Corpora als Ressourcen für die maschinelle Übersetzung

Luzius Thöny
Brunnenwiesenstr.19
8610 Uster
079 779 40 86
luzi1@gmx.net

Seminar *Maschinelle Übersetzung*
Dr. S. Jekat
Sommersemester 04

Inhaltsverzeichnis

Abstract: Diese Arbeit hat die Frage nach dem Nutzen von Corpora als Ressourcen für die maschinelle und maschinengestützte Übersetzung zum Thema. Neben der Verwendung von einsprachigen und zweisprachigen (parallelen) Corpora für die maschinengestütze Übersetzung wird auch der Nutzen dieser Corpora für die eigentliche maschinelle Übersetzung diskutiert. Dafür sollen die drei MT-Teilgebiete *Rule-Based Machine Translation* (RBMT), *Statistical Translation* (SMT) und *Example-Based Translation* (EBMT) betrachtet werden.
Informationen zu Aufbau, Erstellung, Codierung und Annotation von Corpora machen den ersten Teil aus. Danach folgen Darstellungen zur Verwendbarkeit zunächst von einsprachigen und im anschliessenden Teil zu parallelen Corpora. Den Schluss bildet das konkrete Beispiel des OPUS Corpus und einige Details zum Konkordanzprogramm TextSTAT.

1 Einführung: Corpora

1.1 Die Relevanz von Corpora

Obwohl die Beschäftigung mit Textcorpora noch nicht sehr lange zurückgeht, sind Corpora beim heutigen Stand der Technik in der maschinellen Übersetzung (MT) und in der maschinengestützten Übersetzung (CAT) kaum mehr wegzudenken. Corpora werden zwar auch im grossen Stil für klassische linguistische Untersuchungen, wie zum Beispiel für die Lexikographie, genutzt, doch sind sie heute speziell für Computerlinguisten als Ressourcen für die maschinelle Übersetzung unentbehrlich. Corpora werden in hybriden Systemen oft gleich für mehrere Module im Übersetzungsvorgang verwendet oder bilden, wie in der *beispiel-basierten Übersetzung* (EBMT), sogar die einzige und ausschliessliche Ressource im Übersetzungsprozess.

1.2 Typen von Corpora

Corpora sind "...essentially large collections of text in electronic form"(BOWKER / PEARSON 2002:1). Man kann sie nach verschiedenen Kriterien klassifizieren. Meine Aufstellung folgt derjenigen von BOWKER/PEARSON(2002:11ff). Jedes der folgenden sechs Kurzkapitel beschreibt ein solches Kriterium.

1.2.1 Generelle vs. Fachsprachliche Corpora

Wenn die ausgewählten Texte aus so breit gefächterten Themenbereichen stammen, dass möglichste viele Facetten einer Sprache damit abgedeckt werden, um sie damit als Ganzes zu erfassen, nennt man dies ein generelles Corpus (*general reference corpus*). Die darin enthaltene Sprache bezeichnet man als *language for general purposes* (LGP). Nur wenige Corpora sind so konzipiert. Die meisten spezialisieren sich auf eine bestimmte Fachsprache. Man nennt ein solches Corpus *special purpose corpus*. Die entsprechende Sprache trägt den Namen *language for special purposes* (LSP). Weil sich Anwendungen maschineller Übersetzungen sowieso fast immer auf eine Fachsprache beschränken müssen, sind auch die Corpora meistens auf eine LSP ausgerichtet.

1.2.2 Geschriebene vs. Gesprochene Texte in Corpora

Je nach dem ob die enthaltenen Texte ursprünglich gesprochene sind, wie etwa bei Transkriptionen von Sitzungs- oder Parlamentsgesprächen, oder aber ob es sich um ursprünglich geschriebene Texte handelt, nennt man die Corpora entsprechend gesprochene bzw. geschriebene Corpora. Die geschriebenen Corpora sind dabei stark in der Mehrzahl. Es gibt auch Textcorpora, die Texte beider Arten enthalten.

1.2.3 Einsprachige vs. Mehrsprachige Corpora

Besonders für die MT sind die mehrsprachigen Corpora von grossem Nutzen. Man unterscheidet bei den mehrsprachigen nach gängiger Terminologie zwischen parallelen und vergleichbaren Corpora. Im ersten Fall spricht man von Texten und ihren direkten Übersetzungen in eine andere (oder mehrere andere) Sprache(n). Beim zweiten Fall, bei den vergleichbaren Corpora, handelt es sich nicht um direkte Übersetzungen, sondern lediglich um Texte verschiedener Sprachen, die ins gleiche Genre gehören und der gleichen Fachsprache angehören. Sie sind jeweils von Muttersprachlern verfasst.

1.2.4 Synchrone vs. Diachrone Corpora

Auch für diachrone Linguisten haben Corpora viel zu bieten. Die Erstellung von diachronen Corpora, welche die Entwicklung einer Sprache über grössere Zeiträume weg leicht verfolgbar machen, eröffnen auch hier neue Möglichkeiten. Allerdings erschweren die oft nur spärlich vorhandenen Texte aus früheren Jahrhunderten sowie die schwere Verfügbarkeit dieser Quellen in digitalisierter Form den Aufbau eines diachronen Korpus beträchtlich. In Sprachen wie etwa dem Griechischem wäre theoretisch ein Corpus denkbar, dass die Entwicklung der Sprache über mehr als 2000 Jahre hinweg dokumentiert.

1.2.5 Offene vs. Geschlossene Corpora

'Offen vs. geschlossen' markiert nur, ob dem Corpus laufend neue Texte zugefügt werden, oder ob der Umfang nach einmaliger Konzipierung nicht mehr geändert wird.

1.2.6 Lerner-Corpora

Schliesslich gibt es auch noch sogenannte Lerner-Corpora, welche sich aus Texten von Personen konstituieren, die nicht in ihrer Muttersprache schreiben. Damit erhofft man sich z.B. Einsichten über die Lernweise von Fremdsprachen zu gewinnen, womit man dann die Entwicklung von Lern- und Übungsmaterial für den Fremdsprachenunterricht verbessern könnte. Für die MT sind Lerner-Corpora aber unwichtig.

1.3 CES and TEI

Bei der Zusammenstellung eines Corpus wird man bald vor die Frage gestellt, in welcher Form man die Textdaten in der Sammlung halten möchte. Auf jeden Fall will man eine Gliederung und gewisse Metainformationen zum Text haben, damit sie für die *Corpus-Tools*, die man zum Arbeiten mit den Corpora benötigt, auch leicht verarbeitet werden können. Dafür bieten sich XML (Extended Markup Language) und SGML (Standard Generalized Markup Language) an. Schon früh wurden zur Standardisierung dieser Metainformationen in Corpora zwei Initiativen auf dem World Wide Web

gegründet, die einen Standard dazu herausgebracht haben. Es handelt sich um die die *Text Encoding Initiative* (TEI) und den *Corpus Encoding Standard* (CES).

1.3.1 *Text Encoding Initiative* (TEI)

Die TEI wurde 1987 gegründet und hat es sich zum Ziel gesetzt, Richtlinien für das Erstellen digitaler, maschinenlesbarer Texte herauszugeben. Diese Richtlinien gelten nicht nur für Corpora sondern für literarische und linguistische Texte aller Art, die elektronisch zur Verfügung gestellt werden sollen. Seit 2002 und der Version P4 sind die Richtlinien XML kompatibel. Zur Zeit ist die Version P5 im Entstehen. Dass sich TEI als Standard durchsetzen wird steht wohl ausser Frage, denn die Notwendigkeit für eine einheitliche Codierung der Texte ist offensichtlich und die TEI Richtlinien bieten genau das, was man von ihnen erwartet. Den Erfolg von TEI zeigt vor allem anderen die enorme Anzahl an Projekten, die sich an die Richtlinien hält.
TEI konforme Texte sollen nach Meinung des Consortium [1]:

- ... genügen, um die Texteigenschaften für den wissenschaftlichen Gebrauch zu annotieren.
- ... einfach, klar und konkret sein.
- ... einfach zu benutzen sein, auch ohne speziell darauf ausgerichtete Software.
- ... strenge Definitionen und effizientes Arbeiten ermöglichen.
- ... benutzerspezifische Erweiterungen unterstützen.
- ... mit schon vorhandenen und zukünftigen Standards kompatibel sein.

Die Struktur eines TEI Textes Der grundsätzliche Aufbau eines TEI konformen Textes sieht folgendermassen aus:

```
<TEI.2>
  <teiHeader> [ TEI Header information ] </teiHeader>
  <text>
    <front> [ front matter ... ] </front>
    <body> [ body of text ... ] </body>
    <back> [ back matter ... ] </back>
  </text>
</TEI.2>
```

Mehrere `<TEI.2>` Tags können mit einem `<teiCorpus>`-Tag in einem Dokument zusammengefasst werden.
Neben diesen elementaren Tags steht eine grosse Menge an weiteren Tags zur Verfügung, die es ermöglichen, den Text nach verschiedensten Kriterien zu gliedern. Darunter sind beispielsweise die Tags `<p>` zur Markierung eines Paragraphen, `<div>` für eine Subdivision innerhalb des `<front>`-,`<body>`- oder `<back>`-Elementes oder `<s>` für einen Satz. Die meisten Tags, wie etwa die `<div>`-Tags, kann man mit Attributen genauer spezifizieren, z.B. kann man einen 'id'-Wert angeben, der dann innhalb des Dokumentes eindeutig eine Stelle markiert und Verweise darauf ermöglicht.

[1] BURNARD 2002:1ff.

Zur Textformattierung gibt es die Tags `<emph>` (herausheben), `<term>` (Spezialterminus) oder `<titel>` (Formattierung als Titel).
Für poetische Texte gedacht sind Tags wie `<l>` für eine Zeile, `<lg>` für eine Gruppe von Zeilen, etwa einer Strophe, oder `<speaker>` zur Identifikation des Sprechers in einem Theaterstück.
Weitere Tags für Daten, zur Markierung von editorialen Eingriffen, Bibliographien oder für Querverweise, um nur ein paar zu nennen, machen die Mächtigkeit der TEI Richtlinien aus. Man kann die vollständigen Tagsets auf der TEI Homepage nachlesen (TEI CONSORTIUM 2001).

1.3.2 *Corpus Encoding Standard* (CES)

Während der TEI Standard für eine breite Palette an Texten konzipiert ist, richten sich die Vorgaben des *Corpus Encoding Standards* spezifisch an Textcorpora im Bereich der maschinellen Verarbeitung von natürlicher Sprache. CES ist eine SGML Anwendung und seit dem Jahr 2000 gibt es auch eine XML-Variante (XCES). CES ist im wesentlichen eine Erweiterung (oder konkrete Anwendung) der TEI Richtlinien.

Struktur eines CES Dokumentes Die Strukturierung des Textes hält sich eng an die vom TEI-Consortium vorgeschlagene. Beim `<header>`-Element kommen allerdings einige Informationen dazu, die man mit mit einem typischen Korpus gerne mitgeliefert bekommen würde. Die Benutzer eines Corpus benötigen unter Umständen genaue Angaben zur Veröffentlichung des verwendeten Textes, zur Textsorte, etc. Die Struktur eines typischen CES Headers ist hier abgebildet:

```
<cesHeader version="2.0">
  <fileDesc>
    <titleStmt>
      <h.title></h.title>
    </titleStmt>
    <publicationStmt>
      <distributor></distributor>
      <pubAddress></pubAddress>
      <availability></availability>
      <pubDate></pubDate>
    </publicationStmt>
    <sourceDesc>
      <biblStruct>
        <monogr>
          <h.title></h.title>
          <h.author></h.author>
          <imprint>
            <pubPlace></pubPlace>
            <publisher></publisher>
            <pubDate></pubDate>
          </imprint>
        </monogr>
      </biblStruct>
    </sourceDesc>
```

```
    </fileDesc>
  </cesHeader>
```

Mit diesem Header lassen sich eine grosse Anzahl von Metainformationen zum Corpus spezifizieren:

`<titleStmt>`	gruppiert Informationen zum Titel der verwendeten Quelle.
`<publicationStmt>`	beschreibt die Veröffentlichung des Corpus, i.e. Verfügbarkeit, Publikationsjahr, etc.
`<sourceDesc>`	Bibliographische Angaben zu den verwendeten Quellen, nicht zum Corpus.
`<titleStmt>`	gruppiert Informationen zum Titel der verwendeten Quelle.
...	...

Selbstverständlich gibt es eine grosse Anzahl von erlaubten Tags in der `<cesHeader>`-Umgebung, die im Beispiel oben nicht verwendet wurden. So zum Beispiel:

`<extent>`	gibt die Länge des Corpus in Worten oder Bytes an.
`<encodingDesc>`	dokumentiert die Art der Codierung, z.B. eine Beschreibung des ganzen Projektes, Verwendung des Tagsets, editorialer Richtlinien, etc.
...	...

Die Struktur der `<body>`-Umgebung folgt derjenigen von TEI sehr eng und braucht somit nicht noch einmal beschrieben zu werden.

1.4 Weiterführende Annotation

In manchen Fällen möchte man eine weiterführende Annotation der Texte vornehmen. Dies zum Beispiel, wenn man dem Text eine syntaktische oder morphologische Analyse mitgeben möchte. Auch das kann mit Tags gelöst werden. Bei vielen Texten werden die Resultate einer Wortartenbestimmung gleich im Text selber eingefügt. Diesen Prozess nennt man *part-of-speech* tagging (POS tagging) und er wird in der Regel von speziell dafür geschriebenen Programmen mit einer erstaunlich hohen Zuverlässigkeit ausgeführt. Die Webseite des *British National Corpus* (BNC) listet für ihren Tagger eine Fehlerquote von nur 1.15% auf (LEECH 2000).
Der Output eines klassischen Taggers sieht so aus[2]:

```
We_PP will_MD focus_NN initially_RB on_IN markup_NN
which_WDT ...
```

Die Annotation folgt in diesem Beispiel nicht in Form von SGML-Tags, was auch denkbar wäre, sondern direkt mit einem Underscore an die Wortform angehängt. Dies entspricht natürlich keiner Standardisierung und bringt beispielsweise dann Probleme, wenn im Text Worte vorkommen, die bereits einen Underscore aufweisen.
Für die Wortartenbestimmung gibt es unzählige Methoden und auch unzählige Tagsets. Sogar in der Anzahl der verwendeten Tags unterscheiden sie sich markant. Der CLAWS

[2] BOWKER/PEARSON 2002:87

Tagger, welcher im *British National Corpus* Verwendung findet, kennt in manchen Varianten über 160 Tags (UCREL 2004). Dagegen kommt das Stuttgart-Tübingen Tagset (STTS) mit rund einem Drittel davon aus, nämlich 54 Tagtypen (FELDWEG 1996).
Für flektierende Sprachen genügt aber unter Umständen eine reine Wortartenbestimmung noch nicht. Der nächste Schritt ist dann eine vollständige morphologische Analyse. Im Deutschen muss folglich pro Nomen die Bestimmung von Kasus, Genus und Numerus durchgeführt werden. Für Verben braucht man die Kategorien Tempus, Person, Numerus, Modus und Genus Verbi. Eine morphologische Analyse erfordert noch einmal einen beträchtlichen Mehraufwand, der aber wie mir scheint für das Arbeiten mit Corpora, wie etwa das Extrahieren einer Grammatik, von grosser Wichtigkeit ist.
Wenn einem die morphologische Annotation noch nicht genügt, kann man auch noch eine syntaktische vornehmen. Bei diesem Schritt muss man die komplette syntaktische Struktur der Sätze finden und in irgendeiner Form dem Text beifügen. Wenn man soweit geht, wird die Menge der Annotationsdaten so gross, dass sie die Menge der Textdaten klar übersteigt. Man benötigt deshalb ein neues Konzept, um die verschachtelte Syntax darzustellen, ohne das sie Lesbarkeit für Mensch oder Maschine zu stark beeinträchtigt wird. Syntaktisch analysierte Texte speichert man daher meistens in einer *treebank*, die durch Klammern die hierarchische Syntaxstruktur abbildet. Die Markup Sprachen XML und SGML sind für eine derartig strukturierte Datenmenge nicht geeignet. Das folgende Beispiel 'I think they should either do that,...' stammt aus der *Penn Treebank* (MARCUS 1999):

```
( (S (NP-SBJ I)
     (VP think
         (SBAR 0
               (S (NP-SBJ-1 they)
                  (VP should
                        (VP either
                             (VP do
                                 (NP that))
                              ,
                              (...)
```

Inwiefern es sich bei einer solchen Datenstruktur überhaupt noch um ein Corpus handelt, ist fragwürdig. Mit üblichen Corpus-Tools lassen sich *treebanks* jedenfalls nicht mehr vernünftig bearbeiten.

2 Einsprachige Corpora

2.1 Aufbau

Corpusgrössen reichen von wenigen tausend bis 1.6 Mia Wörtern (TRABOLD 2004). Doch die Grösse allein macht noch kein gutes Corpus aus. Ebenso wichtig ist die **Auswahl der Texte**. Für eine konkrete Übersetzungsanwendung in einer bestimmten **Fachsprache** kann beispielsweise ein kleines, aber gut ausgewähltes Corpus viel bessere Resultate liefern als ein grosses, dafür **sprachlich extrem breites Corpus**. Manche Corpora enthalten **komplette Texte**, andere jeweils nur **Textfragmente**, die dafür von einer grösseren **Anzahl Autoren** stammen. Wichtig sind desweiteren das **Publikationsdatum** der Texte sowie selbstverständlich die **Textsprache**.

2.2 Anwendung in der *Computer Aided Translation* (CAT)

Einsprachige Corpora können die Arbeit eines Übersetzers unerstützen. Die drei hauptsächlichen Hilfsandwendungen, mit denen der Übersetzer die adäquate Verwendung eines Wortes überprüfen kann, sind Konkordanzprogramme, Wortlisten und Worthäufigkeitslisten.

2.2.1 Konkordanzprogramme

Ein Konkordanzprogramm erlaubt die Eingabe eines Suchterms und extrahiert dann aus dem Corpus Textstellen, die den Suchterm beinhalten. Typischerweise werden die Sätze und Satzfragmente so angeordnet, dass der gefundene Suchterm in der Mitte des Fensters in jedem Satz genau untereinander steht. Links und Rechts wird der nähere Kontext zum Suchterm angezeigt. Hier ein Beispiel zum Suchterm *angel* in der englischen Bibel des Bibelprojekts der Universität Maryland (RESNIK 1999). . Als Konkordanzprogramm diente das TextSTAT-Tool (HÜNING 2004).

```
rding to the wisdom of an ANGEL of God, to know all thing
erod was dead, behold, an ANGEL of the Lord appeareth in
tood by me this night the ANGEL of God, whose I am, and w
.28.5' type=verse>And the ANGEL answered and said unto th
' type=verse>And unto the ANGEL of the church of the Laod
ike the countenance of an ANGEL of God, very terrible: bu
.5.19' type=verse>But the ANGEL of the Lord by night open
s land; he shall send his ANGEL before thee, and thou sha
' type=verse>And when the ANGEL which spake unto Corneliu
an: but if a spirit or an ANGEL hath spoken to him, let u
rse>And there appeared an ANGEL unto him from heaven, str
ype=verse>And I heard the ANGEL of the waters say, Thou a
 type=verse>And the sixth ANGEL sounded, and I heard a vo
 type=verse>And the third ANGEL poured out his vial upon
```

Solche Konkordanzen sind ausserordentlich nützlich im Erkennen von wiederkehrenden Mustern. Die Suche nach *death* bringt einem beispielsweise schnell auf die häufigen Formulierungen *shadow of death*, *put to death* oder *worthy of death*. Zur Veranschaulichung folgt ein Auszug aus der Konkordanz zu *death*, nach dem Kontext links sortiert (gleiches Corpus):

```
f Israel should be put to DEATH, whether small or great,
wner also shall be put to DEATH. </seg> <seg id='b.EXO.21
 Adonijah shall be put to DEATH this day. </seg> <seg id=
they shall condemn him to DEATH, and shall deliver him to
ourge him, and put him to DEATH: and the third day he sha
l they cause to be put to DEATH. </seg> <seg id='b.LUK.21
 both into prison, and to DEATH. </seg> <seg id='b.LUK.22
together, and were put to DEATH in the days of harvest, i
house, he shall be put to DEATH: but be ye with the king
led with him to be put to DEATH. </seg> <seg id='b.LUK.23
r, shall be surely put to DEATH. </seg> <seg id='b.EXO.21
l there any man be put to DEATH this day in Israel? for d
hall not Shimei be put to DEATH for this, because he curs
```

Manche Konkordanzprogramme haben eine Spezialfunktion zur Berechnung von Kollokationen. Für obiges Beispiel bekäme man eine Aufstellung geliefert, in der man genau sieht, wie oft im Corpus *put to* links von *death* steht oder wie oft *prevented* rechts davon steht.
Für den Übersetzer eröffnen sich hier grosse Möglichkeiten: Ist er etwa unsicher, ob man im biblischen Kontext *Rache* eher als *vengeance* oder *revenge* übersetzt, bringt ihn eine kurze Suche mit der Konkordanzsoftware in kürzester Zeit auf die richtige Spur. *Vengeance* kommt im Corpus 45 mal vor, *revenge* nur 5 mal. Sehr wahrscheinlich ist also *vengeance* die bessere Wahl.

Konkordanzen bringen dem Übersetzer also genau das, was ihm in einem herkömmlichen Wörterbuch fehlt: Informationen über die *tatsächliche Verwendung* von bestimmten Wörtern. Erst der Kontext und die Häufigkeit des Vorkommens eines Wortes kann manchmal die Wahl der richtigen Übersetzung ermöglichen. Ein Wörterbuch kann zu einem Lemma lediglich verschiedene Bedeutungen aneinanderreihen. Erst der Blick in die tatsächlichen Verwendungen der einzelnen Bedeutungen mithilfe einer Konkordanzsoftware erlauben es einem in einem solchen Fall, aus der Reihe von Bedeutungen die passende auszuwählen.

2.2.2 Wortlisten

Wortlisten können alphabetisch von hinten oder von vorne sortiert erstellt werden:

Beispiel 1

```
  accursed - 20
accusation - 10
    accuse - 16
   accused - 14
   accuser - 01
  accusers - 08
  accuseth - 01
  accusing - 01
```

Beispiel 2

```
  incurable - 06
    durable - 02
 honourable - 30
 favourable - 04
inexcusable - 01
      table - 73
 delectable - 01
```

Die alphabetische Anordung zeigt verwandte Wörter und ihre Häufigkeiten an. Im Beispiel 1 sehen wir, abgesehen von *accursed*, alle Wörter zur Wortfamilie *accuse*. Wir sehen auch gleich, dass die altertümliche Form mit der Endung *-th* in der 3.Pers.Sing. nur einmal vorkommet, genauso wie die *ing*-Form oder das Nomen Agentis *accuser*. Die ambigue Form *accuse* ist hingegen mit 16 Einträgen viel häufiger.
Die Rückwärtssortierung offenbart Wörter, die das selbe Derivationssuffix aufweisen. Im Beispiel haben wir es mit dem aus dem Französischen entlehnten, adjektivbildenden Suffix *-able* zu tun. Man muss aber auf Interferenzen aufpassen: Das mit 73 Einträgen

häufigste Wort *table* gehört natürlich nicht dazu.

2.2.3 Worthäufigkeiten

Eine weitere Informationsquelle sind Worthäufigkeitslisten. Sie zeigen, wie oft welches Wort im Corpus vorkommt (Immer noch Maryland Corpus und TextSTAT-Tool):

```
   the - 63890
   and - 51669
    of - 34614
    to - 13563
  that - 12902
    in - 12666
    he - 10443
 shall - 09819
  unto - 08988
   for - 08972
     i - 08867
   his - 08467
     a - 08207
  lord - 07952
  they - 07369
    be - 07013
    is - 06995
   him - 06656
```

Schon ein Blick in die ersten 20 Einträge zeigt, um was für eine Textsorte es sich handelt. Zwar sind die absolut häufigsten Wörter Artikel, Konjunktionen, Präpositionen und Pronomen, doch zeigen die Häufigkeiten von *shall* oder *lord*, dass es sich um einen biblischen Text handelt. Noch deutlicher wird das Spezifische an diesem Text, wenn man diese Worthäufigkeitenliste mit einer Liste vergleicht, die über einem breiten, unspezifischen Corpus erstellt worden ist. Dann bekommen wir als Resultat keine absolute Häufigkeit mehr, sondern die relative Häufigkeit der einzelnen Wörter im Vergleich zur durchschnittlichen Häufigkeit in anderen Texten.

2.3 Anwendung in der *Machine Translation* (MT)

In der MT werden einsprachige Corpora für verschiedene Teilschritte verwendet:

Extraktion von Lexika Wie oben gesehen lassen sich Wortlisten leicht aus Corpora gewinnen. Was für ein Wörterbuch noch zu tun bleibt, ist die Reduktion der Wortformen auf Lemmata.

Extraktion von Grammatiken Dazu kann man einsprachige Corpora verwenden, sofern sie hinreichend annotiert sind (vgl. Abschnitt 1.4). Die Grammatikregeln aus einem Corpus abzuleiten ist relativ gut möglich, wenn die Syntaxstruktur des Corpus wie etwa in einer *treebank* in maschinenlesbarer Form aufbewahrt wird.

Training von Taggern Viele Tagger sind im Prinzip sprachunabhängig und lassen sich an einem Corpus für eine bestimmte Sprache mit einem bestimmten Tagset trainieren. Voraussetzung dazu ist ein von Hand getaggtes Corpus, an dem man die Software dann *lernen* lassen kann.

3 Parallele Corpora

3.1 Einleitung

Unter einem parallelen Corpus versteht man normalerweise ein Corpus, das Texte in einer Originalsprache und Übersetzungen des Textes in eine/mehrere andere Sprachen enthält. Tatsächlich gibt es ja viele Texte, die sowieso in andere Sprachen übersetzt werden, wie beispielsweise Gebrauchsanleitungen, technische Spezifikationen für Geräte und Maschinen, internationale Zeitung, Regierungsdokumente oder auch literarische Texte. Wenn man nun diese Texte in einem zweisprachigen Corpus zusammenstellt, erhält man ein Hilfsmittel, das sich hervorragend für die Unterstützung der maschinellen Übersetzung eignet.
Neben parallelen Corpora gibt es auch noch die sogenannten vergleichbaren Corpora. Die sind aber weniger nützlich für MT, da es sich nicht um Texte und ihre direkte Übersetzung handelt. Vielmehr geht es dabei um Corpora, die ähnliche Texte in verschiedenen Sprachen enthalten. Ähnlich heisst in diesem Fall: gleiche/ähnliche Textsorte, gleiche/ähnliche Fachsprache, gleiches/ähnliches Thema. Diese sind vor allem in der CAT bzw. der menschlichen Übersetzung nützlich.

3.2 Alignment

Eine wesentliche Aufgabe in der Erstellung eines parallelen Corpus ist der Prozess der genauen parallelen Anordnung der einzelnen Textfragmente im Corpus. Denn ein paralleles Corpus nützt ja wenig, wenn man nicht weiss, welche Textstücke die Übersetzungen von welchen anderen Textstücken sind. Diese parallele Anordnung nennt man *alignment* und sie kann mit statistischen Methoden ziemlich gut erreicht werden (vgl. VÉRONIS 2000). Die *alignment-tools* ordnen meistens auf der Basis der Anzahl Zeichen pro Satz zu. Es werden auch die Anzahl Sätze bzw. Paragraphen pro Text, die Interpunktion und anderes verwendet. Es gibt auch Ansätze, die Resultate eines POS-Taggers für das *alignment* zu benutzen (BORIN 2002).

3.3 Anwendung in der *Computer Aided Translation* (CAT)

Konkordanzsoftware, die mit parallelen Corpora umgehen kann, kann die Arbeit eines Übersetzes stark unterstützen. Stellen wir uns einmal einen Übersetzer vor, der die Dokumentation einer Software von Englisch nach Schwedisch übersetzen soll. Er nimmt dazu ein Corpus wie dasjenige von OPUS (TIEDEMANN 2004) zur Hand, welches aus frei erhältlicher Dokumentation von verschiedenen Softwareanwendungen kompiliert wurde. Solche Dokumentationen sind ja problemlos in verschiedenen Übersetzungen zu haben.
Wenn der Übersetzer jetzt beispielsweise unsicher ist, wie man *keyboard* ins Schwedische zu übersetzen hat, und er dafür in ein Wörterbuch schaut, bekommt er *klaviatur* oder *tangentbord* zur Auswahl. Doch welches ist in diesem Fall die bessere Wahl? Ein

Blick ins Corpus kann ihm weiterhelfen:

```
For example , if you want to load personal key settings
from a file , go to the Keyboard tab.
Om Du t ex vill ladda en egen tangentkombination från en
fil , måste Du byta till fliken Tangentbord.
-----------------------------------------------------------
Click the Save button on the Keyboard tab Klicka i så
fall på kommandoknappen Spara på fliken Tangentbord.
-----------------------------------------------------------
Later , after reinstalling %PRODUCTNAME on a new host,
you can reload the key configuration ( Load button on
the Keyboard tab ).
Senare , när Du t ex har gjort en komplett
nyinstallation av %PRODUCTNAME på en ny dator , kan
du åter ladda tangentkonfigurationen ( kommandoknappen
Ladda på fliken Tangentbord ).
-----------------------------------------------------------
Keyboard
Tangentbord
-----------------------------------------------------------
On this tab page you can customize the keyboard
configuration to your own preferences.
Under den här fliken kan du anpassa tangentbordet efter
dina behov.
-----------------------------------------------------------
Enter text using the keyboard.
Mata in text via tangentbordet.
-----------------------------------------------------------
...
```

(Aus dem Onlineinterface des OPUS Corpus, Suchterm *keyboard*)

Der Befund zeigt eindeutig, dass *tangentbord* hier die richtige Wahl ist. Man sieht auch sehr schön, wie entscheidend die richtige Alignierung der Sätze ist. Wären hier Satzpaare aufgelistet worden, die gar nicht ihre gegenseitige Übersetzung sind, wäre es nicht möglich, die richtige Übersetzung für *keyboard* zu finden.
Ein zweites Beispiel betrifft die Verwendung einer Übersetzung im richtigen Kontext. Oft findet man ja die richtige Übersetzung eines Wortes oder einer Phrase schnell, doch ist das Wissen, wie man diese Übersetzung nun passend in einen Satz einbindet, noch einmal ein zusätzliches Stück Information, das man in einem Wörterbuch normalerweise nicht findet. Nehmen wir das Beispiel *template*. Dieses Wort wird üblicherweise als *dokumentmallar* übersetzt wird. Doch welche Verben verwendet man im Zusammenhang mit *dokumentmallar*? Die Suche nach *template* im Corpus ergibt folgende Satzpaare:

```
A template contains the following :
En dokumentmall innehåller följande delar :
-----------------------------------------------------------
Opens the selected template for editing.
Öppnar den valda mallen för redigering.
-----------------------------------------------------------
If you used a template , the name of the template is
displayed here.
Om du har använt en dokumentmall visas dokumentmallens
namn här.
-----------------------------------------------------------
Click here to load the selected template for editing.
Med det här kommandot laddar du dokumentmallen som är
markerad i listrutan.
-----------------------------------------------------------
If you want to import a template , select the folder
in the list box in which the imported template is to be
stored , and then select Import template.
Om du vill importera en dokumentmall , väljer du mappen
i den listruta där den importerade dokumentmallen ska
placeras . Sedan väljer du Importera mall.
-----------------------------------------------------------
Export Template
Exportera mall
-----------------------------------------------------------
Click this button to load the template selected in the
template area for editing .
Om du klickar på den här kommandoknappen laddas
dokumentmallen som är markerad i mallområdet för
redigering
-----------------------------------------------------------
...
```

(Aus dem Onlineinterface des OPUS Corpus, Suchterm *template*)

Gute Formulierungen sind folglich *öppnar mallen* 'open the template', *använda en dokumentmall* 'use a template', *laddar dokumentmallen* 'load the template', *importera en dokumentmall* 'import a template' oder *exportera mall* 'export template'.
Auch Fragen des Stils eignen sich für solche Untersuchungen. Mit einer Suche nach *I* oder *we* kann man schnell herausfinden, ob in einer bestimmten Textsorte, wie etwa wissenschaftlichen Untersuchungen, die Nennung der Autorschaft mit einem Personalpronomen üblich ist[3]. Eine andere Stilfrage wäre diejenige nach der üblichen Satzlänge in der entsprechenden Textsorte. Softwaredokumentationen tendieren eher zu kurzen Sätzen, während Rechtstexte besonders lange Sätze favorisieren. Auch solche Fragen lassen sich mit einem Corpus leicht beantworten.

[3] BOWKER/PEARSON 2002:197 tun genau das und kommen zum Schluss, das die Nennung des Personalpronomens der 1. Person Singular oder Plural vermieden wird.

3.4 Anwendung in der *Machine Translation* (MT)

Die maschinelle Übersetzung kann man grob in die drei Hauprichtungen *Rule-Based Machine Translation* (RBMT), *Statistical Translation* (SMT) und *Example-Based Translation* (EBMT) einteilen. Parallele Corpora finden in allen drei Gebieten grosse Verwendung:

Rule-Based Machine Translation ...bezeichnet die Übersetzung nach Grammatikregeln. In diesem Gebiet werden parallele Corpora hauptsächlich verwendet, um Übersetzungslexika zu gewinnen (TIEDEMANN 1998) oder Grammatiken aus syntaktisch annotierten Corpora zu extrahieren.

Statistical Translation Statistische Übersetzungsprogramme lernen normalerweise das Übersetzungsmodell von einem parallelen Corpus. Das Einzelsprachenmodell hingegen lernen sie von einem einsprachigen Corpus. Corpora sind also unter Umständen die exklusive Ressource eines SMT-Systems.

Example-Based Translation Solche Systeme verlassen sich ausschliesslich auf zweisprachige Corpora, um Textpassagen übersetzen zu können. Der Grundgedanke dabei ist, das in parallelen Corpora zwar nicht die gesamten Sätze aber doch immerhin die Textfragmente, die zu übersetzen sind, bereits schon mit einer Übersetzung vorhanden sind. Man muss nur die genaue Entsprechung dafür finden, und die erhaltenen Fragmente dann wieder richtig zusammensetzen. Die Schritte beim EBMT lauten also (1) Beispiele im Corpus finden (*matching*), (2) die entspechende Übersetzung in der anderen Sprache lokalisieren und (3) die Textstücke wieder zu einem sinnvollen Satz zusammenbauen.(SOMERS 2003:7). Ein solches System kommt, wie auch manche statistischen Ansätze, gänzlich ohne Grammatik aus.

4 Praxis

4.1 Verfügbarkeit von Corpora

Verschiedene Gründe machen die Verfügbarkeit von Corpora schwierig. Zunächst gibt es oft Urheberrechtsprobleme, weil Autoren zwar ihre Texte für wissenschaftliche Untersuchungen freigeben, aber die freie und unkontrollierte Verbreitung übers Internet verhindern wollen. Darum sind Corpora oft nur für die Mitglieder eines bestimmten Projektes, das die Erstellung des Corpus veranlasst hat, verfügbar.
Viele Corpora kann man kommerziell erwerben, was man im Lichte des Aufwandes, der hinter manchen Corpora steckt, sehen muss. Trotzdem ist dies für die Wissenschaft hinderlich.
Es gibt auch Projekte, die ihre Textsammlungen nicht herausgeben, aber immerhin ein Online-Interface anbieten, wo man Anfragen stellen kann. Leider eignen sich diese Interfaces aber jeweils nur für die Benutzung durch Menschen, nicht aber durch Maschinen.

4.2 Ein frei erhältliches Corpus: OPUS

Das OPUS Corpus (TIEDEMANN 2004) besteht aus problemlos im Internet erhältlichen Dokumentationen von verschiedener Software. Es handelt sich bei dieser Software um das Office-Paket OpenOffice, die Skriptsprache PHP und den UNIX Desktop KDE. Die Texte sind in den Sprachen Englisch, Französisch, Spanisch, Schwedisch, Deutsch und Japanisch vorhanden, wobei nicht von jeder Sprache sämtliche Texte enthalten sind. Alle Sprachen zusammengerechnet ergeben eine Corpusgrösse von 2,612,144 Wörtern. Die Codierung des Corpus ist in XML vorgenommen, wobei die Standards von CES und TEI - aus welchen Gründen auch immer - nicht beachtet wurden. Die Gliederung eines solchen XML Files ist aber auch ohne genaue Entsprechung zur Norm gut zu lesen: es gibt ein `<head>`-Tag, das Titel, Sprache und Filename der Datei beschreiben, und dann mit `<p>` abgetrennte Paragraphen mit dem eigentlichen Textinhalt. Die Sätze im Corpus sind morphologisch analysiert. Dabei befindet sich jede einzelne Wortform in einem Tag `<w>`, was für *word* steht, und bekommt die Wortenartenbestimmung, welche duch den POS-Tagger vorgenommen wurde, sowie das Lemma, auf welches die Form zurückgeführt wurde, als Attribute mitgeliefert.

4.2.1 Eine Beispieldatei

Hier folgt der Anfang der deutschsprachigen Datei main0108.xml aus der OpenOffice Dokumentation. Abgebildet sind der Header und der erste Satz 'Das Hilfemenü dient zum Aufruf und Steuern des Hilfesystems von %PRODUCTNAME':

```
<?xml version='1.0' encoding='utf-8'?>
<document>
  <head>
    <title>Hilfe</title>
    <meta name='filename' content='text/common/main0108' />
    <meta name='language' content='de-DE' />
    <help:css-file-link xmlns:help=
    'http://openoffice.org/2000/help' />
  </head>
  <help:to-be-embedded Eid='hilfe' xmlns:help=
  'http://openoffice.org/2000/help'>
    <p class='Head1' id='1'>
      <help:help-id value='SID_HELPMENU' />
      <help:link Id='65587'>
        <s id='s1.1'>
          <w tree='NN' lem='Hilfe' tnt='NN'>Hilfe</w>
        </s>
      </help:link>
    </p>
    <p class='Paragraph' id='2'>
      <help:help-text value='visible'>
        <s id='s2.1'>
          <w tree='ART' lem='d' tnt='ART'>Das</w>
          <w tree='NN' tnt='NE'>Hilfemen\"u</w>
          <w tree='VVFIN' lem='dienen' tnt='VVFIN'>dient</w>
```

```
        <w tree='APPRART' lem='zum' tnt='APPRART'>zum</w>
        <w tree='NN' lem='Aufruf' tnt='NN'>Aufruf</w>
        <w tree='KON' lem='und' tnt='KON'>und</w>
        <w tree='NN' lem='Steuer|Steuern' tnt='NN'>Steuern</w>
        <w tree='ART' lem='d' tnt='ART'>des</w>
        <w tree='NN' tnt='NN'>Hilfesystems</w>
        <w tree='APPR' lem='von' tnt='APPR'>von</w>
        <help:productname>
          <w tree='NN' tnt='NE'>%PRODUCTNAME</w>
        </help:productname>
        <w tree='$.' lem='.' tnt='$.'>.</w>
      </s>
    </help:help-text>
  </p>
```

4.2.2 Vorteile von OPUS

- OPUS ist frei verfügbar und kann von der Homepage des Projektes heruntergeladen werden.
- Es sind 6 Sprachen vertreten.
- OPUS hat eine beträchtliche Grösse.
- Die Fachsprache 'Softwaredokumentation' kann ein Vorteil sein, wenn man das Corpus ebenfalls für diese Fachsprache einsetzen will.
- Es gibt ein komfortables Online-Interface auf der Homepage.

4.2.3 Nachteile von OPUS

- Die Fachsprache 'Softwaredokumentation' limiert aber die Anwendung stark.
- OPUS hält sich nicht an die Standards von CES oder TEI.
- Es gibt keine manuelle Überarbeitung der Resultate des Taggings und der morphologischen Analyse.

4.3 Ein frei erhältliches Corpus-Tool: TextSTAT

Das TextSTAT Programm (HÜNING 2004) ist ein frei verfügbares Tool für den Umgang mit elektronischen Corpora. Es wurde von Matthias Hüning an der Abteilung für Niederlandistik der FU Berlin entwickelt und erlaubt die Erstellung von Konkordanzen, Wortlisten und Häufigkeitslisten. Desweiteren unterstützt es die automatische Erstellung von Corpora aus Internet-Seiten.

4.3.1 Vorteile von TextSTAT

- TextSTAT ist frei verfügbar.
- Die Software bietet Konkordanzen, Wortlisten und Worthäufigkeitslisten.
- Sie erlaubt die Erstellung eigener Corpora.
- Die Resultate sind nach verschiedenen Kriterien sortierbar.
- Die Plattformunabhängigkeit ist dank Programmierung in Python garantiert.

4.3.2 Nachteile von TextSTAT

- Das Tool ist nicht sehr schnell.
- Es kann keine parallelen Corpora verarbeiten.

5 Schlusswort

Die Relevanz von elektronischen Textdaten für MT und CAT ist also beträchtlich, und wird in der Zukunft eher noch ansteigen. Es bleibt zu hoffen, dass sich die Anzahl und Verfügbarkeit von Corpora in nächster Zeit deutlich verbessern wird. Denn wer sich heute mit MT oder CAT beschäftigt, ist auf gute, möglichst standardisierte Corpora angewiesen.

Literatur

[1] Bowker, Lynne and Pearson, Jennifer: *Working with Specialized Language: A practical guide to using corpora*. London/New York 2002.

[2] Ebeling, Jarle: *The Translation Corpus Explorer: A browser for parallel texts*. In: Johansson, Stig und Oksefjell, Signe (eds.): *Corpora and Cross-linguistic Research*. Amsterdam 1998.

[3] Véronis, Jean (ed.): *Parallel Text Processing: Alignment and Use of Translation Corpora*. Text, Speech and Language Technology Volume 13. Dordrecht/Boston/London 2000.

[4] Borin, Lars: *Alignment and tagging*. In: Borin, Lars (ed.): *Parallel corpora, parallel worlds*. Amsterdam/New York 2002.

[5] Burnard, Lou: *TEI Lite: An Introduction to Text Encoding for Interchange*. 2002. Von: <http://www.tei-c.org/Lite/>.

[6] Carl, Michael and Way, Andy: *Recent Advances in Example-Based Machine Translation*(Introduction). Text, Speech and Language Technology Volume 21. Dordrecht/Boston/London 2003.

[7] Somers, Harold: *An Overview of EBMT*. In: Carl, Michael/Way, Andy (ed.): *Recent Advances in Example-Based Machine Translation*. Text, Speech and Language Technology Volume 21. Dordrecht/Boston/London 2003.

[8] Tiedemann, Jörg: *Extraction of translation equivalents from parallel corpora.* In: *Proceedings of the 11th Nordic conference on computational linguistics, Copenhagen 28-29 January 1998.* Copenhagen 1998.

Webseiten:

[9] Feldweg, Helmut: *Die Wortformen der geschlossenen Wortarten im Stuttgart-Tübingen Tagset (STTS).* Geändert am 23. Juli 1996. <http://www.sfs.nphil.uni-tuebingen.de/Elwis/stts/stts.html>.

[10] Hüning, Matthias: *TextSTAT - Konkordanz-Software für Windows und Linux.* Geändert am 24. August 2004. <http://www.niederlandistik.fu-berlin.de/textstat/>.

[11] Ide, Nancy: *Corpus Encoding Standard.* 20. Mai 2000. <http://www.cs.vassar.edu/CES/>.

[12] Leech, Jeffrey und Smith, Nicholas: *BNC2: POS Tagging error rates.* 17. März 2000. <http://www.comp.lancs.ac.uk/ucrel/bnc2/bnc2error.htm>.

[13] Marcus, Mitchell: *The Penn Treebank Project.* Geändert am 2. Februar 1999. <http://www.cis.upenn.edu/ treebank/>.

[14] Resnik, Philip: *University of Maryland Parallel Corpus Project: The Bible.* 1999. <http://benjamin.umd.edu/parallel/>.

[15] TEI Consortium: *Text Encoding Initative: Welcome to the TEI Site.* 15. August 2001 (geändert 6. September 2003). <http://www.tei-c.org>.

[16] Tiedemann, Jörg: *OPUS - an open source parallel corpus.* Geändert am 8. Juni 2004. <http://logos.uio.no/opus/>.

[17] Trabold, Annette: *COSMAS 1 - Korpusrecherche- und -analysesystem.* Geändert am 26. Juni 2004. <http://www.ids-mannheim.de/kt/projekte/cosmasI/>.

[18] UCREL (University Centre for Computer Corpus Research on Language): *CLAWS part-of-speech tagger for English.* Geändert am 2. August 2004. <http://www.comp.lancs.ac.uk/ucrel/claws/>.