend ein & quot ; \ & quot ; stehen !
Ref: drive and path where the XML data file is to be created . make sure there is always a ‘ \ ’ at the end !
Hyp: [drive and path [for the creation]SYN of the XML data file . [at the end must be]SYN a [& quot ;]PUN \ [& quot ;]PUN [are]EXW !]ORD

Satz 1953: wird im GUI aus manuell eingefügt oder aus dem Portfolio gelesen
Ref: manually inserted in the GUI , or read from the portfolio
Hyp: [[is in the GUI [from]EXW manually inserted []PUN or from the portfolio read]ORD]TIM

Satz 1868: AL1ZEIT_EINH Steuercode zulassen von Zeiteinheiten)
Ref: AL1ZEIT_EINH control code : allowable time units)
Hyp: AL1ZEIT_EINH control code []PUN [allow]SYN of time units)

Satz 2773: voll qualifizierter Queuename des eingangs .
Ref: fully qualified IN queue name
Hyp: [full]SYN qualified [Queuename]COM [of the already]WRO [.]PUN

Satz 3965: Aufbereitung der Daten für die SAP-Abacus-Schnittstelle
Ref: date preparation for the SAP Abacus interface
Hyp: [preparation of the [data]WRO for the [SAP-Abacus-Schnittstelle]COM]ORD

Satz 1561: der Parameter dient der Datumsbestimmung ab wann die Kurse neu geladen werden sollen .
Ref: the parameter is used to determine the date from when the rates are to be reloaded .
Hyp: the parameter [[serves]WRO]TIM of [the]EXT [Datumsbestimmung]COM from when the rates [new]EXW [should]SYN be loaded .

Satz 2796: nein (Checkbox nicht selektiert)
Ref: no (checkbox not selected)
Hyp: no ([check box]COM not selected)

Satz 3219: Parameter unter Programm-Id : „ KT _ GRP _ BERAT “
Ref: parameters under programme ID ‘ KT _ GRP _ BERAT ’
Hyp: [parameter]NUM under [program]DIF ID [:]PUN ‘ KT _ GRP _ BERAT ’

Satz 1747: Rendite bis Verfall
Ref: return to maturity
Hyp: return [until]SYN [expiration]SYN

Satz 2385: DEBUG_YIELD : enthält die ermittelten Renditen / Basisdaten
Ref: DEBUG_YIELD : contains the determined returns / basic data
Hyp: DEBUG_YIELD : contains the determined returns / basic data

Satz 3375: -- Portfolios ohne zugewiesene Beratger --
Ref: -- Portfolios without assigned
Hyp: -- [portfolios]TRU without assigned [Beratger --]DIF

Satz 3146: der Parameterwert entspricht einer in KD1KDKRK erfassten Rolle .
Ref: parameter value corresponds with a role recorded in KD1KDKRK .
Hyp: [[the]EXT parameter value corresponds [to]WRO a recorded in KD1KDKRK role .]ORD

Satz 439: mit diesem PPAR wird gesteuert ob in der Lasche Bilanzeinreichung im Kundenstamm (PPAR KD_STAMM / BILANZEINR muss auf 2 sein !) das Panel mit der Dokumentzuordnung ebenfalls angezeigt wird , oder nicht .
Ref: this parameter controls whether the balance sheet submission tab in the client master window (KD_STAMM/BILANZEINR must be = 2) also includes a document allocation panel .
Hyp: [with this program parameter is controlled whether in the Balance sheet submission tab in the client master (PPAR KD_STAMM / BILANZEINR must be set to 2 !) the panel with the document also is displayed , or not .]DIF

Satz 3562: elektronische Destination des Kunden (Freigeber Schätzung)
Ref: electronic destination of the client (approver assessment)
Hyp: electronic destination of the client ([Freigeber]UNK appraisal)

Satz 441: & LT ; leer & gt ; = = & gt ; Lasche Bilanzeinreichung wird in der MAsicht nicht angezeigt . (Default)
Ref: 0 / & LT ; blank & gt ; = Do not display ‘ Balance sheet submission ’ tab in the staff view (default)
Hyp: []DIFF & LT ; blank & gt ; = [= & gt ;]DIF [[[]PUN Balance sheet submission []PUN tab is in the staff view (default) not displayed [.]PUN (default)]ORD]TIM

Satz 1032: die Aktivitäten können auch via “ Start Job “ gestartet werden .
Ref: the activity can also be started via ‘ Start Job ’ .
Hyp: [the [activities]DIFF can also via ’ start job ’ be started .]ORD

Satz 2232: Übernahme ohne Historie
Ref: apply excluding the history
Hyp: [transfer]SYN [without]SYN history

Satz 2694: Parameter , von welchen das Feld „ Obligatorisch “

Ref: ‘ Mandatory ’ field is
Hyp: [parameter , of which the field ‘ mandatory ’]DIF

Satz 3932: Bilanzwährung für Export
Ref: balance sheet currency for export
Hyp: balance []GAP currency for export

Satz 342: der Parameter ANZ_TAGE definiert die Anzahl Tage , für welche die Daten in der Tabelle OM_CHK_SALDO erhalten bleiben .
Ref: parameter ANZ_TAGE defines how many days to retain data in table OM_CHK_SALDO .
Hyp: [[the]EXT parameter ANZ_TAGE defines [the number of days , for which the data]SYN in the table OM_CHK_SALDO receive remain .]ORD

Satz 1292: Kontoart , Titelart und Risikoland für die Restliquidität zum Anlagevorschlagsparameter 82 aus der Tabelle PF1REBAL_PAR
Ref: account type , securities type , and country of risk for excess liquidity with Investment Proposal parameter 82 from table PF1REBAL_PAR
Hyp: account type , [security type]COM []PUN and [Risk country]COM for [the]EXT [remaining]SYN liquidity [to]WRO [the]EXT [Anlagevorschlagsparameter]COM 82 from [the]EXT table PF1REBAL_PAR

Satz 1234: dieser Parameter steuert die Nachkommastellen fuer die Performanceattribution .
Ref: this parameter controls the number of decimal places for performance attribution .
Hyp: this parameter controls the [decimal places]SYN for [the]EXT [Performanceattribution]COM .

Satz 3711: - & LT ; kein & gt ; oder 0 = Output über Printing System
Ref: - & LT ; none & gt ; or 0 = Output via Printing System
Hyp: - & LT ; none & gt ; or 0 = Output via [printing]TRU [system]TRU

Satz 2463: @ Uebrige Strukt . Produkte
Ref: @ Remaining real estate
Hyp: @ [other]SYN [structured products]DIF

Satz 1674: Beginn der Modellzuordnung . falls das Datum leer (NULL) gelassen wird , erfolgt die
Ref: model allocation start . if you leave this field empty (NULL) ,
Hyp: [start of the Modellzuordnung . if the date blank (NULL) is left , the]ALI

Satz 2110: Bewertung / Rendite auf Ebene Asset Allokation (seit Portfoliostart) Nachrechnen , wenn Positionen aufgrund von Risikoattributmutationen auf ein anderes Segment verschoben wurden .
Ref: calculates evaluation / returns on asset allocation level (since portfolio opening) if positions have been moved to a different segment on the basis of modifications to the risk attributes .
Hyp: [[valuation]SYN / [return]NUM on level asset allocation (since [Portfoliostart]COM) [through ‘ Recalculate]EXW [,]PUN if positions basis of [Risikoattributmutationen]COM to [another]SYN segment moved have been .]ORD

Satz 1652: das Datum muss & gt ; = EROEFF_DAT sein .
Ref: the date must be & gt ; = EROEFF_DAT .
Hyp: [the date must & gt ; = be EROEFF_DAT .]ORD

Satz 3232: Beispiel : / * + index (b KT_GRP_BERATI5) * /
Ref: example : / * + index (b KT_GRP_BERATI5) * /
Hyp: example : / * + index (b KT_GRP_BERATI5) * /

Satz 258: Datensatz ist ersichtlich
Ref: data record suppressed
Hyp: [Suppress is can]DIF

Satz 1120: der Vorschlagswert der beim Erfassen des Anlagevorschlages angezeigt wird ist im Program Parameter PFREBAL / DEF_VL_STUECK abgelegt .
Ref: the default value displayed during entry of the investment proposal is stored in the program parameter PFREBAL / DEF_VL_STUECK . the suggested values provided are set to " Datensatz unterdrücken "
Hyp: [the default value of the []GAP [when]SYN [entering]WRO at is displayed is in []GAP program parameter PFREBAL / DEF_VL_STUECK stored .]TIM []ALI

Satz 982: die dritte Stelle des PP OMSZV.ABLOES_JCS_JOBS.13 ist auf den Wert 2 zu setzen .
Ref: you have to set the third position of the program parameter OMSZV.ABLOES_JCS_JOBS.13 to value 2 .
Hyp: [[[]GAP the third position of the program parameter OMSZV.ABLOES_JCS_JOBS.13 [that is on]WRO [the]EXT value 2 to be set .]ORD]TIM

Satz 2461: @ Uebrige Strukt . Produkte
Ref: @ Remaining real estate
Hyp: @ [other]SYN [structured products]DIF

Satz 2660: Finnova FIX Schnittstelle
Ref: the two bundles (plug-ins) required by the Finnova FIX interface are now installed on the Finnova OSKi server and are immediately available .
Hyp: [Finnova FIX interface]ALI

Satz 3864: mit diesem PrgPar wird im IB gesteuert , ob bei der Erfassung von einem Dauerauftrag vom Typ „ Kontoübertrag " auf ein Vorsorgekonto die automatische Betragsmutation zugelassen ist oder nicht .
Ref: this program parameter controls in IB whether automatic amount modifications are permissible when recording a standing order of the type ' account transfer ' to a pension account .
Hyp: [[[with]EXW this program parameter is in [the]EXT IB controlled [,]PUN whether [for]WRO [the]EXT recording [of]EXW a standing order of the type ' account transfer ' [on]WRO a pension provision account [the]EXT automatic Betragsmutation is allowed [or not]DIF .]TIM]ORD

Satz 1217: mit diesem Parameter , wird für jede Aenderung (Ueberschreibung) eines aktiven Stichtags ein Info Log Eintrag mit den Daten des zu ändernden Stichtags geschrieben .
Ref: this parameter defines whether an info log entry about the date to be changed is created whenever an active reference date is changed (overwritten) .
Hyp: [[[with]EXW this parameter [,]PUN is for each change ([Ueberschreibung]UNK) of a active reference date an information log entry with the data of the steps to reference date written .]ORD]DIF

Satz 3760: in diesem Dokument werden nur die für die Einführung des Dauerauftrags notwendigen Schritte und Anforderungen beschrieben .
Ref: this document does only include those steps and requirements that are related to the implementation of standing orders .
Hyp: [[[in]EXW this document only the for the [introduction]SYN of [the]EXT standing [order]NUM necessary steps and them according described .]ORD]DIF

Satz 1535: Anlagevorschlag Sparkonto aggregieren :
Ref: aggregate savings account in investment proposal .
Hyp: [money (investment proposal pseudo security]WRO savings account aggregate [:]PUN

Satz 424: dieser muss vorgängig parametriert werden .
Ref: the submission deadline for the balance sheet is calculated by means of the decision tree KDBE , which has to be parameterised in advance .
Hyp: [[this]WRO [must]SYN be parameterised in advance .]ORD

Satz 3660: Strasse des Grundstückes
Ref: street of the property
Hyp: street of the property

Satz 678: Ablösung JCS-Jobs in ZV
Ref: replacement of JCS jobs in Payment Transactions (ZV)
Hyp: replacement of JCS jobs in [ZV]DIF

Satz 470: 2 = = & gt ; neue Bilanzeinreichung aktiv .
Ref: 2 = Enable new balance sheet submission .
Hyp: [2 = [= & gt ;]DIF new balance sheet submission [active]SYN .]ORD

Satz 4: der zweite Teil definiert die Datenstrukturen , die bei der Abarbeitung der konfigurierten Restriktionen – dem Prüfprozess – verwendet werden .
Ref: data structures for configuring restrictions
Hyp: [the second part defines the data structures , the for the processing the configured restrictions – the Prüfprozess – be used .]ALI

Satz 2027: 11.6.1.2 PRG_PAR DELDEP_FUTEVENTS = 0
Ref: 11.4.1.2 PRG_PAR DELDEP_FUTEVENTS = 0
Hyp: 11.6.1.2 [Prg_Par]TRU DELDEP_FUTEVENTS = 0

Satz 2272: unabhängig ob die Risikokennzahlen im Batch oder Online angefordert werden , die Berechnung erfolgt immer im Hintergrund .
Ref: regardless of the fact whether risk key figures are requested in batch or online , the calculation is always done in the background .
Hyp: [[[irrespective]SYN of []GAP whether [the]EXT risk key figures have in [the]EXT batch or online be requested , the calculation is always in the background .]TIM]ORD

Satz 948: die Aktivität kann auch via " Start Job " gestartet werden .
Ref: the activity can also be started via ' Start Job ' .
Hyp: [the activity can also via ']PUN start job ' be started .]ORD

Satz 3057: verfügbar unter Menupunkt : Extras – Gruppierungen / Entscheidungen – (Allg.Gruppe – Gruppe Zuordnung)
Ref: navigation : Extras – Grouping / Decisions – (Gen. Group – Group assignment)
Hyp: navigation : Extras – Groupings / Decisions – ([Allg.Gruppe]TOK – [group]TRU [allocation]SYN)

Satz 1622: (KT_GRP.eroeff_dat)
Ref: portfolio opening date . (
Hyp: [(KT_GRP.eroeff_dat)]ALI

Satz 1722: es können maximal 999 Lieferanten erfasst werden .
Ref: up to a maximum of 999 providers can be entered .
Hyp: [[]GAP can maximum []GAP 999 [the]EXT [issuer]WRO be [recorded]SYN .]ORD

Satz 778: bei der alten Verarbeitung via JCS-Job bs.zv017 war dies nicht nötig , weil dort einfach der Job restartet werden konnte und die offenen Aufträge verarbeitet wurden .
Ref: this was not required with the former processing via JCS job bs.zv017 , because at that time , the job could simply be restarted and the open orders were processed .
Hyp: [for the [old]SYN processing via JCS job bs.zv017 this was not [necessary]SYN , because [there]WRO simply []PUN the job be restarted could and the open orders [been processed]TIM .]ORD

Satz 1779: Anzahl der Coupon-Perioden pro Jahr
Ref: number of coupon periods per year
Hyp: number of coupon periods per year

Satz 1336: , AN ' – Anlageberater
Ref: ' AN ' – investment advisor
Hyp: ' AN ' – Investment advisor

Satz 300: Check beim Speichern einer Restriktion
Ref: check when saving a restriction
Hyp: check when saving a restriction

Satz 787: die Entfernung des letzten Schrittes aus dem JCS-Job kann auch erst vorgenommen werden , wenn die Umstellung bereits erprobt ist .
Ref: the removal of the last step from the JCS job can also only be carried out if the conversion is already tested .
Hyp: the [Removal]TRU of the last step from the JCS job can also only be [made]SYN [,]PUN [if the conversion already [tried]SYN is]ORD .

Satz 2805: bestimmt , ob das DataDictionary für die Validierung (Gültigkeit) von Meldungen verwendet werden soll .
Ref: determines whether DataDictionary is to be used to validate messages .
Hyp: [determines whether the DataDictionary [for]WRO [the]EXT [validation (validity) of messages]DIF is to be used .]ORD

Satz 1043: die Aktivitäten ZVSTA_E * sind so anzupassen , dass die Periodizität der Ausführung jener des JCS-Jobs entspricht .
Ref: you have to adjust activity ZVSTA_E * in such a way that the periodicity of the execution corresponds to that of the JCS job .
Hyp: [the [activities]DIF ZVSTA_E * are to []GAP [,]PUN that the periodicity of the execution [the one]SYN of the JCS job corresponds to .]ORD

Satz 1505: 1 = Baum wird geschlossen dargestellt d.h. es sind nur die Knoten auf Portfolioebene sichtbar .
Ref: 1 = Tree is displayed closed ; i.e. only nodes on portfolio level are visible .
Hyp: [1 = [tree]TRU is closed displayed []PUN i.e. [it]EXW are only [the]EXT [knot]SYN on portfolio level visible .]ORD

Satz 1908: Druck wird definitiv auslösen
Ref: definite printing triggered
Hyp: [printing is definitely trigger]ORD

Satz 173: PMS-Investitionsprofil - SAA
Ref: PMS Investment profile
Hyp: [PMS-Investitionsprofil]COM [- SAA]DIF

Satz 3251: 1 = Neues Beraterkonzept mit KDQUTIL1.F_GetKDBerater aufrufen
Ref: 1 = Call new advisor concept with KDQUTIL1.F_GetKDBerater
Hyp: 1 = [Calling]TIM new advisor concept with KDQUTIL1.F_GetKDBerater

Satz 2129: selektiert die zu aktivierenden Benchmark Events . (die vorgängig für die Zukunft erfasst wurden)
Ref: selection of the benchmark events to be activated (which were entered in advance for the future)
Hyp:][selected the to [aktivierenden]UNK benchmark events [.]PUN ([the]WRO in advance for the future have been entered)]ORD]TIM

Satz 2640: Installation des FIX Engine Bundles
Ref: the FIX engine dequeuer retrieves the messages and sends them to the Stock Exchange application .
Hyp: [installation of the FIX engine bundle]ALI

Satz 3525: Macro muss kundenseitig erstellt werden .
Ref: macro must be created by the customer .
Hyp: macro [needs]SYN to be created by the customer .

Satz 620: Abschluss mutieren (Superwisor) = = & gt ; Diese Funktion wird benötigt falls ein einzelner Record bearbeitet werden Muss .
Ref: edit financial statement (supervisor)==> This function is required if a single record is to be edited
Hyp: [balancing Modify]WRO ([Superwisor]UNK) = = & gt ; This function is required if a [individual]SYN record [be [processed]SYN]TIM [You need]EXW .

Satz 1601: Printing System , Design
Ref: printing system / Design
Hyp: printing system [,]PUN [design]TRU

Satz 2974: Einführung nicht möglich
Ref: introduction not possible
Hyp: [implementation]SYN not possible

Satz 1375: steuert ob die Funktion & quot ; Segmentdetail öffnen & quot ; in der Maske Strategieprüfung aufgerufen werden kann .
Ref: controls if the " Open segment detail " function can be selected in the Strategy Check screen or not .
Hyp: [controls [whether]SYN the function & quot ; open & quot ; [Segmentdetail]COM in the screen []GAP strategy can be [called]SYN []DIF .]ORD

Satz 2565: * Eigene Konti CHF
Ref: own Savings 3 accounts
Hyp: [* Own CHF accounts]ALI

Satz 1407: 1 (Methode 1)
Ref: 1 (method 1)
Hyp: 1 (method 1)

Satz 408: Vgl . Kap . 1.3 Allgemeine Hinweise .
Ref: see chapter 1.3 General notes
Hyp: [enter]WRO [.]PUN chapter [.]PUN 1.3 General notes [.]PUN

Satz 341: definiert die Anzahl Tage , für die die Daten in der Tabelle OM_CHK_SALDO erhalten bleiben
Ref: defines the number of days to retain the data in table OM_CHK_SALDO .
Hyp: [defines the number of days [,]PUN for [the]EXW the data in [the]EXT table OM_CHK_SALDO [remain receive]SYN []PUN]ORD

Satz 3101: der Parameterwert entspricht einer in KD1KDKRK erfassten Rolle .
Ref: parameter value corresponds with a role recorded in KD1KDKRK .
Hyp: [[the]EXT parameter value corresponds [to]WRO a recorded in KD1KDKRK role .]ORD

Satz 3511: pro Dokumenttyp muss eine Dokumentenvorlage erstellt werden , welche dann durch das Macro mit Daten gefüllt werden muss .
Ref: you must create a document template per document type , which the macro will then fill with data .
Hyp: [[[]GAP per document type a document template [needs]SYN to be created , which then [by]EXW the macro [needs]SYN to be filled with data .]ORD]TIM

Satz 1694: wird der betreffende Record * nicht * verarbeitet und die Daten bleiben unverändert .
Ref: the corresponding record is " not " processed and the data is left unchanged .
Hyp: [the []GAP record [*]PUN not [*]PUN processed and the data [remain]SYN unchanged .]TIM

Satz 2921: Erhaltene mit erwarteter Sequenznummer vergleichen .
Ref: compare received with expected sequence number
Hyp: [received with expected sequence number compare [.]PUN]ORD

Satz 1586: & LT ; Mod Dur Eff & gt ; wird ausschliesslich dann befüllt , wenn ein Sollwert vorhanden ist .
Ref: & LT ; Mod Dur eff & gt ; is only filled if there is a target value .
Hyp: & LT ; [Change of Dur securiti]WRO & gt ; is only [then]EXW filled [,]PUN if []GAP [a [Sollwert]COM is [available]EXW]ORD .

Satz 2986: Einführung nicht möglich
Ref: introduction not possible
Hyp: [implementation]SYN not possible

Satz 2913: Ungeordnete Gruppenfelder überprüfen .
Ref: check unordered group fields
Hyp: [Ungeordnete]UNK [Gruppenfelder]COM check [.]PUN

Satz 1384: max .
Ref: can be max .
Hyp: [max .]ALI

Satz 1386: dieser Paramter bestimmt , wo die generierte CSV-Datei gespeichert wird .
Ref: this parameter indicates where the generated CSV file is saved .
Hyp: this parameter [determines]SYN where the generated CSV file is saved .

Satz 3608: Gebäude (techn .
Ref: building (tech .
Hyp: [building (techn .]ALI

Satz 741: für das ablösen einiger JCS-Jobs ist ein generelles Konzept erarbeitet worden (Kapitel 2.1) .
Ref: a general concept has been developed for the replacement of some JCS jobs (chapter 2.1) .
Hyp: [[for the replace some JCS jobs is a general concept has been developed [been]EXW (chapter 2.1) .]ORD]TIM

Satz 3295: wenn PrgPar KT_GRP/UPDATE_SITZ auf 1 steht und PrgPar DEF_KD_STAMM/SITZ_DEFAULT auf 2 steht , wird zusätzlich auch der Sitz des Defaultberaters (Rolle 5000) übernommen .
Ref: if the program parameter KT_GRP/UPDATE_SITZ is at 1 , and the program parameter DEF_KD_STAMM/SITZ_DEFAULT is at 2 , the branch of the default advisor (Role 5000) is additionally adopted .
Hyp: [[if []GAP program parameter KT_GRP/UPDATE_SITZ is [on]SYN 1 and [PrgPar]SYN DEF_KD_STAMM/SITZ_DEFAULT is set to 2 [is]EXW , is [also]EXW additionally the [office]WRO of the default adviser [,]PUN [as]WRO ([role]TRU 5000) []GAP .]ORD]TIM

Satz 2336: geloggt wird nach AL_LOG , dass ein Portfolio gelöscht wurde , unter Angabe des Benutzers .
Ref: the user specifies that a portfolio has been deleted and logs this information according to AL_LOG .
Hyp: [[logged is after AL_LOG , that a portfolio has been deleted , under indication of the user .]TIM]DIF

Glossar

Abdeckung (engl.: recall) Ein Mass, welches aussagt, wie viel der gelieferten Information relevant ist.

***n*-Gramm** Das Ergebnis einer Zerlegung eines Textes in Fragmente (beispielsweise Worte), wobei davon n Fragmenten zusammen genommen werden zu einem n-Gramm.

Paralleler Korpus die Texte liegen mindestens in zwei Sprachen vor und die Sätze sind miteinander verknüpft.

Präzision (engl.: precision) Ein Mass, welches aussagt, wie viel der gelieferten Information wirklich korrekt ist.

Stammformreduktion (engl.: Stemming) Ein Wort auf seinen Wortstamm reduzieren.

Literaturverzeichnis

Bird, S. (Juli 2006). NLTK: the natural language toolkit. *Proceedings of the COLING/ACL on Interactive presentation sessions* , 69-72.

Brown, P. F., Pietra, V. J., Pietra, S. A., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics , 19.2*, 263-311.

Clark, J., Deyer, C., Lavie, A., & Smith, N. (Juni 2011). Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. *49th Annual Meeting of the Assocation for Computational Linguistics:shortpapers* , 176-181.

Federico, M., Bertoldi, N., & Cettolo, M. (2008). *IRSTLM: an Open Source Toolkit for Handling Large Sacle Language Models.* Brisbane, Australia: Proceedings of Interspeech.

Google Übersetzer. (11. 8 2014). Von https://translate.google.de abgerufen

Jekat, S., & Volk, M. (2010). Maschinelle und computergestützte Übersetzung. In K.-U. Carstensen, C. Ebert, C. Ebert, S. Hekat, R. Klabunde, & H. Langer, *Computerlinguistik und Sprachtechnologie - Eine Einführung* (3. Auflage Ausg.). Heidelberg: Spektrum.

Koehn, P. (September 2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *MT summit* , 79-86.

Koehn, P. (2010). *Statistical Machine Translation.* Cambridge University Press.

Koehn, P., & Hoang, H. (Juni 2008). Design of the Moses Decoder for Statistical Machine Translation. *Software Engineering, Testing, and Quality Assurance for Natural Language Processing* , 58-65.

Koehn, P., & Knight, K. (April 2003). Empirical Methods for Compound Splitting. *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics* , 187-193.

Koehn, P., Axelrod, A., Mayne, A. B., Callison-Burch, C., Osborne, M., & Talbot, D. (Oktober 2005). Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. 68-75.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., et al. (June 2007). Moses: Open Source Toolkit for Statistical Machine Translation. *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions* , 177-180.

Koehn, P., Och, F. J., & Marcu, D. (Mai 2003). Statistical phrase-based translation. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* , 48-54.

Lavie, A. (2010). *Evaluating the Output of Machine Translation Systems.* Denver, Colorado, USA: AMTA 2010 Tutorial.

LEO.org. (11. 8 2014). Von http://dict.leo.org abgerufen

LinguaTec Sprachtechnologien. (11. 8 2014). Von http://www.linguatec.net abgerufen

Och, F. J. (Juli 2003). Minimum Error Rate Training in Statistical Machine Translation. *Proc. of the 41th Annual Meeting of the Association for Computional Linguistics* , 160-167.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics* (29.1), 19-51.

Sennrich, R. (April 2012). Perplexity Minimization for Translation Model Domain Adaptation in Statistical Machine Translation. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* , 539-549.

Sennrich, R. (2009). *Syntactically Enriched Statistical Machine Translation from English to German.* Zürich: Universität Zürich.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., et al. (Mai 2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation* , 2142-2147.

Tiedemann, J. (2009). News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. N. Mitkov, *Recent Advances in Natural Language Processing* (S. 237-248). Amsterdam/Philadelphia.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., & Nagy, V. (2005). Parallel corpora for medium density languages. In *In Proceedings of the RANLP 2005* (S. 590-596).

Wikipedia. (11. 8 2014). Von http://de.wikipedia.org abgerufen