ESTUDIOS LINGÜÍSTICOS

Revista de **Filología Alemana**
ISSN-e: 1988-2823

http://dx.doi.org/10.5209/RFAL.60148

EDICIONES COMPLUTENSE

Das PaGeS-Korpus, ein Parallelkorpus der deutschen und spanischen Gegenwartssprache[1]

Irene Doval[2]

Recibido: 18 de enero de 2017 / Aceptado: 15 de marzo de 2017

Zusammenfassung. Das Korpus-PaGeS ist ein zweisprachiges Parallelkorpus, das aus einer Sammlung von spanischen und deutschen Texten der Gegenwartssprache besteht. Der Aufsatz beschreibt die einzelnen Arbeitsphasen in der Erstellung des Korpus. Die Beschreibung umfasst die manuelle Vorverarbeitung, die linguistische Aufbereitung und das automatische und manuelle Verfahren für die Alignierung der Texte. Es wird auf den Zugriff und die Visualisierung der Daten eingegangen und die verschiedenen Suchmöglichkeiten werden erläutert. Abschließend werden die geplanten nächsten Schritte skizziert.
Schlüsselwörter: Parallelkorpora; Kontrastive Linguistik; Übersetzung; Korpuslinguistik; Computerlinguistik; Deutsche Sprache; Spanische Sprache.

[en] The PaGeS Corpus, a Parallel Corpus of the Contemporary German and Spanish Language

Abstract. The corpus PaGeS is a bilingual parallel corpus, that comprises a collection of contemporary Spanish and German texts. This paper describes the different steps in the construction of the corpus. The description includes the manual preparation process of the texts to make the documents suitable for further processing, the linguistic annotation and the manual and automatic procedure of the sentence alignment of the texts. It is dealt with the access and the visualization of the data and the different search possibilities are explained. Finally, the next steps of future work are outlined.
Keywords: Parallel Corpus; Contrastive Linguistics; Translation; Corpus Linguistics; Computer Linguistics; German Language; Spanish Language.

[es] El corpus PaGeS, un corpus paralelo de textos alemanes y españoles contemporáneos

Resumen. El corpus PaGeS es un corpus bilingüe paralelo que incluye una colección de textos contemporáneos alemanes y españoles. Este artículo describe las sucesivas fases en la elaboración del corpus. Esta descripción incluye el proceso de preparación de los textos para procesarlos, la anotación lingüística y el procedimiento de alineación automático y manual. Se aborda el tema del acceso y de

---

[1] Der vorliegende Aufsatz geht von einem vom spanischen Ministerio de Ciencia e Innovación (FFI2013-42571-P) finanzierten Projekt aus (s. Fussnote 3).

[2] Universidade de Santiago de Compostela (Spanien)
E-Mail: i.doval@usc.es

la visualización de los resultados, así como las diferentes posibilidades de búsqueda. Finalmente se esbozan los pasos futuros.
Palabras clave: Corpus paralelos; lingüística contrastiva; traducción; lingüística de corpus; lingüística computacional; lengua alemana; lengua española.

Cómo citar: Doval, I.,«Das PaGeS-Korpus, ein Parallelkorpus der deutschen und spanischen Gegenwartssprache», *Revista de Filología Alemana* 26 (2018), 181-197

1. Einleitung. Korpustypologie

Die linguistische Forschung hat zwar immer von Textkorpora als empirischer Basis Gebrauch gemacht, aber erst die Verfügbarkeit von großen Textdatenbanken in elektronischer Form hat der Korpuslinguistik, die sich mit dem Aufbau und der Auswertung von Korpora beschäftigt, zum wirklichen Durchbruch verholfen.

Früher waren die manuelle Korpuserstellung und die manuelle Untersuchung der Daten fehleranfällig, sehr zeitraubend und teuer. Außerdem handelte es sich meist um Belegsammlungen in Form von Sätzen bzw. kurzen Textabschnitten, die zu ihrer Sortierung auf einzelne Zettel übertragen wurden und als Basis zur Erstellung von Wörterbüchern oder Grammatiken dienten.

Die ersten Textkorpora in elektronischer Form sind in den 60er Jahren entstanden. Zu nennen ist hier vor allem das Brown Corpus, mit dessen Erstellung Nelson Francis und Henry Kučera 1961 begannen. Seit der rasanten Verbreitung von Computern in den 90er Jahren sind digitale Textkorpora für linguistische Fragestellungen unumgänglich geworden. Sie haben eine völlig neue Ausgangslage und neue Perspektiven der Forschung im Bereich von Grammatik, Lexikographie, Fremdsprachendidaktik oder Übersetzung u.a. eröffnet. Dabei handelt es sich immer um computerlesbare Korpora. So definieren Lemnitzer / Zinsmeister (2010: 40) ein linguistisches Korpus folgendermaßen:

> Sammlung schriftlicher oder gesprochener Äußerungen in einer oder mehreren Sprachen. Die Daten des Korpus sind digitalisiert, d.h. auf Rechnern gespeichert und maschinenlesbar. Die Bestandteile des Korpus bestehen aus den Daten selber sowie möglicherweise aus Metadaten, die diese Daten beschreiben und aus linguistischen Annotationen, die diesen Daten zugeordnet sind.

Korpora lassen sich nach unterschiedlichen Kriterien klassifizieren. Bezüglich der Sprachenauswahl (Lemnitzer / Zinsmeister 2010: 102ff. Zinsmeister 2010: 6) können Korpora monolingual – sie enthalten nur eine Sprache- oder bi- bzw. multilingual – sie enthalten zwei oder mehrere Sprachen- sein. Bei den letzteren unterscheidet man nach der allgemein anerkannten Terminologie (McEnery / Xiao 2007: 2-3) vergleichbare und parallele Korpora. Vergleichbare Korpora setzen sich aus einsprachigen Texten in verschiedenen Sprachen zusammen, die Thema,Textsorte und Register teilen und eine ähnliche Herkunft und einen ähnlichen Umfang haben

wie bspw. Wetterberichte, Stellenangebote oder Zeitschriftenartikel. Die Texte eines vergleichbaren Korpus sind keine Übersetzungen voneinander, aber sie werden nach gemeinsamen Auswahlkriterien ausgewählt. Parallelkorpora enthalten dagegen Quelltexte in einer Sprache und deren Übersetzung in eine oder mehrere andere Sprachen. Eines der ersten und bekanntesten Parallelkorpora ist das Hansard Korpus (http://www.tsrali3.com/), das aus den Protokollen kanadischer Parlamentsdebatten in Englisch und Französisch besteht. Parallelkorpora können zwei- oder mehrsprachig sein, je nachdem, ob sie Texte aus zwei oder mehr Sprachen beinhalten. Sie können entweder unidirektional, wenn die Originalsprache immer die gleiche ist (z. B. spanische Originaltexte und deutsche Übersetzungstexte), bidirektional (z. B. spanische und deutsche Originaltexte und Übersetzungstexte) oder multidirektional, wenn Originaltexte in mehrere Sprachen übersetzt sind, bspw. ein englischer Text wie etwa eine EU-Verordnung, der in verschiedene Sprachen übersetzt wird. Zum Abschluss dieses Abschnittes ist die oben vorgestellte Typologie noch einmal in Abbildung 1 aufgeführt.

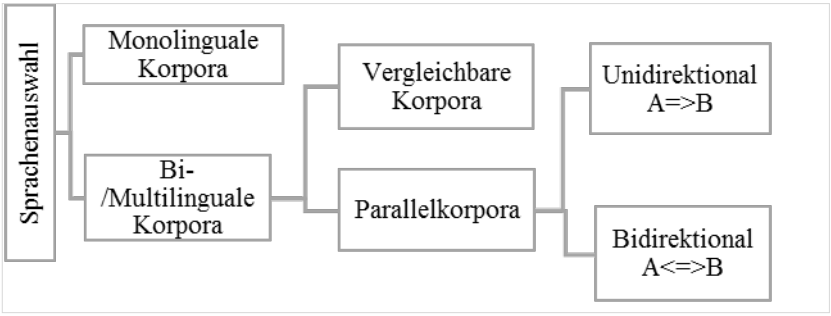


Abbildung 1. Typologie der Parallelkorpora
(in Anlehnung an Lemnitzer / Zinsmeister 2010)

Im nächsten Abschnitt wird ein kurzer Überblick über vorhandene Parallelkorpora gegeben und die Motivation zur Erstellung des PaGeS – Korpus[3] (**Pa**rallel Corpus **Ge**rman / **S**panish) wird aufgezeigt. Nachfolgend werden die Arbeitsschritte in der Konstruktion des Korpus beschrieben: Design und Datenerhebung (3.), Vorverarbeitung der Texte, die Metadaten (4.), die Alignierung (5.) und die linguistische Annotation (6.). Anschließend werden die Suchmöglichkeiten der Daten dargestellt (7.) und schließlich wird eine Zusammenfassung der distinktiven Merkmale des Korpus erstellt (8.).

---

[3] Das Korpus PaGeS (www.corpuspages.eu) wird vom Forschungsteam SpatiAlEs unter der Leitung von Prof. Irene Doval an der Universität Santiago de Compostela erstellt. Es wird im Rahmen des vom Spanischen Ministerium für Wirtschaft und Wettbewerbsfähigkeit geförderten Forschungsprojekts FFI2013-42571-P durchgeführt.

2. Andere Parallelkorpora Deutsch / Spanisch. Motivation für die Erstellung des PaGeS-Korpus

Die Erstellung des Korpus PaGeS ist Teil eines Forschungsprojekts zur Analyse der spatialen Ausdrücke[4] im Spanischen und Deutschen. Im Gegensatz zu den meisten bisherigen kontrastiven Arbeiten, die weitgehend auf einzelsprachlichen Grammatiken sowie auf der Introspektion der Linguisten basieren, wollten wir uns korpuslinguistischer Methoden bedienen, um auf einer breiten empirischen Datenbasis die Analyse durchzuführen. Bestehende Korpora haben sich aber für unser Forschungsvorhaben als wenig geeignet erwiesen und deshalb haben wir mit der Erarbeitung eines eigenen Korpus begonnen, das unseren Anforderungen möglichst gut entspricht. Im Folgenden werden die wichtigsten frei verfügbaren Parallelkorpora kurz vorgestellt und dabei wird auf deren Beschränkungen für unsere Recherche eingegangen.

Die vorhandenen umfangreichen Parallelkorpora, die Spanisch und Deutsch enthalten, sind ausschließlich mehrsprachige Korpora. Die bei Weitem wichtigsten multilingualen Textkorpora sind die Sprachressourcen der Europäischen Union. Steinberg *et al*. (2014: 679ff.) bieten einen vergleichenden Überblick über die Ressourcen, die von der Europäischen Kommission und vom Europäischen Parlament bereitgestellt worden sind. Europarl ist ein Parallelkorpus, das die Protokolle des Europaparlaments seit 1996 in 21 offiziellen EU-Sprachen enthält (http://www.statmt.org/europarl). Die neueste Version (Release v7) umfasst bis zu 50 Millionen Wörter pro Sprache. Die Texte sind auf Satzebene aligniert. Die deutschen Texte enthalten 47.236.849 Wörter und die spanischen Texte 54.806.927 Wörter.

Das Digital Corpus des Europäischen Parlaments (DCEP) (https://ec.europa.eu/jrc/en/language-technologies/dcep) umfasst eine Vielzahl von Dokumenten, die auf der offiziellen Website des Europäischen Parlaments veröffentlicht worden sind: Pressemitteilungen, Tagesordnungen, Plenartagungsprotokolle, Berichte usw. Die aktuelle Version des DCEP enthält 162.141 Dokumente in der Leitsprache Englisch mit insgesamt über 100 Millionen Wörtern. Das Sprachenpaar Deutsch / Spanisch verfügt über 103.016 Dokumente. Die Texte wurden für jedes einzelne Sprachenpaar auf der Satzebene aligniert.

Das JRC-Acquis (https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis) ist ein großes Parallelkorpus bestehend aus Rechtstexten der Europäischen Union. Die Textsammlung wächst stetig und umfasst derzeit 464.000 Texte ab 1950. Die aktuelle Version 3.0 liegt in 22 Sprachen mit durchschnittlich je 23.000 Texten vor.

Besonders erwähnenswert ist OPUS (http://opus.lingfil.uu.se/), eine enorme Sammlung von frei verfügbaren mehrsprachigen Parallelkorpora. Es ist eine wachsende Ressource und bietet auch Werkzeuge für die Verarbeitung von parallelen

---

[4] Spatiale Ausdrücke bringen spatiale Relationen, d.h. Lokalisierungsereignisse zum Ausdruck. In einer spatialen Relation wird ein Objekt (Figur) im Verhältnis zu einem zweiten Referenzobjekt (Grund) lokalisiert. Nimmt die Figur nur einen Eigenort für einen bestimmten Zeitraum ein, handelt es sich um eine statische Relation, z.B. *Das Buch liegt auf dem Tisch*. Findet ein Ortswechsel der Figur statt, handelt es sich um eine dynamische Relation oder ein Bewegungsereignis, z.B. *Er legt das Buch auf den Tisch*. (s. Krause / Doval 2011: 15ff. und Doval 2016: 210-211).

Sprachdaten sowie Schnittstellen zum Durchsuchen. Das am stärksten vertretene Sprachenpaar ist Spanisch-Englisch mit etwa 36 Millionen parallelen Sätzen (Tiedemann 2012: 2214ff.). Die Bereiche, die von OPUS insbesondere abgedeckt werden, sind Gesetzgebungs- und Verwaltungstexte, hauptsächlich aus Institutionen der Europäischen Union. Darüber hinaus gibt es Untertitel, technische Dokumentationen, eine Vielzahl von Zeitungstexten sowie einige andere kleinere Sammlungen aus verschiedenen Online-Quellen.

In den letzten Jahren werden immer mehr multilinguale Textsammlungen online verfügbar gemacht, die durch ein Computerprogramm, einen sogenannten Crawler[5], der automatisch das Web nach zweisprachigen Webseiten durchsucht, zusammengestellt werden. Das erfolgsreichste Beispiel dafür ist Linguee, „eine einzigartige Kombination aus einem redaktionellen Wörterbuch und einer Suchmaschine“ (http://www.linguee.de/deutsch-englisch/page/ about.php). Die Suchergebnisse werden entsprechend in zwei Teile unterteilt. Oben werden übersichtliche Vokabeltreffer angezeigt, wie in einem Online-Wörterbuch, weiter unten eine große Anzahl von zweisprachigen Satzbeispielen aus Internetquellen, die durch einen eigenen Algorithmus auf Qualität geprüft worden sind. Die meisten Texte stammen jedoch aus dem administrativen, ökonomischen bzw. kommerziellen Bereich. Nach eigenen Angaben werden etwa eine Milliarde übersetzte Satzbeispiele in 25 Sprachen in Linguee übernommen.

Hier soll noch einmal das primäre Ziel unseres Vorhabens in Erinnerung gerufen werden. Es geht darum, die Ausdrücke der Lokalisierungsreignisse sowie deren semantische und syntaktische Eigenschaften in Deutsch und Spanisch zu analysieren. Nach einigen Stichproben zeigte sich schon, dass die vorgenannten Korpora für unser Forschungsziel starken Beschränkungen unterliegen.

Einerseits decken die genannten Parallelkorpora mit Spanisch und Deutsch nur einige Spezialgebiete ab, hauptsächlich administrative und kommerzielle Fachsprachen, in denen Belege für Bewegungsausdrücke nicht in einer ausreichenden Zahl vorkommen.

Andererseits lässt sich in den genannten Korpora die Richtung der Übersetzung nicht immer eindeutig bestimmen, denn es ist nicht immer klar, welcher Text der Quelltext ist und welcher die Übersetzung. Für viele Texte der Europäischen Union ist davon auszugehen, dass sie oft in englischer Sprache verfasst oder aus der Originalsprache ins Englische übersetzt wurden und dann aus dem Englischen in andere Sprachen übersetzt wurden. Im Fall von Textsammlungen aus dem Internet wie Linguee ist die Bestimmung der Originalsprache noch problematischer. Man könnte annehmen, dass die Texte zuerst in der Landesssprache der Webseite geschrieben und dann in verschiedene Sprachen übersetzt worden sind, aber es fehlt offenbar jeder Beweis, der diese Vermutung bestätigt. Es kann davon ausgegangen werden, dass in den vorhandenen mehrsprachigen Korpora eine direkte Übersetzung zwischen Deutsch und Spanisch wahrscheinlich etwas ungewöhnlich ist und dass die meisten Übersetzungen indirekt über eine dritte Sprache erfolgen.

---

[5] Die Bezeichnung Crawler leitet sich von der Suchmaschine WebCrawler ab, die 1994 als erste öffentlich verfügbare Suchmaschine mit Volltextindex-Suche arbeitete. Durch den Einsatz von Crawlern können Suchmaschinen neue Websites hinzufügen, nicht mehr vorhandene Seiten löschen und geänderte Seiten aktualisieren. (s. www.gruenderszene.de/lexikon/begriffe/crawler).

Dazu kommt noch, dass zahlreiche Texte (Originale und Übersetzungen), die in den vorgestellten Korpora aufgenommen wurden, keiner Qualitätskontrolle unterzogen worden sind. Öfter werden die Texte nicht von Muttersprachlern verfasst. Doch gerade dies ist eines der Hauptanliegen unserer Forschung: eine fundierte Datengrundlage mit einer gesicherten Qualität von Original- und übersetzten Texten als empirische Basis zu schaffen.

3. Korpus-Design und Datenbeschaffung

Ein Korpus ist nicht nur eine Sammlung von elektronischen Texten, sondern die Texte müssen nach bestimmten Kriterien gesammelt werden, die gewährleisten sollen, dass das Korpus für das geplante Forschungsziel geeignet ist. In diesem Abschnitt werde ich die Sprachdaten des PaGeS-Korpus bezüglich Größe, Textsorten, Sprachvarietäten, Anzahl und Erscheinungsdatum der Texte beschreiben.

Das PaGeS-Korpus ist als zweisprachiges Korpus in Deutsch und Spanisch konzipiert worden, obwohl die Möglichkeit einer weiteren mehrsprachigen Erweiterung des Korpus nicht ausgeschlossen ist. Von Anfang an haben wir auf höchste Textqualität der Originale und der Übersetzungen großen Wert gelegt. Die einzige Möglichkeit, die Qualität zu gewährleisten, besteht darin, schriftliche Texte von angesehenen Verlagen zu verwenden, bei denen sowohl Originaltexte als auch Übersetzungen einer anspruchsvollen Qualitätskontrolle unterzogen werden.

Das Korpus besteht aus nach 1960 erschienenen Büchern, mit besonderem Schwerpunkt auf Werken aus den letzten zwei Jahrzehnten. Dabei sind fiktionale Werke und Sachbuchtexte enthalten, entweder als gesamte Werke oder als Textausschnitte. Fiktionale Texte bilden die große Mehrheit der Sprachdaten (80 %), da sie eher in andere Sprachen übersetzt werden und damit den größten Teil der verfügbaren Ressourcen darstellen. Bei der Auswahl der Originalwerke wurde für eine bestimmte dialektale Vielfalt gesorgt. So enthält das Korpus neben Büchern von deutschen und spanischen Schriftstellern auch andere von amerikanischen, österreichischen oder Schweizer Autoren. Unter ihnen sind berühmte Autoren wie García Márquez oder Vargas Llosa bis hin zu jungen Schriftstellern wie Joël Dicker.[6]

Geplant ist die Aufnahme von etwa 280 Werken (140 pro Sprache) mit einer Gesamtgröße von ca. 25 Millionen Textwörtern[7], von denen 86 Prozent annährend gleichmäßig auf deutsche und spanische Originaltexte verteilt sind (s. unten Abb. 3). Somit ist das Korpus bezüglich der Übersetzungsrichtung ausgeglichen. Den restlichen Anteil von 14 Prozent bilden Werke, die aus einer dritten Sprache ins Deutsche und Spanische übersetzt wurden (Abb. 3: GX und SX). So werden nicht nur Originale und Übersetzungen, sondern auch zwei Übersetzungen parallelisiert, wie Abbildung 2 zeigt. Die schwarzen Pfeile geben die Richtung der Übersetzung an und die Pfeile mit Doppelspitze die parallelisierten Texte.

---

[6] Für eine vollständige Liste der Autoren und Werke siehe www.corpuspages.eu/.

[7] In Anlehnung an die DWDS-Korpora (www.dwds.de/d/korpora) benutze ich hier den Begriff ‚Textwort' für jedes Vorkommen eines Wortes in einem fortlaufenden Text. Als alternative Bezeichnungen findet man auch ‚Token', ‚Wortvorkommen' oder ‚laufendes Wort'. (S. Störrer 2013: 219)

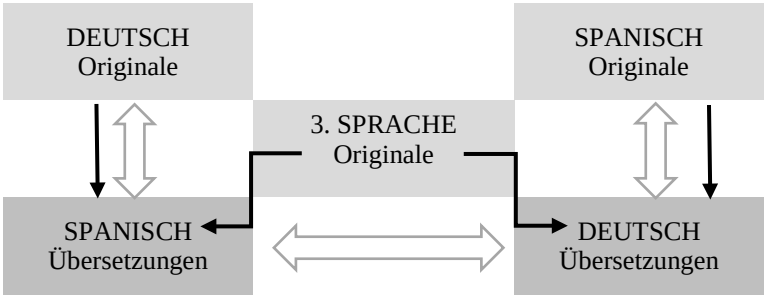
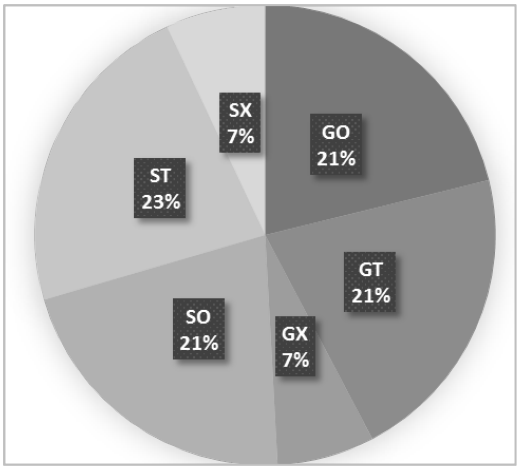


Abbildung 2. Übersetzungsrichtung und parallelisierte Texte

Tabelle 1 gibt einen Überblick über die aktuelle Zusammensetzung des Korpus und Abbildung 3 zeigt den Anteil der Textwörter nach Sprachen sortiert.

Tabelle 1: Anzahl der Werke und Textwörter nach Sprachen und Originaltexten (Stand: November 2016)

Sprache	**Werke**	**Textwörter**
Deutsch Orig.(GO)	54	3965930
Deutsch Übers aus dem Spanischen (GS)	40	3961328
Deutsch Übers. aus einer 3. Sprache (GX)	10	1282368
Deutsch gesamt	104	9209626
Spanisch Orig. (SO)	40	3997581
Spanisch Übers. Deutsch (SG)	54	4222008
Spanisch Übers. aus einer 3. Sprache (SX)	10	1312162
Spanisch gesamt	104	9531751



GO	Deutsch Orig.
GT	Deutsch Übers. <Spanisch
GX	Deutsch Übers. <3. Sprache
SO	Spanisch Orig.
ST	Spanisch Übers. <Deutsch
SX	Spanisch Übers. <3. Sprache

Abbildung 3: Anteil der Textwörter nach Sprachen und Originaltexten

4. Textvorverarbeitung und Metadaten

Die folgenden Abschnitte beschreiben die Arbeitsschritte, die wir für die Erstellung des PaGeS-Korpus ausgeführt haben. Nachdem die Texte ausgewählt und digitalisiert wurden, falls sie nicht schon in elektronischer Form vorlagen, müssen diese einem manuellen Prozess unterzogen werden, um sie für die Alignierung vorzubereiten. Dies besteht im Wesentlichen darin, so viel Parallelität wie möglich zwischen Quell- und Zieltext zu erzielen, um die besten Ergebnisse bei der Alignierung zu erhalten. Dies beinhaltet drei Aufgaben: (a) das Entfernen von nicht korrespondierenden Textausschnitten, fehlerhaften Zeichen und Bildern, (b) Korrekturlesen und (c) Markieren und Annotieren von Metadaten.

Zum einen werden alle Texte, die nicht Teil des Textkörpers sind, entfernt, wie bibliographische Informationen, Seitenköpfe, Widmungen und Anmerkungen des Autors oder des Übersetzers. Auch jeder Anhang ohne Entsprechung in der anderen Version wird gelöscht.

Zum anderen werden beide Versionen (Original und Übersetzung) auf Fehler überprüft, die mit dem Digitalisierungsprozess verbunden sind. Typische Fehler sind dabei unter anderen das Einfügen eines Leerzeichens innerhalb eines Wortes, das Löschen von Leerzeichen zwischen Wörtern oder die gelegentliche Verwechslung bestimmter Zeichen.

In der eingangs erwähnten Definition wurden drei Bestandteile eines Korpus genannt: die Sprachdaten, die linguistischen Annotationen und die Metadaten, die die Sprachdaten beschreiben. Die Metadaten werden verwendet, um relevante Informationen über die Texte zu erfassen und sie aus dem Korpus abrufen zu können. Jeder der im PaGeS-Korpus enthaltenen Werke wird mit einer Metadatenliste versehen, die unter anderen folgende Informationen enthält: Autor und Übersetzer, Titel, Erscheinungsjahr und andere bibliographische Information, Originalsprache und Versionssprache, Genre, manuellen Prüfer sowie Informationen zu den Grundstatistiken (Anzahl der Zeichen, der Textwörter und der Bisegmente) der Dokumente. Diese zusätzlichen Metadaten-Tags werden an die einzelnen Textdateien angehängt und lokal zusammen mit jedem Textdokument gespeichert.

Die Markierung für die Aufteilung der Bücher (wie Teile, Kapitel oder Unterkapitel) wird manuell eingefügt und dadurch wird die Lokalisierung der Textfolge erleichtert. Diese internen Aufteilungen sind im Gegensatz zu den Seitennummern in jeder Ausgabe und im Quell- und Zieltext immer konstant.

Nach dem Korrekturlesen und Markieren werden die Texte in einem gemeinsamen Kodierungsschema UTF-8 als Textdateien gespeichert.

5. Segmentierung und Alignierung

Ein entscheidender Schritt bei der Konstruktion und Nutzung eines Parallelkorpus ist die Alignierung. Tiedemann (2011: 123) definiert Alignierung „as a process of making symmetric correspondences explicit in order to enable further processing of parallel resources“. Durch die Alignierung werden Segmente der Übersetzung den entsprechenden Segmenten des Originals zugeordnet. Die Segmente können abhängig von der vorherigen Segmentierung Absätze, Sätze oder Wörter darstellen

und die nachfolgende Alignierung erfolgt entsprechend auf Paragraph-, Satz- oder Wortebene. Derzeit ist das PaGeS-Korpus auf Satzebene aligniert.

Bei diesem Alignierungsprozess werden zwei Aufgaben kombiniert: Zuerst erfolgt eine Tokenisierung und Segmentierung in Satzsegmente der monolingualen Texte und dann werden die deutschen und spanischen Satzsegmente verknüpft.

Im PaGeS-Korpus wird für die Satz-Alignierung das Open-Source-Programm LF-Aligner (http://sourceforge.net/projects/aligner/) verwendet, weil es in mehreren Tests die besten Ergebnisse erreicht hat. Es basiert auf Hunalign (Varga, *et al*., 2007), eine sehr verbreitete Alignierungs-Software für mehrsprachige Korpora. LF-Aligner verwendet für die Alignierung sowohl Satzlänge[8] als auch lexikalische Entsprechungen.[9] Da aber die lexikalischen Korrespondenzen selbst automatisch abgeleitet werden, bedarf es nicht eines von außen bereitgestellten Lexikons. Die Alignierung erfolgt in vier Schritten: (a) Die Texte werden tokenisiert und interpunktionsbasiert segmentiert. (b) Dann werden sie unter Verwendung einer modifizierten Version von Brown *et al*.'s (1993) satzlängenbasiertem Modell aligniert. (c) Das Programm erstellt ein automatisches Wörterbuch, das auf dieser ersten Alignierung basiert, und (d) schließlich verfeinert es die Alignierung in einem zweiten Durchlauf unter Verwendung des automatischen Wörterbuchs.

Die Satzalignierung wäre trivial, wenn ein Satz immer in genau einen Satz, Entsprechung 1:1, übersetzt würde. Während des Übersetzungsprozesses können aber Sätze durch den Übersetzer geteilt (Entsprechung 1:2, s. Beispiel 1), zusammengeführt (Entsprechung 2:1), gelöscht (Entsprechung 1:0, s. Beispiel 2), eingefügt (Entsprechung 0:1, s. Beispiel 3) oder neu geordnet werden, um eine natürliche Übersetzung in der Zielsprache zu erzeugen.

1. a. Tenía varios dedos ennegrecidos, tal vez congelados, que parecían aferrar un pequeño objeto. (Sierra, Seg. 50)
   b. Mehrere seiner Finger waren schwarz angelaufen, vielleicht erfroren. Sie schienen einen kleinen Gegenstand zu umklammern.
2. a. «Wem?», fragte ich. «Wem werde ich gefallen?» «Ihm.» (Saffier, Seg. 1206)
   b. —¿A quién? —A él.
3. a. La terra dos mortos. (Sierra, Seg. 4359)
   b. A terra dos mortos‹, wie die Galicier sagen, das Land der Toten.

All diese Fragen sind beträchtliche Herausforderungen für die automatische Satzalignierung. Ihre Trefferquote hängt ganz von der Qualität des Ausgangsmaterials ab, denn der Korrespondenzgrad zwischen den Quell- und Zieltexten variiert erheblich in Abhängigkeit von den Texten selbst, den Übersetzern und der Richtung der Übersetzung. So erreicht das LF-Aligner im PaGeS-Korpus in einigen Werken eine Trefferquote von 98 %, in anderen kann es jedoch auf Werte unter 90 %

---

[8] Längenbasierte Ansätze vergleichen die Satzlänge der Ausgangssprache gemessen in Zeichen (Gale / Church 1993) oder Wörtern (Brown *et al.* 1993) mit der der Zielsprache, um die Wahrscheinlichkeit der Satzentsprechung zu bewerten.

[9] Lexikalische Ansätze (Kay / Roscheisen 1993) schlagen vor, die Sätze unter Verwendung eines lexikonbasierten Verfahrens zu alignieren. Sichere Ankerpunkte werden durch zweisprachige Wörterbücher oder Oberflächenähnlichkeiten von Wortformen identifiziert.

sinken. Im letzteren Fall haben wir die Werke verworfen, da sich der große Aufwand der manuellen Nachkorrektur nicht gelohnt hätte. Besonders problematisch sind in dieser Hinsicht die Werke, in denen die deutsche und die spanische Fassung Übersetzungen aus einer dritten Sprache sind, da sie zwei unabhängige Übersetzungsprozesse durchlaufen haben.

Nach der Alignierung exportiert LF-Aligner die Texte in eine Excel Tabelle, wo die manuelle Prüfung erfolgt. Nur dadurch lassen sich die von uns angestrebten Ergebnisse, eine Fehlerquote von unter 0,5%, erzielen. Die Nachkorrektur umfasst drei Schritte: Zuerst werden die Segmente, die über 350 Zeichen enthalten, geteilt. In einem zweiten Schritt werden die Segmente, die keine Entsprechung in der anderen Sprache haben, gefiltert. Hier kann es sich um eine falsche Alignierung handeln oder um Löschungen bzw. Einfügungen im Übersetzungstext. Ist das Bisegment falsch aligniert, werden die erforderlichen Korrekturen gemacht. Wenn der Text nicht übersetzt oder eingefügt wurde, wird dies entsprechend markiert. Tabelle 2 zeigt mehrere falsch alignierte Bisegmente.

Tabelle 2: Falsch alignierte Bisegmente (Saffier: Seg. 5108-5112)

Ich blickte auf die beiden armen Kerle, die ich verunstaltet hatte.	
Leider besaß ich keinerlei Fähigkeit, um sie zu heilen.	Miré a los pobres tipos a los que había desfigurado.
Es würde gewiss Wochen dauern, bis sie wieder gesund waren.	Por desgracia, no poseía la habilidad de curarlos.
Was war ich nur für eine hohle Nuss.	Seguramente tardarían semanas en recuperarse.
Ich war kopfüber in eine Situation gestürzt, ohne sie zu umreißen.	Qué idiota era. Me había lanzado de cabeza a una situación sin analizarla antes.

Schließlich wird, um das manuelle Prüfverfahren zu entlasten, für jedes Bisegment ein Wahrscheinlichkeitswert angegeben, der sich aus dem Quotient der Summe und der Differenz der Längen der beiden Segmente (in Zeichen) berechnet.[10] Dieser Wert wird zur Sortierung der Bisegmente verwendet. Dies beruht auf der Überlegung, dass die Originalsätze und deren Übersetzung tendenziell eine ähnliche Zeichenzahl haben. Die meisten falsch alignierten Bisegmente haben einen Wahrscheinlichkeitswert zwischen -5 und 5. Deswegen wird bei der manuellen Prüfung diesen Bisegmenten besondere Beachtung geschenkt. Diese Vorgehensweise ist weniger arbeitsintensiv und weniger zeitaufwändig als eine vollständige Überprüfung der Alignierung. Trotzdem ist sie unseres Erachtens in der Lage eine hohe Genauigkeit zu sichern, indem sie ein Kompromiss zwischen Wünschbarem und Machbarem darstellt.

Beim aktuellen Stand wurden 600.146 Segmente (s. Tab. 3) automatisch aligniert, von denen weniger als ein Drittel manuell überprüft worden ist.

---

[10] Die Formel lautet: W=Za+Zb/Za-Zb, wobei W für den Wahrscheinlichskeitwert steht, Za für die Zeichenzahl des Segments in der Augangssprache und Zb für die Zeichenzahl des Segments in der Zielsprache.

Tabelle 3: Zahl der Bisegmente (Stand: November 2016)

Originalsprache	Ungeprüft	Geprüft
Deutsch	261.627	79.896
Spanisch	294.689	75.001
Dritte Sprache	43.830	8.226
Total	600.146	163.123

6. Linguistische Annotationen

Linguistische Annotationen sind Informationen zu linguistischen Merkmalen, die den Primärdaten des Korpus in digitaler Form beigefügt sind. (Störrer 2013: 220). Das PaGeS-Korpus ist auf morphosyntaktischer Ebene annotiert, d.h. jedes Textwort wurde einer Wortart (engl. *part of speech*) zugewiesen. Diese Annotation, die meist in der Korpuslinguistik mit dem englischen Begriff PoS (Part-of-Speech)-Tagging bezeichnet wird, wurde automatisch vom IMS TreeTagger ausgeführt.[11]

Im Vorfeld des Taggings erfolgt die Tokenisierung der Texte. Dadurch wird festgelegt, welche Zeichenfolgen (Tokens) als eine Einheit betrachtet werden. Tokens schließen neben den Textwörtern auch Zahlen in Ziffern, Satzzeichen und Sonderzeichen wie &, $ oder § ein. Die Wortartenannotation erfolgt in zwei Schritten: (a) Nach der Tokenisierung wird jedem Token die entsprechende Menge der möglichen Tags zugeordnet. (b) Tag-Disambiguierung: Durch verschiedene Verfahren wird von der Menge der Tags auf eins reduziert (Schmidt 1995).

Das verwendete Inventar von Wortartbezeichnungen wird als Tagset bezeichnet, das sprachspezifisch ist, da es von den spezifischen grammatischen Gegebenheiten der einzelnen Sprachen abhängt. Sprachen mit einer reicheren Flexionsmorphologie haben in der Regel längere Tagsets. Das Tagset des TreeTaggers, STTS (Stuttgart-Tübingen-TagSet), das sich als Standard durchgesetzt hat, umfasst elf Wortkategorien: Nomina, Verben, Artikel, Adjektive, Pronomina, Kardinalzahlen, Adverbien, Konjunktionen, Adpositionen, Interjektionen und Partikel. Jede Wortkategorie wird nach distributionellen, morphologischen und syntaktischen Kriterien noch weiter unterteilt. Das STTS-Tagset hat insgesamt 54 Tags für das Deutsche und 75 Tags für das Spanische einschließend Tags für Interpunktion, numerische Angaben und Daten.

Als dritter Schritt werden die Tokens lemmatisiert. Bei der Lemmatisierung wird jedem Textwort eine bestimmte Grundform oder Lemma zugewiesen. So werden flektierte Formen (wie *pudo, podía, podemos*) auf das Lemma *poder* zurückgeführt. TreeTagger erstellt eine Textdatei mit drei tabulatorgetrennten Spalten: In der ersten stehen die Tokens, in der zweiten die Wortarten und in der dritten

---

[11] Der Tree-Tagger wurde am Institut für maschinelle Sprachverarbeitung (IMS) der Universität Stuttgart von H. Schmidt (1994, 1995) entwickelt. Die ursprüngliche Version des Tree Taggers wurde für das Englische entwickelt. Inzwischen gibt es Versionen für mehr als 20 Sprachen (siehe http://www.cis.uni-muenchen.de/~schmid/tools/ TreeTagger/). Wir erwägen die Verwendung eines anderen PoS-Taggers für das Spanische, FreeLing (http://nlp.lsi.upc.edu/freeling/), der von L. Padró (2011) entwickelt wurde. Bisher ist noch keine endgültige Entscheidung getroffen.

die Lemmata. Wie in Tabelle 4 ersichtlich, sind die Tags-Abkürzungen auch sprachspezifisch.

Tabelle 4. PoS-Tagging und Lemmatisierung unter Verwendung des TreeTaggers

Token	**PoS-Tag**	**Lemma**	**Token**	**PoS-Tag**	**Lemma**
Wer	PWS	wer	En	PREP	en
nicht	PTKNEG	nicht	su	PPO	suyo
zaubern	VVINF	zaubern	opinión	NC	opinión
konnte	VMFIN	können	,	CM	,
,	$,	,	quienes	REL	quien
war	VAFIN	sein	no	NEG	no
seiner	PPOSAT	sein	podían	VLfin	poder
Meinung	NN	Meinung	emplear	VLinf	emplear
nach	APPR	nach	la	ART	el
wertlos	ADJD	wertlos	magia	NC	magia
,	$,	,	eran	VSfin	ser
und	KON	und	seres	NC	ser

Die Annotation erweitert die Suchmöglichkeiten erheblich. Auf ihrer Basis können nicht nur Wortformen, sondern auch Lemmata gesucht werden, bei der alle flektierten Formen zur Grundform ausgegeben werden. Sie ermöglicht auch eine Suche nach Wortarten oder Folge von Wortarten.

Das PoS-Tagging dient auch dazu, homonyme Wortformen zu unterscheiden, z. B. *sein* als Infinitiv und *sein* als Possessivpronomen. Man kann aber nicht fehlerfreie Zuordnungen erwarten. Besonders problematisch ist im Deutschen die Disambiguierung von Relativpronomen vs. Artikel, finiten Vollverben vs. deren Infinitiv sowie Eigennamen vs. normalem Nomen.

Außer der morphosyntaktischen Annotation sind noch verschiedene andere grammatische Annotationen entwickelt worden. So gibt es syntaktische Annotationen zur Konstituentenstruktur und zu grammatischen Funktionen und semantische Annotationen zu Lesarten und semantischen Rollen. (vgl. Zinsmeister 2010: 484, Lemnitzer / Zinsmeister 2010: 60ff.). Derzeit beschränkt sich jedoch die linguistische Annotation des PaGeS-Korpus auf Wortarten und Lemmata.

7. Zugriff und Visualisierung

Das PaGeS- Korpus ist seit März 2016 online verfügbar und für wisseschaftliche Zwecke frei zugänglich (abrufbar unter www.corpuspages.eu). Die alignierten und annotierten Texte werden als Textdateien gespeichert und in der Suchmaschine

Solr[12] indiziert. Die Benutzeroberfläche ist eine webbasierte Anwendung, erstellt durch das Grails Framework (https://grails.org/) (s. Abb. 4).

Auch wenn PaGeS für einen spezifischen Forschungszweck, wie eingangs schon erwähnt, erzeugt wurde, war es von Beginn an unser Ziel, dass das Korpus über seinen intendierten Zweck hinaus vielfach genutzt werden kann. Volk *et al*. (2014) erwähnen drei Hauptnutzergruppen der Parallelkorpora, nämlich Fremdsprachenlerner, Übersetzer und Linguisten, jede davon mit unterschiedlichen Bedürfnissen. Während Fremdsprachenlerner wohl schnell und unkompliziert Übersetzungsvorschläge bekommen möchten, brauchen Übersetzer womöglich zusätzlich Suchfilter, die die Suche nach bestimmten Kriterien wie Datum oder Autor eingrenzen. Schließlich verlangen wahrscheinlich Linguisten oder Translationswissenschaftler höhere Anforderungen, um komplexe Suchanfragen formulieren zu können.

Diesen unterschiedlichen Nutzergruppen differenzierte Suchmöglichkeiten anzubieten, um deren jeweiligen Anforderungen gerecht zu werden, stellt eine große Herausforderung dar. Im Hinblick darauf haben wir eine Benutzeroberfläche entwickelt, die relativ komplexe Suchanfragen erlaubt, ohne dadurch andere Nutzer abzuschrecken.

Abbildung 4. Suchmaske der einfachen Suche (www.corpuspages.eu)

Abbildung 4 zeigt die einfache Suchmaske. Sie besteht nur aus einem Eingabefeld, in das der Suchbegriff (auf Deutsch oder Spanisch) eingegeben wird. Der Nutzer kommt schnell und ohne jeglichen Aufwand zu einer übersichtlichen Trefferliste. Standardmäßig ist die Suche lemmatisiert, d. h. es wird nach allen flektierten Formen des betreffenden Wortes gesucht. Wird der Suchbegriff in Anführungszeichen eingeschlossen, schränkt man die Suche auf die genaue Wortform

[12] Solr (http://lucene.apache.org/solr/) ist ein Enterprise Search Server auf der Basis der Lucene Java-Bibliothek und kann unabhängig vom Portal betrieben werden. Die Indexierung und Abfrage der Texte erfolgen über eine HTTP/XML

ein. Werden mehrere Wörter eingegeben, werden alle Textabschnitte gefunden, die alle eingegebenen Wörter in einem Abstand von maximal fünf Wörtern enthalten.

Die einzelnen Treffer werden zusammen mit ihrer Übersetzung untereinander aufgelistet und der Suchbegriff ist durch Fettdruck hervorgehoben. Ein Ausschnitt aus dem vorhergehenden und dem nachfolgenden Text wird auch jeweils angezeigt.

Wenn man [Quelle] anklickt, gelangt man auf eine Seite, die die bibliographischen Angaben der Textstelle, wie Autor, Titel, Verlag und Kapitelangabe aufweist, wie die untenstehende Abbildung zeigt. Hier lässt sich auch der Kotext um y Sätze vor und nach dem gesuchten Wort erweitern.

Abbildung 5: Quellenangabe und Erweiterung des Kotextes

Was die Sortierung der Ergebnisse angeht, werden zuerst die Treffer in der Originalsprache und dann die Treffer in den Übersetzungstexten angezeigt. Die nach Eingabe einer Suche angezeigten Ergebnisse lassen sich als Text- oder Excel-Datei auf den eigenen Rechner herunterladen.

Abbildung 6 zeigt die Suchmaske der erweiterten Suche. Sie ist noch nicht implementiert und wird voraussichtlich ab März 2017 voll funktionsfähig sein. In der erweiterten Suche gibt es die Möglichkeit, Suchfilter zu verwenden, die die Suche nach bestimmten Kriterien einschränken. Zur Suchverfeinerung sind über eine Eingabemaske bzw. eine Dropdown-Auswahlliste die wichtigsten Parameter wählbar: Version (Original oder Übersetzung), Datum, Autor und Werk. Alle Angaben werden miteinander kombiniert. Die Treffer können auch hier nach unterschiedlichen Kriterien sortiert werden.

Abbildung 6: Suchmaske der erweiterten Suche

Schließlich können über die Optionen der erweiterten Suche hinaus noch komplexere Suchanfragen gestellt werden, indem weitere Parameter entsprechend der

Abfragesyntax. der Suchmaschine Solr in das Eingabefeld eingetragen werden. Durch die Verwendung von regulären Ausdrücken in der Form /REGEX/ können verschiedene Parameter und Optionen in den Suchanfragen kombiniert werden, was eine Vielzahl komplexer Korpusabfragen ermöglicht.

8. Fazit und Ausblick

Dieser Artikel beschreibt die verschiedenen Schritte, die bis zum Zeitpunkt seiner Abfassung in den Aufbau des zweisprachigen Parallelkorpus PaGeS abgeschlossen wurden. Es handelt sich hierbei um ein laufendes Projekt, und wir werden nun einerseits dem Korpus neue Werke hinzufügen, um die Datengrundlage zu erweitern. Auf der anderen Seite wollen wir weitere wichtige Funktionalitäten wie die Wortalignierung oder die erweiterte Suche implementieren. Geplant ist auch eine Aufbereitung der vorliegenden Metadaten nach den Richtlinien der TEI (Text Encoding Initiative)[13], die sich zu einem Standard innerhalb der Korpuslinguistik und der Geisteswissenschaften entwickelt haben, um so die Nachhaltigkeit der Daten besser zu sichern (Lüdeling / Walter 2010). Zusätzlich sollen Funktionalitäten zur statistischen Analyse der Daten eingebaut werden.

Trotz der Verfügbarkeit anderer bestehender Parallelkorpora hebt sich das Korpus PaGeS durch eine Reihe von distinktiven Merkmalen ab:

1) Texttyp und Textsorte der Daten. Die Texte von PaGeS beschränken sich nicht auf den Verwaltungs-, Gesetzgebungs- oder kommerziellen Bereich, sondern sie enthalten fiktionale Literatur und Sachbücher, die den aktuellen allgemeinen Sprachgebrauch darstellen.

2) Hohe Qualität der Quelltexte und der Übersetzungen wird gewährleistet, denn sie alle wurden in Verlagen veröffentlicht, die die Texte sorgfältig prüfen.

3) Qualitätskontrolle des ganzen Prozesses. Für jeden Schritt wird ein Qualitätskontrollsystem durchgeführt. Das Korpus wird manuell auf verschiedenen Ebenen überprüft, einschließlich Vorverarbeitung der Texte, Segmentierung, Alignierung und Annotation.

4) Linguistische Annotation. Die Texte sind mit Lemma-Informationen und Wortarten-Tags angereicht, die es erlauben, das Korpus nach linguistischen Kriterien zu durchsuchen.

5) Verfügbarkeit und Multifunktionalität. Das Korpus wird über ein benutzerfreundliches Web-Interface zur Verfügung gestellt, das auf den Benutzer-Suchgewohnheiten basiert und auf unterschiedliche Nutzertypen ausgerichtet ist.

Aufgrund all dieser Merkmale ist das Korpus-PaGeS für eine Vielfalt an Anwendungsmöglichkeiten sehr gut geeignet, wie als empirische Basis für die linguistische und translatorische Forschung, als Datengrundlage für bilinguale Wörterbücher und für Übungsmaterial im Sprach- und Übersetzungsunterricht und nicht zuletzt als unkompliziertes Tool für die Suche nach kontextualisierten Übersetzungsvorschlägen.

---

[13] Die TEI-Richtlinien (http://www.tei-c.org/Guidelines/P5/) inzwischen in der Version P5, ist ein Set an XML-Regeln, um beliebige Texte einheitlich zu codieren.

9. Literaturverzeichnis

Braune, F. / Fraser, A., «Improved Unsupervised Sentence Alignment for Symmetrical and Asymmetrical Parallel Corpora», in: Huang, Ch. / Jurafsky, D. (ed.), *Coling.* Beijing 2010, 81-89.

Brown, P. *et al*., «The Mathematics of Statistical Machine Translation: Parameter Estimation», *Computational Linguistics* 19/2 (1993), 263-311.

Doval, I., «Raumerfassung kontrastiv Deutsch / Spanisch», in: Ogawa, A. (Hg.), *Raumerfassung – Deutsch im Kontrast*. Tübingen: Stauffenburg Verlag 2016, 209-236.

Fabricius-Hansen, C., «Paralleltext und Übersetzung in sprachwissenschaftlicher Sicht», in: Kittel, H. *et al.* (Hg.), *Übersetzung, Translation, Traduction*, vol 1, Berlin / New York: de Gruyter 2004, 322-29.

Gale, W. / Church, K., «A program for aligning sentences in bilingual corpora», *Computational Linguistics* 19/1 (1993), 75-102.

Kay, M. / Röscheisen, M., «Text-Translation Alignment», *Computational Linguistics* 19/1 (1993), 121-142.

Koehn, P., EuroParl, «A parallel corpus for statistical machine translation». *Proceedings of the machine translation summit*, Thailand, Phuket 2005, 79-86. http://www.statmt.org/europarl/ [15.07.2017].

Krause, M. / Doval, I., *Spatiale Relationen – kontrastiv Deutsch – Spanisch.* Tübingen: Groos 2011.

Lemnitzer, L. / Zinsmeister, H., *Korpuslinguistik. Eine Einführung.* Tübingen: Narr, 2. Aufl. 2010.

Lüdeling, A. / Walter, M., «Korpuslinguistik für Deutsch als Fremdsprache Sprachvermittlungund Spracherwerbsforschung», 2009, https://www.linguistik.hu-berlin.de/de/institut/ professuren/korpuslinguistik/ mitarbeiter-innen/anke/pdf/Luedeling-WalterDaF.pdf [12.06.2017].

Mcenery, A. / Xiao, Z., «Parallel and comparable corpora: What are they up to?», *Incorporating corpora: Translation and the linguist. Translating Europe.* Multilingual matters, Chap XX, Clevedon, UK, 2007. http://someya-net.com/104-IT_Kansai_Initiative/corpora_and_translation.pdf [15.07.2017].

Padró, L., «Analizadores Multilingües en FreeLing», *Linguamatica* 3/ 2 (2011), 13-20.

Schmid, H., «Improvements in Part-of-Speech Tagging with an Application to German», *Proceedings of the ACL SIGDAT-Workshop*. Dublin: 1995. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf [15.07.2017].

Schmid, H., «Probabilistic Part-of-Speech Tagging Using Decision Trees», *Proceedings of International Conference on New Methods in Language Processing*, Manchester 1994. Reviewed version: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf [15.07.2017].

Steinberger, R. *et al*., «An overview of the European Union's highly multilingual parallel corpora», *Language Resources and Evaluation,* 48, 4 (2014), 679-707. doi:10.1007/s10579-014-9277-0.

Storrer, A., «Korpusgestützte Sprachanalyse in Lexikographie und Phraseologie», in: Knapp, K. *et al*. (Hg.), *Angewandte Linguistik. Ein Lehrbuch.* 3. Auflage. Tübingen: Francke 2013, 216-239.

Tiedemann, J., «Parallel Data, Tools and Interfaces in OPUS». *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC-2012) ELRA 2012, 2214-2218, www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf [12.06.2017].

Tiedemann, J., *Bitext Alignment*. Toronto: Morgan & Claypool 2011.

Varga, D. *et al*., «Parallel corpora for medium density languages», *Proceedings of the RANLP* 2005, 590-596 https://doi.org/10.1016/j.protcy.2014.11.024. [15.07.2017].

Volk, M. / Graën, J. / Callegaro, E., «Innovations in parallel corpus search tools», in: *Ninth International Conference on Language Resources and Evaluation* (LREC'14), Reykjavik, 2014, http://www.zora.uzh.ch/id/eprint/97282/1/Volk_Graen_Callegaro_LREC_2014_v06.pdf [15.07.2017].

Zinsmeister, H., «Corpora», in: Carstensen, K.-U. *et al.* (Hg.), *Computerlinguistik und Sprachtechnologie: Eine Einführung.* Heidelberg: Spektrum, Akad. Verl., 3. Aufl., 2010, 481-492.

Primärquellen

Safier, D., *Happy family*. Hamburg: Rowohlt 2011.
[Safier, D., *Una familia feliz.* Barcelona: Seix Barral 2012.]

Sierra, J., *El ángel perdido*. Barcelona: Planeta 2011.
[Sierra, J., *Die Rache der Engel.* München: Blanvalet 2013.]