CROCO

LINGUISTIC PROPERTIES OF TRANSLATIONS
A CORPUS-BASED INVESTIGATION FOR THE LANGUAGE PAIR ENGLISH-GERMAN

Title **Multidimensionales Korpus-Alignment**

Author Mihaela Vela, Silvia Hansen-Schirra

Deliverable No. 3
Work package 1.4
Status *final version*
Availability *local*
Date *26 April 2006*

DFG project STE 840/5-1
http://fr46.uni-saarland.de/croco/

Mulidimensionales Korpus-Alignment

1 Einleitung

Um Explizierung und andere übersetzungsspezifische Eigenschaften analysieren zu können, wird das CroCo-Korpus auf verschiedenen Ebenen aligniert: Wörter, Einzelsätze und Sätze. In diesem Bericht geht es allerdings nicht um die Regeln, entsprechend derer das Korpus aligniert wird (siehe hierzu die „Richtlinien für die Alignierung“, die als Dokumentation zum CroCo-Korpus mitgeliefert werden), sondern vielmehr um die technische Umsetzung während des Alignierungsprozesses. Es wird beschrieben, welche (halb-) automatischen Alignment-Programme verwendet werden und wie diese angepasst werden mussten bzw. inwiefern Nach-Bearbeitung erforderlich war. Aus diesem Grund ist der hier vorliegende Bericht eher als eine Art Handbuch zu sehen, dass den Umgang mit den für den Alignierungsprozess notwendigen Programmen beschreibt sowie die Aufbereitung des Korpus für die Alignierung.

Kapitel 2 beschäftigt sich mit dem Satz-Alignment, Kapitel 3 mit dem Einzelsatz-Alignment und Kapitel 4 mit dem Wort-Alignment. Darüberhinaus wird die Repräsentation (siehe Kapitel 5) und Abfrage (siehe Kapitel 6) des alignierten Korpus diskutiert.

2 Satz-Alignment

Für das Satz-Alignment wird WinAlign verwendet. Dieses Alignment-Programm ist im Paket der Translator’s Workbench, dem Translation Memory von Trados enthalten (vgl. Heyn 1996). Der Alignierungsprozess verläuft halbautomatisch: WinAlign schlägt für jedes Alignment-Projekt automatisch eine Alignierung vor. Dabei erstellt das Werkzeug aus den ausgangs- und zielsprachlichen Texten zweisprachige Satzpaare. Zur Erkennung von Satzgrenzen wir die Interpunktion zugrunde gelegt. Hierbei dienen allerdings auch Abkürzungslisten als Unterstützung, um zu verhindern, dass nach jeder Abkürzung eine Satzgrenze identifiziert wird. Außerdem spielt die relative Länge von Sätzen bei der automatischen Alignierung auch eine Rolle, so dass beispielsweise bei einer Satz-Aufsplittung zwei kurze Sätze in der Übersetzung mit einem langen Satz im Original aligniert werden. Der eigentliche Alignierungsprozess wird von einem Annotierer überwacht, damit möglichst viele und vor allem sinnvolle Satzpaare ermittelt werden können.

Das Programm lässt, wie schon erwähnt, 1:n-Alignierungen zu (siehe Abbildung 1). Aus technischen Gründen ist es nicht möglich, beispielsweise Satz 1 im Deutschen mit Satz 1 und 2 im Englischen und gleichzeitig den englischen Satz 2 mit dem deutschen Satz 2 zu alignieren. Wenn diese Fälle auftreten, werden die zu alignierenden Sätze entweder an einer semantisch sinnvollen Stelle getrennt oder sie werden zusammengezogen. Wenn ein Satz keine Entsprechung hat, bei so genannten „Empty Links“, muss der Annotierer ein leeres Segment einfügen, da ansonsten das nicht-alignierte Segment verloren gehen würde. Das Output wird als Text-Datei gespeichert, in der der Originalsatz und der übersetzte Satz jeweils in einer eigenen Zeile stehen und durch einen Semikolon getrennt werden. Für die weitere Verarbeitung im CroCo-Projekt werden diese Dateien mittels Perl-Skript in XCES überführt (siehe Kapitel 5).

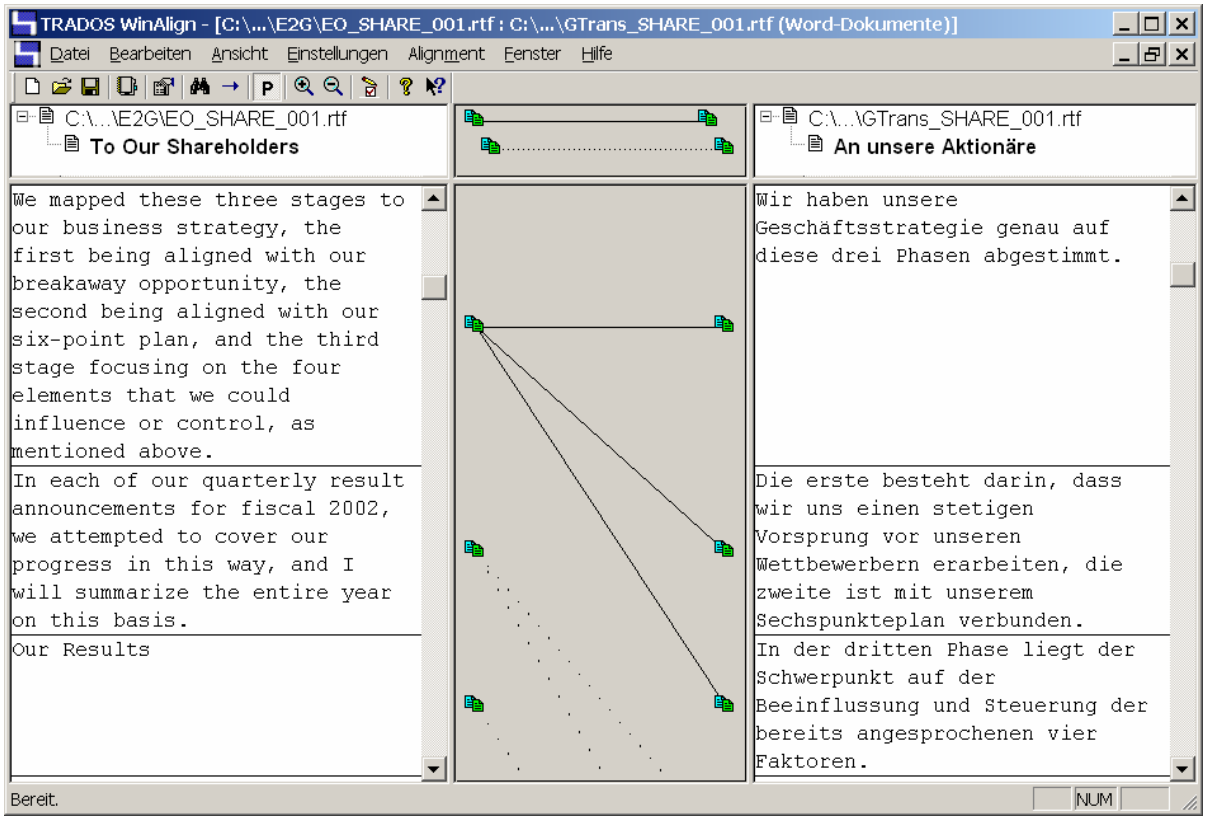

Abbildung 1: Satz-Alignment

3 Einzelsatz-Alignment[1]

Der Segmentierungs- und Alignierungsprozess der Einzelsätze wird in CroCo mit MMAX2 (Müller & Strube 2003) in einem Schritt durchgeführt. Das heisst, zuerst wird ein Einzelsatz segmentiert und im nächsten Schritt wird der entprechende übersetzte Einzelsatz damit aligniert. In MMAX2 werden linguistische Einheiten wie Phrasen, Einzelsätze, Sätze und Paragraphen *Markables* gennant. Jedes *Markable* entspricht einer Datei und dadurch auch einer Ebene im MMAX2-Projekt.

Um annotieren und alignieren zu können müssen zuerst die mit WordAlign (TRADOS) alignierten Sätze in ein für MMAX2 entsprechendes Format gebracht werden. Zusätzlich zu dieser Formatkonvertierung werden für MMAX2 notwendige Konfigurationsdateien erstellt, die dann mit der Datei **filename.mmax** verknüpft werden. Die Datei **filename.mmax** selber wird in eine ausführbare Datei **filename.bat** eingebunden, die durch doppelten Mausklick aufgerufen wird.

Alle benötigte Konvertierungen, Erzeugungen von Dateien und Verknüpfungen zwischen den Dateien werden automatisch mit einem Shell Skript erzeugt.

[1] Die Annotation der syntaktischen Funktionen wird – technisch gesehen – ähnlich durchgeführt wie die Einzelsatz-Alignierung. Die Beschreibung der Vorbereitung des Korpus und der Nutzung von MMAX2 ist daher sowohl für das Einzelsatz-Alignment sowie für die Annotation der syntaktischen Funktionen gültig.

3.1. Konfigurationsdateien für die Alignierung von Einzelsätzne mit MMAX2

Zu den Konfigurationsdateien, die für die Annotation und Alignierung von Einzelsätze notwendig sind, gehören: **clause_customization.xml**, **croco_style.xsl**, **clause_scheme.xml** und **common_paths.xml**. Die beiden Dateien **clause_customization.xml** und **croco_style.xsl** werden benötigt, um die schon bearbeiteten Phrasen in MMAX2 graphisch von den noch nicht bearbeiteten zu unterscheiden.

Die Datei **clause_scheme.xml** wird benötigt, um eine Liste von möglichen Attributen für einen Einzelsatz zu definieren. Einzelsätze werden in CroCo nicht weiter kategorisiert, nachdem sie als solche markiert werden. Dies wird in Abbildung 2 sichtbar, wo **type** nur einen einzigen Wert zugewiesen bekommt, nämlich **clause**. In dieser Datei werden auch mögliche Relationen zwischen *Markables* definiert wie zum Beispiel die Regel, dass nur Einzelsätze miteinander aligniert werden können.

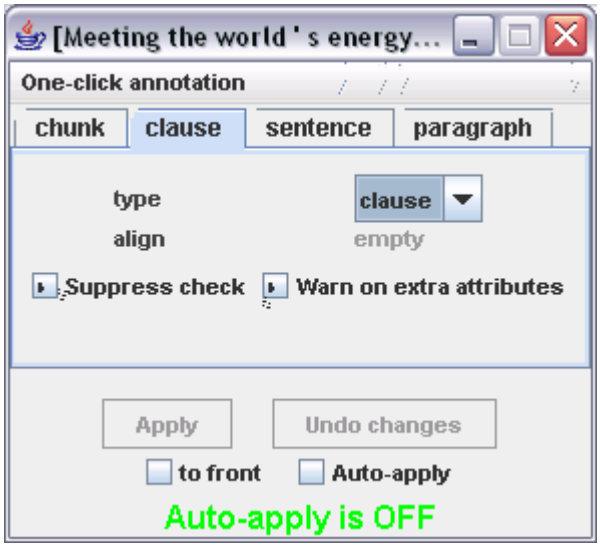

Abbildung 2: Attribut-Fenster für Einzelsätze

Die Datei **common_paths.xml** wird von MMAX2 gebraucht, um die Verknüpfungen zwischen den einzelnen Ebenen (*Markables*) in einem MMAX2-Projekt herzustellen.

3.2 Markable-Dateien für die Alignierung von Einzelsätzen mit MMAX2

Die *Markable-Datei*en für Einzelsätze kodieren in XML die verschiedenen Ebenen, die in MMAX2 bearbeitet werden können. Jede Ebene wird in eine separate Datei geschrieben und entspricht einem *Markable*. Für die Einzelsätze werden aus den satzalignierten Dateien, mittels eines Perl Skripts, *Markable-Datei*en für Paragraph, Satz, Phrase und Wort konvertiert: **filename_paragraph_level.xml**, **filename_sentence_level.xml**, **filename_clause_level.xml** und **filename_words.xml**.

Die Paragraph-, Satz- und Wortebenen werden in MMAX2 schon vorgegeben und müssen in dem MMAX2-Browser nicht mehr bearbeitet werden. Diese Dateien dienen dazu, Paragraphen und Sätze anhand der Wörterindexzahlen zu markieren. Um im Alignierungsprozess die Verschiebung der Paragraph- und Satzgrenzen zu vermeiden, müssen in der Konsole in Abbildung 3 nur die Einzelsätze als aktiv geschaltet werden. Die anderen Ebenen müssen sichtbar sein, aber nicht veränderbar, was mit der *visible* Einstellung gewährleistet wird.

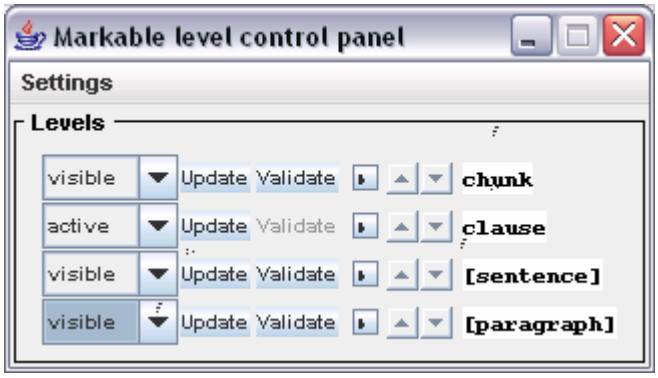

Abbildung 3: Aktivierung der *Clause*-Ebene in MMAX2

In der Datei **filename_paragraph_level.xml** werden in XML mit Hilfe des Attributes **span** und der Wortindexzahlen die Grenzen eines Paragraphes festgelegt:

```
<markable id="1" span="word_1..word_34" type="paragraph" />
<markable id="2" span="word_35..word_89" type="paragraph" />
<markable id="3" span="word_90..word_125" type="paragraph" />
<markable id="4" span="word_126..word_245" type="paragraph" />
<markable id="5" span="word_246..word_304" type="paragraph" />
<markable id="6" span="word_305..word_312" type="paragraph" />
<markable id="7" span="word_313..word_364" type="paragraph" />
<markable id="8" span="word_365..word_429" type="paragraph" />
<markable id="9" span="word_430..word_490" type="paragraph" />
```

Abbildung 4: Segmentierung in der Datei **filename_paragraph_level.xml**

Für CroCo wird der Paragraph als Satzpaar definiert, indem der erste Satz der Originalsatz ist und der zweite der alignierten Übersetzung entspricht. Auf diese Weise wird die Annotation und die Alignierung der linguistischen Einheiten innerhalb der Sätze vereinfacht. Durch die Einstellungen in den Konfigurationsdatein werden Paragraphen durch grüne runde Klammern und durch eine unterbrochene Linie voneinander abgegrenzt (siehe Abbildung 5).

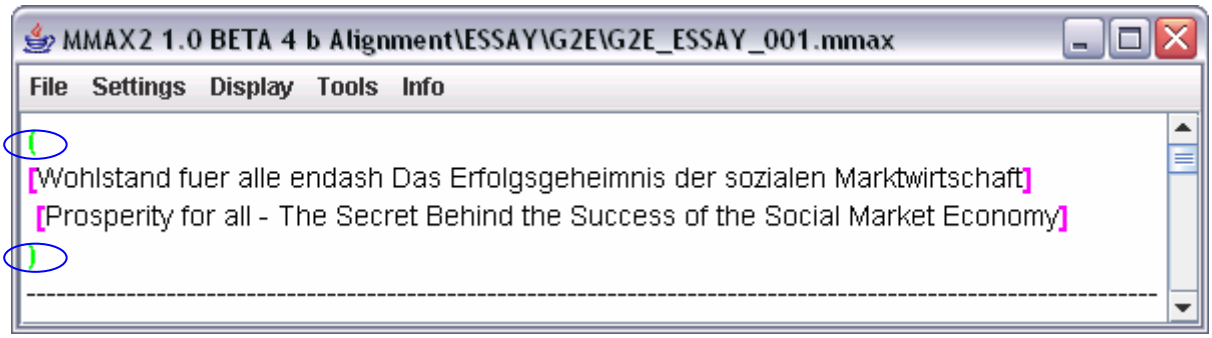

Abbildung 5: Satzpaar im MMAX2-Browser

Analog zum Paragraphen, wird in der Datei **filename_sentence_level.xml** (siehe Abbildung 6) anhand der Wordindexzahlen die Satzgrenze festgelegt. Graphisch wird die Satzgrenze durch magentafarbene eckige Klammern dargestellt (siehe Abbildung 7).

```
<markable id="1" span="word_1..word_20" type="sentence" />
<markable id="2" span="word_21..word_34" type="sentence" />
<markable id="3" span="word_35..word_63" type="sentence" />
<markable id="4" span="word_64..word_89" type="sentence" />
<markable id="5" span="word_90..word_109" type="sentence" />
<markable id="6" span="word_110..word_125" type="sentence" />
<markable id="7" span="word_126..word_183" type="sentence" />
<markable id="8" span="word_184..word_245" type="sentence" />
<markable id="9" span="word_246..word_273" type="sentence" />
<markable id="10" span="word_274..word_304" type="sentence" />
```

Abbildung 6: Segmentierung in der Datei **filename_sentence_level.xml**

Abbildung 7: Satz im MMAX2-Browser

Die Datei **filename_clause_level.xml** ist am Anfang der Alignierungsarbeit leer. In dieser Datei werden die im MMAX2-Browser durchgeführte Alignierung auf der Einzelsatzebene gespeichert. Im Kontext des Croco-Projekts werden Einzelsätze nach Wortbedeutung aligniert. Ein Einzelsatz enthält eine finite oder infinite Verbform sowie die es begleitenden Satzglieder. In der Regel enthält ein Einzelsatz nie mehr als eine finite oder infinite Verbform. Für das Alignment dient das Verb als Grundlage (siehe hierzu die „Richtlinien für die Alignierung“).

Um einen Einzelsatz zu segmentieren, muss man mit der linken Maustaste alle Wörter markieren die zu dem Einzelsatz gehören. Danach wählt man aus dem Merkmalfenster (siehe Abbildung 2) den Einzelsatztyp, welcher nur **clause** sein kann. Diese vom Annotierer getroffene Auswahl wird automatisch von MMAX2 in die Datei **filename_clause_level.xml** (siehe Abbildung 8) eingetragen.

```
<markable id="markable_389" span="word_2451..word_2461" type="clause"
align="empty" />
<markable id="markable_579" span="word_4345..word_4355" type="clause"
align="empty" />
<markable id="markable_429" span="word_2864..word_2877" type="clause"
align="markable_432"/>
```

Abbildung 8: Die Annotation in der Datei **filename_clause_level.xml**

Der Alignierungsprozess in MMAX2 impliziert, dass die Phrasen schon segmentiert sind. Um zu alignieren, muss man zunächst die Phrase im Ausgangstext mit der rechten Maustaste markieren. Dann geht man zur Phrase im Zieltext und markiert diese mit der linken Maustaste. Als Folge erscheint ein Pop-Up-Fenster in dem **Mark as aligned chunk** steht (siehe Abbildung 9). Durch auswählen dieses Textes, mit der linken Maustatste, werden die zwei Einzelsätze miteinander aligniert und die Information wird in die Datei **filename_clause_level.xml** (siehe Abbildung 8) eingetragen.

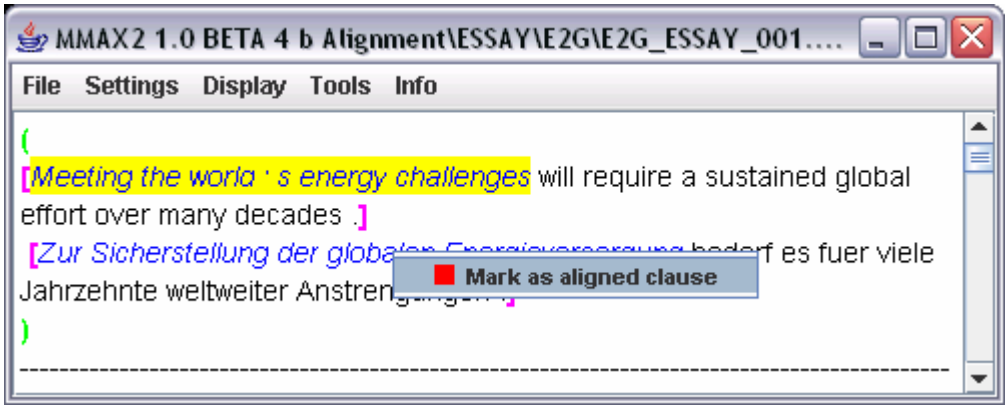

Abbildung 9: Der Alignierungsprozess der Einzelsätze im MMAX2-Browser

In der Datei **filename_words.xml** werden im XML-Format alle Wörter einer Datei aufgelistet. In dieser Datei wird jedem Wort ein Index zugewiesen, so dass die Paragraph-, Satz- und Phrasengrenzen anhand dieser Indexierung festgelegt werden kann (siehe Abbildung 10).

```
<word id="word_18">Zur</word>
<word id="word_19">Sicherstellung</word>
<word id="word_20">der</word>
<word id="word_21">globalen</word>
<word id="word_22">Energieversorgung</word>
```

Abbildung 10: Die Annotation in der Datei **filename_words_level.xml**

Für die weitere Verarbeitung im CroCo-Projekt werden die MMAX2-Dateien mittels Perl-Skript in XCES überführt (siehe Kapitel 5).

4 Wort-Alignment

Für das Wort-Alignment wird in CroCo ein neues Alignment-Verfahren verwendet, das explizit strukturierte Informationen ausnutzen kann (vgl. Schrader 2006). Hierfür werden mehrere linguistische Beschreibungsebenen herangezogen, wie beispielsweise Informationen über Lemmas und Wortarten, oder auch syntaktische Information. Außerdem werden Zuordnungen sowohl anhand statistischer Ähnlichkeiten zwischen Wörtern oder Wortgruppen getroffen als auch anhand linguistisch motivierter Regeln. Als Trainingsmaterial für die statistische Komponente des Programms dient das Europarl-Korpus (vgl. Koehn 2002).

Mit Hilfe dieses Programms können auch Problemfälle, die typologisch bedingt sind, automatisch aligniert werden. So werden zum Beispiel (auch selten vorkommende) Komposita im Deutschen ihren Mehr-Wort-Entsprechungen im Englischen zugeordnet (z.B. Eigentumsbeschädigungen – damage to property (vgl. Schrader 2006, 189)). Dies geschieht auf der Basis von vorher annotierten Part-of-Speech-Tags sowie Wortlängenähnlichkeiten (gemessen in Buchstaben). Die Kombination von statistischen Methoden und linguistischer Intelligenz (in Form eines mehrschichtig annotierten Korpus) scheint viel versprechend im Kontext des CroCo-Projekts. Auf diese Weise können auf großen Datenmengen errechnete Wahrscheinlichkeiten mit linguistischem Input kombiniert werden. Zudem ist das Programm auf die CroCo-Annotation und –Register anpassbar und somit für das Wort-Alignment in CroCo optimierbar.

5 Multidimensionale Repräsentation

Durch den Alignierungsprozess werden drei neue Alignment-Ebenen produziert (Satz-, Einzelsatz-, Wort-Ebene), die in verschiedenen Formaten vorliegen. Um die Alignment-Ebenen vergleichbar zu machen und sie für die Abfrage zu vereinheitlichen werden sie mittels Perl-Skripts in XCES (Corpus Encoding Standard for XML; Ide et al. 2000) überführt. Die verschiedenen Alignment-Ebenen werden im *multi-layer XML stand-off Mark-up* gespeichert. Dies bedeutet, dass die verschiedenen Ebenen voneinander getrennt bleiben und dass das Korpus indexiert wird. Die Indexierung wird auf jeder Alignment-Ebene separat vollzogen. D.h. jedes Wort, jeder Einzelsatz und jeder Satz erhalten eine ID und werden in separaten Index-Dateien gespeichert. Abbildung 11 verdeutlicht die Indexierung auf Wort-Ebene mit dem deutschen Original-Satz „Ihre Hände ließen ihn leise wimmern.“ und dessen englischer Übersetzung „He whimpered softly under her hands.".

```
<document xmlns:xlink="http://www.w3.org/1999/
xlink" name="GO.tok.xml" xml:lang="de"
docType="ori">
<header xlink:href="GO.header.xml"/>
 <tokens>
  <token id="t1" strg="Ihre"/>
  <token id="t2" strg="Hände"/>
  <token id="t3" strg="ließen"/>
  <token id="t4" strg="ihn"/>
  <token id="t5" strg="leise"/>
  <token id="t6" strg="wimmern"/>
 </tokens>
</document>
```

```
<document xmlns:xlink="http://www.w3.org/1999/
xlink" name="ETrans.tok.xml" xml:lang="en"
docType="trans">
<header xlink:href="ETrans.header.xml"/>
 <tokens>
  <token id="t1" strg="He"/>
  <token id="t2" strg="whimpered"/>
  <token id="t3" strg="softly"/>
  <token id="t4" strg="under"/>
  <token id="t5" strg="her"/>
  <token id="t6" strg="hands"/>
 </tokens>
</document>
```

Abbildung 11: Indexierung auf Wort-Ebene

In Abbildung 11 sieht man, dass jedes Token ein Attribut **strg** erhält, in dem der eigentliche Text steht, und ein Attribut für die Indexierung (**id**), das die Position des Tokens im Text anzeigt. Diese **id** dient als Anker für alle XPointer, die aus Annotations- und Alignment-Ebenen auf sie verweisen. Die Ebene wird hierbei durch das entsprechende Kürzel (hier "t“ für Token) in der **id** angegeben. Der Verweis auf die Index-Datei wird durch das Attribute **name** operationalisiert. Das Attribut **xml:lang** gibt die Sprache an und **docType** verrät, ob es ein Original oder eine Übersetzung ist. Für die Einzelsätze und die Sätze werden entsprechende Index-Dateien produziert.

Für das Alignment ist das Attribut **trans.loc** wichtig (siehe Abbildung 12): Dieses Attribut verlinkt die Alignment-Datei mit der dazugehörigen Index-Datei. Darüber hinaus gibt das Attribut **xml:lang** die zu alignierenden Sprachen an, wobei das Attribut **n** die Abfolge der Alignierung anzeigt. Das eigentliche Alignment wird durch ein Elementpaar **align** realisiert, in dem in jeder Sprache durch XPointer auf die alignierten Tokens verwiesen wird. Wörter, für die keine Entsprechung gefunden werden kann, also „Empty Links“, erhalten den Attributwert “#undefined”.

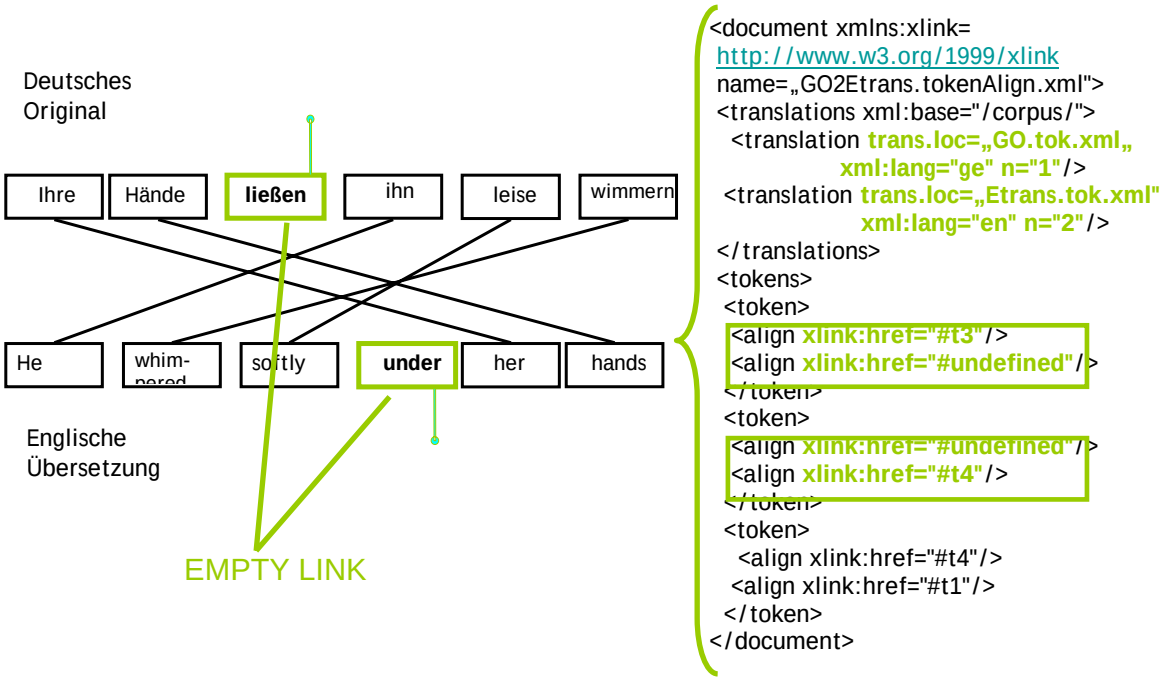

Abbildung 12: Alignierung auf Wort-Ebene

6. Abfrage des Alignments

Zur Zeit erfolgt die Abfrage der in XML gespeicherten, multidimensionalen Korpus-Annotation einschließlich Alignment mit XQuery[2]. Diese Abfragesprache eignet sich besonders gut im Kontext des CroCo-Projekts, da man mit XQuery Informationen aus unterschiedlichen Quelldateien, wie beispielsweise verschiedene Alignment- und Annotations-Ebenen, abrufen kann. Die Anwendung dieser Abfragesprache für multilinguale Korpora, die auf unterschiedlichen Ebenen annotiert sind, wird in (Teich et al. 2001) näher beschrieben. Anhand des Beispiels in Abbildung 12 soll im folgenden exemplarisch eine Abfrage für „Empty Links“ diskutiert werden.

Eine Abfrage für „Empty Links“ auf der Wortebene könnte wie folgt aussehen: Finde alle Wörter, die keine alignierte Entsprechung haben, d.h. die für das Attribut **xlink** den Wert “#undefined” erhalten. Die Realisierung der Abfrage in XQuery ist in Abbildung 13 dargestellt.[3] Diese Abfrage kann natürlich auch für „Empty Links“ auf der Ebene der Einzelsätze und Sätze gestellt werden.

[2] http://www.w3.org/TR/xquery

[3] Eine detaillierte Beschreibung dieser Abfrage ist in (Hansen-Schirra et al. 2006) zu finden.

```
let $doc := .
for $k in $doc//tokens/token
return
 if ($k/align[1][@xlink:href="#un-
  defined"] and $k/align[2]
  [@xlink:href!="#undefined"])
 then local:getString($k/align[1]/
  @xlink:href,$k/align[2]/@xlink:href,
  $doc//translations/translation
  [@n='2']/@trans.loc)
 else if ($k/align[1][@xlink:href
  !="#undefined"] and $k/align[2]
  [@xlink:href="#undefined"])
 then local:getString($k/align[1]/
  @xlink:href,$k/align[2]/@xlink:href,
  $doc//translations/translation
  [@n='1']/@trans.loc)
 else ()
```

Abbildung 13: XQuery für „Empty Links“

Wird diese Abfrage auf die Sätze in Abbildung 12 angewendet, werden als Ergebnis die Wörter *ließen* und *under* angezeigt. Diese Wörter haben keine Entsprechung in der jeweiligen anderen Sprache und erhalten somit „Empty Links“. Die Abfrage und Untersuchung der „Empty Links“ auf verschiedenen Alignment-Ebenen kann wichtige Rückschlüsse für die im Rahmen des CroCo-Projekts untersuchte Übersetzungseigenschaft Explizierung liefern. Um die Korpusanalyse zu erleichtern, soll in der Auswertungsphase des Projekts eine graphische Benutzerschnittstelle für die Abfrage zum Einsatz kommen. Zum jetzigen Zeitpunkt wird noch überprüft, ob hierfür ein bereits existierendes Suchwerkzeug den Anforderungen des CroCo-Projekts genügt (z.B. NITE, Kilgour & Carletta 2006) oder ob eine eigene Suchfunktion entwickelt werden muss.

References

Silvia Hansen-Schirra, Stella Neumann & Mihaela Vela. 2006. Multi-dimensional Annotation and Alignment in an English-German Translation Corpus. *Proceedings of EACL Workshop “Multi-dimensional mark-up in NLP”*, Trento.

Matthias Heyn. 1996. Integrating machine translation into translation memory systems. *European Association for Machine Translation - Workshop Proceedings*, ISSCO, Geneva:111-123.

Jonathan Kilgour & Jean Carletta. 2006. The NITE XML Toolkit: Demonstration from five corpora. *Proceedings of EACL Workshop “Multi-dimensional mark-up in NLP”*, Trento. http://www.ltg.ed.ac.uk/NITE/

Philip Koehn. 2002. *Europarl: A Multilingual Corpus for Evaluation of Machine Translation*. Draft, unpublished. http://people.csail.mit.edu/koehn/publications/europarl/

Christoph Müller and Michael Strube. 2003. Multi-Level Annotation in MMAX. *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, Sapporo, Japan:198-107.

Nancy Ide, Patrice Bonhomme and Laurent Romary. 2000. XCES: An XML-based Encoding Standard for Linguistic Corpora. *Proceedings of the Second Language Resources and Evaluation Conference* (LREC), Athens, Greece:825-830. http://www.xml-ces.org

Bettina Schrader. 2006. How does morphological complexity translate? A cross-linguistic case study for word alignment. *Proceedings of Linguistic Evidence Conference*, Tübingen: 189-191.

Elke Teich, Silvia Hansen, and Peter Fankhauser. 2001. Representing and querying multi-layer annotated corpora. *Proceedings of the IRCS Workshop on Linguistic Databases*. Philadelphia: 228-237.