https://doi.org/10.14201/0AQ0320319341

DAS DEUTSCH-SPANISCHE PARALLELKORPUS *PAGES*: AUFBAU UND NUTZUNGSMÖGLICHKEITEN

The German-Spanish Parallel Corpus PaGeS: *Structure and Possibilities of Use*

Irene Doval
Universidad de Santiago de Compostela
María Teresa Sánchez Nieto
Universidad de Valladolid

ZUSAMMENFASSUNG

In diesem Beitrag wollen wir das Parallel Corpus German Spanisch, *PaGeS*[1], vorstellen. Es ist ein bidirektionales Spanisch<>Deutsch Parallelkorpus, das seit 2017 zu Unterrichts- und Forschungszwecken online frei zugänglich ist. Zuerst wird die Zusammensetzung des Korpus-*PaGeS* vorgestellt. Dann werden die Arbeitsschritte in der Konstruktion des *PaGeS*-Korpus beschrieben: Design und Datenerhebung, Vorverarbeitung der Texte,

[1] Das *PaGeS*-Korpus wurde im Rahmen eines von der Nationalen Forschungsagentur (AEI) des Spanischen Ministeriums für Wissenschaft, Innovation und Universitäten geförderten Forschungsprojekts (FFI2013-42571-P, FFI 2017-85938-R, Projektleitung: Irene Doval) erstellt und kann unter der Webadresse www.corpuspages.eu genutzt werden. Volle Funktionalität ist erst nach der Registrierung möglich. Für nähere Informationen zu dem Korpus *PaGeS* s. Doval (2018, 181-197) und Doval/F.Lanza/Jiménez/Liste/Lübke (2019, 103-121).

Ediciones Universidad de Salamanca / CC BY-NC-ND

Metadaten, Alignierung und linguistische Annotation. Die nächsten Abschnitte sind der Organisation der Textressourcen in *PaGeS* sowie den Wegen gewidmet, auf denen der Benutzer sich der verschiedenen Suchoperatoren bedienen kann, um Suchanfragen zu formulieren. Zu diesem Zweck werden eine Reihe von Suchanfragebeispielen vorgestellt, die Belege abrufen, bei denen typische sprachenpaarbedingte bzw. linguistische Übersetzungsprobleme im Sprachenpaar Deutsch/Spanisch zusammen mit ihren Lösungen durch professionelle Übersetzer beobachtet werden können. Somit wird dem Leser der Weg geebnet, selbständig das Potential von *PaGeS* für Forschung und Lehre zu entdecken. Schließlich wird auf die zukünftige Entwicklung des Korpus hingewiesen.

Schlüsselwörter: *Parallelkorpora; Korpus PaGeS; Deutsch als Fremdsprache; Spanisch als Fremdsprache; Übersetzungsforschung.*

ABSTRACT

The aim of this paper is to present Spanish German Parallel Corpus, *PaGeS*. It is a bidirectional Spanish<>German parallel corpus that has been freely accessible online for teaching and research purposes since 2017. Firstly, the make up of the *PaGeS* corpus will be described. Then the steps in the development of the *PaGeS* corpus are outlined: Design and data collection, pre-processing of the texts, the metadata, the alignment, and the linguistic annotation. The following section addresses the organization of text resources in *PaGeS* and the ways in which users can refine their searches and use search parameters to formulate queries. For this purpose, several search query examples are presented that retrieve evidence where typical or linguistic translation problems in the German/Spanish language pair can be observed along with professional translators' solutions. This helps the reader to independently discover the potential of *PaGeS* for research and teaching. Finally, the future development of the corpus is outlined.

Keywords: *parallel corpora; PaGeS corpus; German as a foreign language; Spanish as a foreign language; translation research.*

1. DAS KORPUS *PAGES*: DESIGN UND DATENBESCHAFFUNG

PaGeS besteht aus zwei Hauptteilen: dem Kernkorpus und den Ergänzungen. Die Ergänzungen sind Texte unterschiedlichen Ursprungs, die nachträglich hinzugefügt wurden und als bloße Ergänzung zum Kernkorpus betrachtet werden sollen. Von daher wird sich unsere Beschreibung auf das Kernkorpus konzentrieren, das vollständig im Rahmen des Projekts erarbeitet wurde.

1.1. Das Kernkorpus

Vorab ein paar Anmerkungen dazu. Ein Korpus ist nicht nur eine beliebige 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 (vgl. Lemnitzer/Zinsmeister, 2015: 14; Hirschmann, 2019: 2).

Bei zweisprachigen Korpora muss jedoch eine wichtige Einschränkung hinsichtlich des verfügbaren Materials bei der Auswahl der Texte beachtet werden. Die überwiegende Mehrheit der Texte wird nicht übersetzt und, wenn sie übersetzt werden, unterliegt nur ein kleiner Teil davon einer Qualitätskontrolle. Daher muss offensichtlich das erste Kriterium für die Erstellung eines Parallelkorpus ein opportunistisches sein, d.h., es ist erforderlich, auf das zurückzugreifen, was vorhanden ist.

Dazu kommt das Kriterium der Textqualität der Originale und der Übersetzungen. Abgesehen von institutionellen Sprachressourcen vor allem der Europäischen Union ist die einzige Möglichkeit, die Textqualität zu gewährleisten, schriftliche Texte von angesehenen Verlagen zu verwenden, bei denen sowohl Originaltexte als auch Übersetzungen einer anspruchsvollen Qualitätskontrolle unterzogen werden.

Deshalb enthält das Kernkorpus Originaltexte auf Deutsch und Spanisch und deren veröffentlichte Übersetzungen sowie einen kleinen Anteil (ca. 7 %) von Texten, die aus einer dritten Sprache ins Deutsche und ins Spanische übersetzt wurden. Es umfasst derzeit (April 2021) eine Sammlung von 169 Werken, überwiegend Belletristik (ca. 80 %) sowie Sachtexte verschiedener Gattungen. Alle Bücher sind nach 1960 erschienen, mit besonderem Schwerpunkt auf Werken aus den letzten zwei Jahrzehnten[2].

Einen Anteil von 7 Prozent bilden Werke, die aus einer dritten Sprache ins Deutsche und Spanische übersetzt wurden. So werden nicht nur Originale und Übersetzungen, sondern auch zwei Übersetzungen parallelisiert, wie Abb. 1 zeigt. Die schwarzen Pfeile geben die Richtung der Übersetzung und die Pfeile mit Doppelspitze die parallelisierten Texte an.

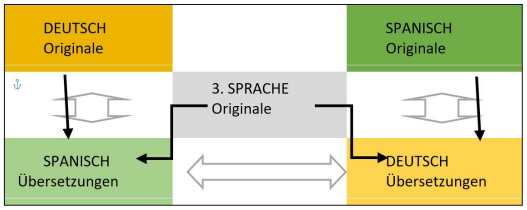

Abbildung 1. *Übersetzungsrichtung und parallelisierte Texte.*

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

Die enthaltenen Werke wurden nicht vollständig, sondern in Auszügen aufgenommen, um die Einschränkungen für urheberrechtlich geschützte Werke zu beachten und dabei auch eine größere Vielfalt an Texten zu erzielen.

Das Kernkorpus von *PaGeS* enthält somit ca. 31.000.000 Textwörter[3] und 1.055.685 Bisegmente, d.h. Paare von alignierten Textchunks (Sätze oder kleinere Segmente). Tabelle 1 gibt einen Überblick über die aktuelle Zusammensetzung des Kernkorpus.

Tabelle 1. *Anzahl der Werke, Textwörter, Types und Bisegmente nach Sprachen und Originaltexten im Kernkorpus (Stand: April 2021).*

SPRACHE	WERKE	TEXT-WÖRTER	TYPES	BI-SEGMENTE
Deutsch Original	81	6.280.994	188.540	461.768
Spanisch Übersetzung	(81)	6.781.481	109.921	
Spanisch Original	70	7.010.327	119.501	442.623
Deutsch Übersetzung	(70)	6.924.157	162.972	
Deutsch Übersetzung < 3. Sprache	18	2.084.860	74.244	151.294
Spanisch Übersetzung < 3. Sprache	(18)	2.149.204	57.389	
Gesamt	169	31.267.205		1.055.685

Abb. 2 zeigt den Anteil der Textwörter nach Sprachen sortiert. Hier wird ersichtlich, dass das Korpus bezüglich der Übersetzungsrichtung ganz ausgeglichen ist.

[3] In Anlehnung an die DWDS-Korpora (www.dwds.de/d/korpora) wird hier der Begriff 'Textwort' für jedes Vorkommen eines Wortes in einem fortlaufenden Text verwendet. Als alternative Bezeichnungen findet man auch 'Token', 'Wortvorkommen' oder 'laufendes Wort' (s. Störrer 2013, 219).

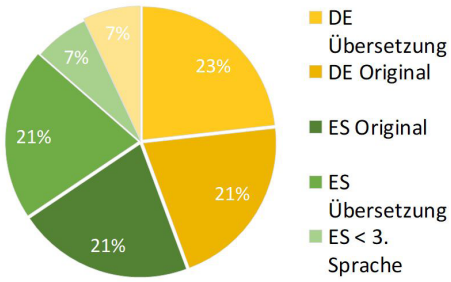

Abbildung 2. *Anteil der Sprachdaten im PaGeS nach Sprachen und Originaltexten.*

1.2. Ergänzungen

Wie bereits die Bezeichnung erkennen lässt, wurde dieser Teil des Korpus lediglich als Ergänzung zum Kernkorpus konzipiert, um bestimmte Textsorten wie Verwaltungs- und Rechtssprache zu berücksichtigen, die nicht durch den Bestand des Kernkorpus (Belletristik und Sachbücher) abgedeckt waren. Im Gegensatz zum Kernkorpus wurden weder die Texte der Ergänzungen noch das Alignment manuell geprüft und es wurde hier kein PoS-Tagging durchgeführt.

Derzeit beinhaltet dieser Teil nur *Europarl*, ein Parallelkorpus, das die ausführlichen Sitzungsberichte des Europäischen Parlaments von 1996 bis 2011 enthält[4]. Die neueste Version (Release v7) umfasst bis zu 50 Millionen Wörter pro Sprache und die Texte sind auf Satzebene aligniert. Folgende Tabelle gibt den Umfang des *Europarls* in Zeichen und Textwörtern an, der in *PaGeS* aufgenommen worden ist.

Tabelle 2. *Anzahl der Textwörter und Bisegmente des Europarl im PaGeS (Stand: April 2021).*

SPRACHE	ZEICHEN	TEXTWÖRTER	BISEGMENTE
Deutsch	219.099.293	35.222.373	1.586.374
Spanisch	205.008.875	39.664.923	
Total	424.108.168	74.887.296	1.586.374

[4] (http://www.statmt.org/europarl) Koehn2005. In *PaGeS* werden die bereinigte und korrigierte Version CoStEP Corpus (https://pub.cl.uzh.ch/wiki/public/costep/start) sowie deren Metadaten verwendet. Weitere Informationen in Graën/Batinic/Volk (2014).

Es ist vorgesehen, dass neue Sammlungen zweisprachiger Texte unterschiedlicher Textsorten und Herkunft hinzukommen. So werden in unmittelbarer Zukunft die Transkriptionen der TED-Talks, deren Texte und Alignment manuell geprüft worden sind, in das Teilkorpus aufgenommen.

2. ARBEITSSCHRITTE BEI DER KONSTRUKTION DES *PAGES*-KORPUS

2.1. Textvorverarbeitung und Metadaten

Die folgenden Abschnitte beschreiben die Arbeitsschritte, die für die Erstellung des *PaGeS*-Kernkorpus ausgeführt worden sind. 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 hauptsächlich drei Aufgaben:

(a) Entfernen von nicht korrespondierenden Textausschnitten, fehlerhaften Zeichen und Bildern

(b) Korrekturlesen

(c) Markieren und Annotieren von Metadaten.

Die Metadaten werden verwendet, um relevante Informationen über die Texte zu erfassen und sie aus dem Korpus abrufen zu können. Jedes der im *PaGeS*-Korpus enthaltenen Werke wird mit einer Metadatenliste versehen, die u. a. folgende Informationen enthält: Autor und Übersetzer, Titel, Erscheinungsjahr und andere bibliographische Information, Originalsprache, Genre und manuellen Prüfer. 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. Nach dem Korrekturlesen und Markieren werden die Texte in einem gemeinsamen Kodierungsschema UTF-8 als Textdateien gespeichert.

2.2. 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 Segmentierung in Satzsegmente der monolingualen Texte und dann werden die deutschen und spanischen Satzsegmente verknüpft, wie in Abb. 3 dargestellt:

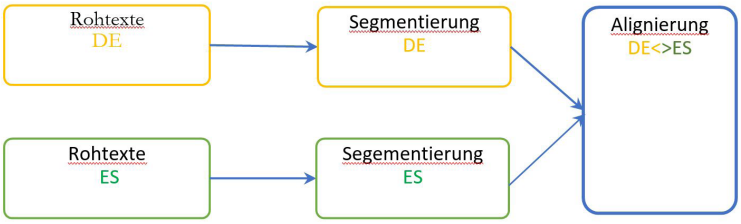

Abbildung 3. *Korpus PaGeS Verarbeitungsprozess 1.*

Im *PaGeS*-Korpus wird für die Satz-Alignierung das Open-Source-Programm LF-Aligner[5] verwendet, weil es in mehreren Tests die besten Ergebnisse gezeigt hat. Es basiert auf Hunalign (Varga et al., 2007), einer sehr verbreiteten Alignierungs-Software für mehrsprachige Korpora. LF-Aligner verwendet für die Alignierung sowohl Satzlänge als auch lexikalische Entsprechungen[6]. Da aber die lexikalischen Korrespondenzen selbst automatisch abgeleitet werden, bedarf es keines von außen bereitgestellten Lexikons.

Die Satzalignierung wäre trivial, wenn ein Satz immer in genau einen Satz (Entsprechung 1:1) übersetzt würde. Während des Übersetzungsprozesses werden jedoch Sätze im Zieltext ausgelassen (Entsprechung 1:0, s. Tab. 3) bzw. eingefügt (Entsprechung 0:1, s. Tab. 4), der Übersetzer kann einen Satz teilen (Entsprechung 1:2, s. Tab. 5), zusammen-

[5] http://sourceforge.net/projects/aligner/

[6] Längenbasierte Ansätze vergleichen die Satzlänge der Ausgangssprache gemessen in Zeichen (Gale/Church 1993) mit der der Zielsprache, um die Wahrscheinlichkeit der Satzentsprechung zu bewerten. Lexikalische Ansätze (Kay/Roscheisen 1993) schlagen vor, die Sätze unter Verwendung eines lexikonbasierten Verfahrens zu alignieren.

führen (Entsprechung 2:1, s. Tab. 6), oder neu ordnen, um eine natürliche Übersetzung in der Zielsprache zu erzeugen.

Tabelle 3. *Entsprechung 1:0 (Safier, Seg. 1206).*

«Wem?»,	*–¿A quién?*
fragte ich. «Wem werde ich gefallen?»	
«Ihm.»	*–A él.*

Tabelle 4. *Entsprechung 0:1 (Sierra, Seg. 4359).*

A terra dos mortos.	*A terra dos mortos,*
	wie die Galicier sagen, das Land der Toten.

Tabelle 5. *Entsprechung 1:2 (Sierra, Seg. 50).*

Tenía varios dedos ennegrecidos, tal vez congelados, que parecían aferrar un pequeño objeto.	*Mehrere seiner Finger waren schwarz angelaufen, vielleicht erfroren.*
	Sie schienen einen kleinen Gegenstand zu umklammern.

Tabelle 6. *Entsprechung 2:1 (Saffier: Seg. 5108-5112).*

Man werde sich einigen, sagte der Herzog.	*Llegarían a un acuerdo, dijo el duque.*
Ein Professorentitel sei möglich.	
Wenn auch nicht bei doppelten Bezügen.	*La cátedra era posible, aunque sin doble sueldo.*

All diese Fragen sind beträchtliche Herausforderungen für die automatische Satz-Alignierung. 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 Übersetzungsrichtung. So erreicht das LF-Aligner im *PaGeS*-Korpus in einigen Werken eine Trefferquote von 98 %, in anderen kann es jedoch auf Werte von unter 80 % sinken. Besonders problematisch sind in dieser Hinsicht die Werke, in denen die deutsche und die spanische Fassung Übersetzungen aus einer dritten Sprache sind, da sie dann 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.

Um das manuelle Prüfverfahren zu entlasten, wird schließlich für jedes Bisegment ein Wahrscheinlichkeitswert angegeben, der sich aus dem Quotienten der Summe und der Differenz der Längen der beiden Segmente (in Zeichen) berechnet[7]. Dieser Wert wird zur Sortierung der Bisegmente verwendet. 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 einen Kompromiss zwischen Wünschenswertem und Machbarem darstellt. Beim aktuellen Stand wurden 1.055.685 Segmente (s. Tabelle 1) automatisch aligniert und manuell überprüft.

2.3. 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 wird meist in der Korpuslinguistik mit dem englischen Begriff *PoS (Part-of-Speech)-Tagging* bezeichnet.

Da es sich bei *PaGeS* um eine bilinguale Textsammlung handelt, wurden zwei verschiedene Tagger verwendet: der IMS TreeTagger[8] für das Deutsche und FreeLing[9] für das Spanische, weil sie die besten Ergebnisse in der jeweiligen Sprache erreicht haben.

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

[8] 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/).

[9] Freeling wurde von L. Padró (2011) entwickelt. (http://nlp.lsi.upc.edu/freeling/).

Dieser Prozess erfolgt einzelsprachlich, deswegen müssen die alignierten Texte mit der entsprechenden Markierung wieder getrennt werden, wie Abb. 4 veranschaulicht.

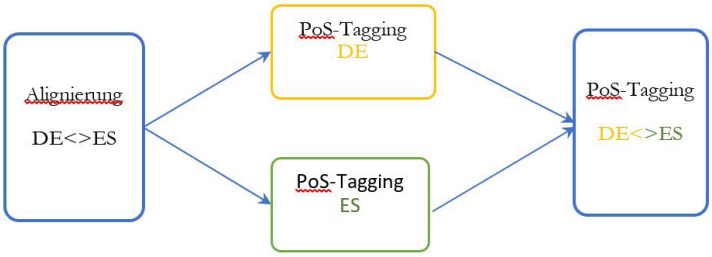

Abbildung 4. *Korpus PaGeS Verarbeitungsprozess 2.*

Die Wortartenannotation erfolgt in zwei Schritten:

(a) Nach der Tokenisierung (dadurch wird festgelegt, welche Zeichenfolgen –Tokens– als eine Einheit betrachtet werden, vgl. Hirschmann 2019, 31) wird jedem Token die entsprechende Menge der möglichen Tags zugeordnet.

(b) Tag-Disambiguierung: Durch verschiedene Verfahren wird die Menge der Tags auf eins reduziert (Schmidt, 1995).

Das verwendete Inventar von Wortartbezeichnungen wird als *Tagset* bezeichnet, das sprachspezifisch ist, da es von den individuellen grammatischen Gegebenheiten der einzelnen Sprachen abhängt. Sprachen mit einer reicheren Flexionsmorphologie haben in der Regel längere Tagsets. Das Tagset des TreeTagger, STTS (Stuttgart-Tübingen-TagSet), das sich für das Deutsche 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, einschließlich Tags für Interpunktion, numerische Angaben und Daten. FreeLing verwendet den EAGLES-Tagset für Spanisch[10]. EAGLES PoS-Tags bestehen aus Tags mit unterschiedlicher Länge, wobei jedes Zeichen einem morphologischen Merkmal entspricht. Das erste Zeichen im Tag ist immer die Kategorie (PoS). Die Kategorie bestimmt die Länge des Tags und die Interpretation jedes Zeichens im Tag.

[10] EAGLES *Recommendations for the Morphosyntactic Annotation of Corpora*. EAGLES Document EAG–TCWG–MAC/R. Version of Mar, 1996. S. https://www.sketchengine.eu/spanish-freeling-part-of-speech-tagset/

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. Der TreeTagger ist aber nicht in der Lage, trennbare Verben korrekt zu lemmatisieren, wenn das Präfix vom Verbstamm abgetrennt vorkommt. Volk et al. (2014) haben einen Algorithmus entwickelt, um das getrennte Präfix wieder an das Verb anzufügen[11], wie in Tab. 7 bei *aus+sehen* zu finden ist.

TreeTagger erstellt eine Textdatei mit drei tabulatorgetrennten Spalten: In der ersten stehen die Tokens, in der zweiten die Wortarten und in der dritten die Lemmata (s. Tab. 7).

Tabelle 7. *PoS-Tagging und Lemmatisierung unter Verwendung des TreeTaggers.*

Token	PoS-Tag	Lemma
Diese	PDAT	dies
Inschrift	NN	Inschrift
stand	VVFIN	stehen
auf	APPR	auf
der	ART	die
Glastür	NN	Glastür
eines	ART	eine
kleinen	ADJA	klein
Ladens	NN	Laden
,	$,	,
aber	KON	aber
so	ADV	so
sah	VVFIN	aus+sehen
sie	PPER	sie
natürlich	ADV	natürlich
nur	ADV	nur
aus	PTKVZ	aus

[11] Das Skript wurde uns freundlicherweise von Prof. Volk zur Verfügung gestellt. Ihm sei hier ganz herzlich für seine Unterstützung gedankt.

FreeLing zeigt in einer ähnlichen Textdatei in der ersten Spalte die Tokens, in der zweiten die Lemmata und in der dritten die PoS-Tags. (s. Tab. 8)

Tabelle 8. *PoS-Tagging und Lemmatisierung unter Verwendung des Freelings.*

Token	Lemma	PoS-Tag
Esta	este	PD0FS00
era	ser	VSII3S0
la	el	DA0FS0
inscripción	inscripción	NCFS000
que	que	PR0CN00
había	haber	VMII3S0
en	en	SP
la	el	DA0FS0
puerta	puerta	NCFS000
de	de	SP
cristal	cristal	NCMS000
de	de	SP
una	uno	DI0FS0
tiendecita	tienda	NCFS00V
,	,	Fc

Wie aus Tabellen 7 und 8 ersichtlich, sind die Tag-Abkürzungen auch sprachspezifisch. Die Annotation erweitert die Suchmöglichkeiten erheblich. Auf ihrer Basis können nicht nur Wortformen, sondern auch Lemmata gesucht werden, bei denen alle flektierten Formen zur Grundform ausgegeben werden. Sie ermöglicht auch eine Suche nach Wortarten oder Folgen von Wortarten.

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

3. *PAGES*: SUCH- UND ANWENDUNGSMÖGLICHKEITEN

3.1. Wo kann gesucht werden? Sektionen in *PaGeS*

PaGeS bietet zwei Schnittstellen für den Datenzugriff. Sie sind auf der *PaGeS*-Webseite mit den Labels «Einfache Suche» und «Erweiterte Suche» identifiziert (siehe Abb. 5 und 6).

Abbildung 5. *Schnittstelle «Einfache Suche» auf www.corpuspages.eu.*

Abbildung 6. *Schnittstelle «Erweiterte Suche» auf www.corpuspages.eu.*

Die Schnittstelle «Erweiterte Suche» gestattet dem Benutzer Zugang zu den verschiedenen Sektionen, aus denen *PaGeS* besteht und aus denen der Benutzer Subkorpora absondern kann.

(1) Zum einen kann man die Suchanfrage – sei sie eine Suche nach spanischsprachigen oder deutschsprachigen Belegen – entweder auf Originaltexte oder auf übersetzte Texte (d. h. auf spanische bzw. auf deutsche Übersetzungen) beschränken. Die Suche kann man aber auch in beiden Subkorpora formulieren, d. h., ohne die Übersetzungsrichtung zu beobachten, indem man auf der Schnittstelle «Erweiterte Suche» die Optionen «Originaltexte» UND «Übersetzungen < DE/ES» anklickt. Es gibt aber weitere Ressourcen in *PaGeS,* in denen man suchen kann, und zwar in übersetzten Texten aus einer dritten Sprache ins Deutsche oder ins Spanische. Dafür muss man auf der Schnittstelle «Erweiterte Suche» die Option «Übersetzungen < 3. Sprache» anklicken. Zuletzt kann man auch Belege aus der *Europarl*-Erweiterung miteinbeziehen bzw. abrufen (siehe Abb. 7).

(2) Zum anderen kann man aber die Textbasis, auf die unsere Suchfrage zielt, weiter unterteilen, indem man nur in bestimmten Werken sucht, oder in Werken eines bestimmten Genres, eines bestimmten Zeitraums bzw. einer bestimmten Dialektalvariante[12]. Natürlich kann man diese Kriterien miteinander verbinden und auf eine Suchanfrage in einer oder mehreren der oben beschriebenen Subsektionen anwenden (siehe Abb. 6).

Eigenschaft der Texte	(Ziel- bzw. Ausgangs)sprache	enthalten in...		
		Kernkorpus	Europarl-Erweiterung	TED-Talks-Erweiterung
Originaltexte	spanische Texte	☑	☑	
	deutsche Texte	☑	☑	
übersetzte Texte	ins Spanische	☑	☑	
	ins Deutsche	☑	☑	
	ins Deutsche aus einer 3. Sprache	☑	☑	☑
	ins Spanische aus einer 3. Sprache	☑	☑	☑

Abbildung 7. *Möglichkeiten der Ressourcenabgrenzung und -kombination.*

Hinsichtlich der Suche in der Ressource *Europarl* muss Folgendes angemerkt werden:

a. Hier kann der Benutzer die Originalsprache (des Redners) ebenfalls bestimmen. Die Konfiguration der SolR-Suchmaschine, der sich *PaGeS* bedient, setzt allerdings bestimmte Beschränkungen in diesem Sinne: Als Originalsprachen gelten ausschließlich Spanisch, Deutsch, eine «dritte» (d. h. jede Sprache außer Deutsch und Spanisch) oder eine «beliebige» Sprache (jede Sprache, inklusive DE und ES). Um nach Belegen z. B. von einem deutschsprachigen Redner zu suchen, muss man auf der Erweiterten Suche den Button «Originaltexte» anklicken und die Suchanfrage im Suchfeld «Deutsch» angeben.

b. Die Ergebnisse einer unter Einbeziehung von *Europarl* formulierten Suchanfrage in Originaltexten enthalten immer Belege aus der Kernkorpus-Textbasis, wie aus Abb. 8 zu entnehmen ist – es sei denn, man beschränkt die Suche ausschließlich auf *Europarl*-Daten (siehe Abb. 5). Dafür muss sich der Benutzer des Felds Werk-ID bedienen und einen oder

[12] Die Dialektalvariante der im Kernkorpus enthaltenen Textauszügen kann anhand der aufgelisteten bibliographischen Angaben festgestellt werden. Dazu gelangt man über die Schaltfläche «Textressourcen» oben rechts auf der Webseite.

mehrere Kodierungen der *Europarl*-Erweiterungsdateien angeben: 9996, 9997, 9998, 9999, 9900, 9901, 9902, 9903, 9904, 9905, 9906, 9907, 9908, 9909, 9910, 9911. Mit «9*» werden alle Europarl-Kodierungen in die Suchanfrage aufgenommen. Bei diesen Kodierungen identifizieren die zwei ersten Zahlen die Zugehörigkeit der Daten zur *Europarl*-Erweiterung und die zwei nächsten jeweils das Jahr, aus dem die Debatten entstammen (9996 => 1996; 9911 => 2011, usw.). Diese Prozedur hat den Vorteil, dass Benutzer mittels eines einzigen Suchfelds die Ergebnisse gleichzeitig auf *Europarl*-Inhalte UND auf bestimmte Jahre beschränken können. Die in *PaGeS* einprogrammierte *Europarl*-Version ist die bereinigte und korrigierte Version *CoStEP Corpus*, die in Graën et al. (2014) näher beschrieben wird. Die Daten stammen aus den Jahren 1996 bis 2011.

Abbildung 8. *Ergebnisse der Suchanfrage mit Angabe der Originalsprache in der Europarl-Erweiterung ohne Ausschluss von Kernkorpus-Daten auf www.corpuspages.eu.*

Abbildung 9. *Ergebnisse der Suchanfrage mit Angabe der Originalsprache in der Europarl-Erweiterung unter Ausschluss von Kernkorpus-Daten auf www.corpuspages.eu*

Zwei Eigenschaften der Ergebnisdarstellung bei *PaGeS* sollen an dieser Stelle erklärt werden.

(1) Über den Kodierungs-Hypertext gelangen Benutzer zu einer weiteren Seite, auf der der spezifische Beleg erscheint, zusammen mit der vollen bibliographischen Angabe (im Falle der *Europarl*-Belege schließt dies die Angabe der Sprache des Redners mit ein) und der Möglichkeit der Kontexterweiterung.

(2) Die gesuchte Charakterkette erscheint immer **fett**, unabhängig von der Ressource, in der sie gesucht wurde. Das kann man sehr gut anhand von Realien verdeutlichen, die oft als Lehnwörter in die Übersetzung aufgenommen werden – z. B. kulinarische Realien wie *Esqueixada*/*Butifarra* (ES) bzw. *Kindergarten* (DE). Die Fettdruckkonvention bei der Ergebnisdarstellung erlaubt es, schnell jene Fälle auszumachen, in denen das Lehnwort in der jeweiligen Übersetzung *nicht* benutzt wurde.

3.2. Wie kann gesucht werden? Einführung in die Suchanfrageformulierung auf *PaGeS*

Die folgende Darstellung erfolgt auf der Basis der Operatoren, deren sich die SolR-Suchmaschine bedient.

- Wortunterbrechungen und Charakterauslassungen: *, ?
- Distanzoperator: ~
- boolesche Operatoren: OR, NOT, AND
- Kombination von booleschen Operatoren und Wortunterbrechungen, Charakterauslassungen bzw. Distanzoperator.

Für jeden Operator werden 1 bis 2 Suchanfragebeispiele dargestellt. Sie sollen LeserInnen dazu inspirieren, Suchanfragen, die ihren eigenen Interessen entsprechen, zu entwerfen und auszuprobieren.

3.2.1. *Suchanfragenbeispiele mit dem Wortunterbrechungsoperator (*)*

Der Wortunterbrechungsoperator (WUO) ersetzt immer einen oder mehrere Charaktere, d. h. er wird nicht ignoriert. Das kann man sich sehr gut anhand von Suchen nach Adjektivadverbien vor Augen führen, wie z. B. im Falle von *offensichtlich*.

- «offensichtlich» > Belege von *offensichtlich* als Adverb /prädikatives Adjektiv (exakte Suche, mit Anführungszeichen)

– offensichtlich > Belege von *offensichtlich* als Adverb UND als attributives Adjektiv (lemmatisierte Suche, ohne Anführungszeichen)
– offensichtlich* > Belege von *offensichtlich* als attributives Adjektiv/Substantiv. Der WUO ersetzt die Deklinations- bzw. die Graduierungsflexion, sowie Derivationssuffixe wie z. B. /keit/.

Der WUO kann Charaktere nicht nur am Ende, sondern auch inmitten oder am Anfang eines Wortes ersetzen. Diese Möglichkeit kann man z. B. ausnützen, um Internationalismen als potenzielle falsche Freunde in der Übersetzung DE>ES>DE näher zu betrachten. Einfache Suchanfragen wie *konzept zeigen, wie sehr die Äquivalenten von *Konzept* mit dem jeweiligen Bestimmungswort und unmittelbaren Kotext im Zusammenhang stehen: *Lichtkonzept* > *iluminación; Briefkonzept* > *esbozo/borrador de carta*; *Gesamtkonzept* > *diseño global*; *Handlungskonzept* > *estrategia de actuación*; *Reformkonzept* > *proyecto de reforma* usw.

Der Charakterersatzoperator (CEO) wird hier aus Platzgründen zusammen mit anderen Operatoren unter Punkt 3.2.2 dargestellt.

3.2.2. *Suchanfragenbeispiele mit WUO und CEO in Kombination mit booleschen Operatoren*

Man kann den WUO (*) zusammen mit dem CEO (?) verwenden und einen oder beide Operatoren mit den booleschen Operatoren *OR* bzw. *NOT* kombinieren. Was den Operator *AND* betrifft, so wird er vom System immer angenommen, wenn kein Operator zwischen zwei durch eine Leertaste getrennten Charakterketten erscheint. Zur Verdeutlichung werden im Folgenden verschiedene Suchkontexte angenommen.

3.2.2.1. Medizinische Fachbegriffe

Medizinische Fachbegriffe werden bekanntlich auf Deutsch oft durch Fremdwörter lateinischen bzw. griechischen Ursprungs ausgedrückt. Die Suche ba?teri* (einfache Suche) ergibt also Gebrauchsbeispiele sowohl von dt. *Bakterie, bakteriell, bakterienähnlich*, als auch von sp. *bacteria, bacteriológico*, usw. – dazu aber auch noch Rauschen wie z. B. dt. *Batterie*, usw.

Um das Rauschen auszuschließen, sind die booleschen Operatoren (BO) *NOT* und *OR* hilfreich (Schnittstelle Einfache Suche):

[SS] ba?teri* NOT (*batterie* OR batterie* OR *batterie OR Batterie)

In der Erweiterten Suche erlauben es die BO z. B. auch, nach deutschen Äquivalenten von spanischen medizinischen Ausdrücken mit einer fremdsprachlichen Wurzel zu suchen (z. B, */bacteri/, /infecci/*), die NICHT mit der fremdsprachlichen Wurzel (*/bakteri/, /infekt/, /infiz/*) wiedergegeben werden:

> Suchfeld DE [SS] NOT (Infekt* OR *infekt OR Infekt OR *infekt* OR Infiz* OR *infiz*)
> Suchfeld ES [SS] (infecci* OR *infecci*)

3.2.2.2. Produktivität umgangssprachlicher Präfixe

Mit der Suchanfrage mega* im Suchfeld Deutsch (Einfache Suche) erhält man Belege von adjektivischen und substantivischen Ableitungen mit diesem Präfix auf Deutsch und mögliche Entsprechungen der Ableitungen im Spanischen. Man stößt dabei auf manche umgangssprachliche Neuprägung, die noch nicht Eingang in zweisprachigen Wörterbüchern gefunden hat.

Häufige irrelevante Wörter (Rauschen) wie *Megaphon* oder Fremdwörter wie *Megalomane* kann man in einem zweiten Schritt mithilfe von BO aus der Suche ausschließen:

> [SS] Mega* NOT (Megaphon OR Megalomane)

3.2.2.3. Aspektuelle Mehrdeutigkeit von Präteritum und Perfekt auf Deutsch

Da *PaGeS* im Kernkorpus größtenteils Material aus narrativen Werken enthält, stellt es eine Datenquelle erster Klasse dar, um angehende Übersetzer auf dieses wichtige sprachenpaarbedingte Übersetzungsproblem aufmerksam zu machen (s. Sánchez Nieto 2012).

Um sich bzw. den Studierenden das Phänomen vor Augen zu führen, kann man in *PaGeS* nach Kontexten suchen, in denen bestimmte Formen eines deutschen Tempus mit bestimmten Formen eines spanischen Tempus bzw. einer Periphrase korrelieren. Es muss jedoch vorab bemerkt werden, dass die im Folgenden vorgeschlagenen Suchanfragenbeispiele auf deutsche regelmäßige Verben und auf einzelne Personenformen des Präteritums und der spanischen *Indefinido* bzw. *Imperfecto* beschränkt sind.

Als Erstes kann man eine Suchanfrage erstellen, bei deren Ergebnissen eine Korrespondenz deutscher Präteritumsformen mit Formen des spanischen *Pretérito Indefinido* zu beobachten ist. Der WUOist hier hilfreich, um ausschließlich nach Konjugationsendungen zu suchen, z. B. deutsch */-te/* (Endung der 1. und 3. Person Singular bei regelmäßigen

Verben) und spanisch /*-ó*/ (anwesender Vokal mit Akzentzeichen bei der 3. Person Singular des *Pretérito Indefinido*). Mit dem BO *NOT* im spanischen Suchfeld vermeiden wir, dass auch spanische Segmente als Ergebnisse mitabgerufen werden, die auf /*-te*/ endende Wörter enthalten[13].

Suchfeld DE *te
Suchfeld ES [SS] *ó NOT *te

Auf ähnliche Weise kann man nach Kontexten suchen, in denen das deutsche Präteritum mit Formen des spanischen *Imperfecto* wiedergegeben wird.

Suchfeld DE *te
Suchfeld ES [SS] (*ía OR *aba) NOT *te

Weiter kann man Suchanfragen nach diesem Muster erstellen, die Segmente abrufen, in denen das deutsche Präteritum sich auf den Anfang einer Situation bezieht und somit im Spanischen mit Verbalperiphrasen wie *empezar a* + *Infinitivo* korreliert.

Suchfeld DE *te
Suchfeld ES [SS] empezar a~0 (*ar OR *er OR *ir) NOT *te

Die Tilde (~) funktioniert hier als Distanzoperator und stellt sicher, dass in den spanischsprachigen Segmenten die Formen des Verbs *empezar* und die Präposition *a* direkt nacheinander erscheinen. Zur Tilde werden noch einige Einzelheiten im nächsten Unterabsatz ausgeführt.

3.2.3. *Suchanfragenbeispiele mit dem Distanzoperator ~*

Im Folgenden werden zwei weitere Suchanfragen angeführt, bei denen die Tilde als Distanzoperator (DO) hilft, Belege abzurufen, die interessante Übersetzungsprobleme im Sprachenpaar Deutsch/Spanisch darstellen: deutsche substantivierte Geräuschverben und spanische aspektuelle Verbalperiphrasen.

Für substantivierte deutsche Geräuschverben gibt es bekanntlich manchmal keine direkte lexikalische Entsprechung im Spanischen, wie z. B. im Satz «Im gleichen Augen-

[13] Bei solchen großangelengten Suchanfragen ist es ratsam, die Suche zu beschränken (z. B. auf ein einziges Werk mittels dem Suchfeld ID-Werk), damit die Anzahl der Ergebnisse in einem handhabbaren Rahmen bleibt.

blick hörte sie **ein Kratzen**, direkt neben sich in der Hüttenwand»[14]. Um zu beobachten, wie professionelle Übersetzer diese Einheit in der Übersetzung behandelt haben, bieten sich hier die Suchanfragen [SS] ein Kratzen~0, oder [SS] das Kratzen~0 an – im Prinzip in Originaltexten (Erweiterte Suche). Der Suchmotor stellt dann Belege bereit, bei denen irgendeine Form des unbestimmten bzw. bestimmten Artikels unmittelbar vor irgendeiner Form des Substantivs *Kratzen* steht, da die Suche lemmatisiert ist. Darunter befinden sich die unten angeführten Bisegmente, bei denen der Valenzunterschied zwischen *Kratzen* und *arañar* zu beobachten ist.

> DE: *Ein Kratzen*. <> ES: *Como si alguien estuviese arañando algo.* [0012, 3, 23].
>
> DE: *Dann hörte sie ein Kratzen*. <> ES: *Luego oyó que arañaban la puerta*. [0037, 5, Viernes ma…].

Einige Einzelheiten bezüglich der Erstellung von Suchanfragen mit dem DO in *PaGeS* werden nun anhand der Suche nach deutschen Entsprechungen von spanischen Verbalperiphrasen verdeutlicht. Ähnlich wie im Falle der BO muss prinzipiell bei Suchanfragen mit dem DO der Befehl *[SS]* am Anfang mitgeschrieben werden, wie z. B. im Falle von *liarse a* + Infinitiv:

[SS] liar a~0

Die angeführte Suchanfrage stellt Belege bereit, die Charakterketten wie *liado a*, *liara a*, *lió a* usw. enthalten, jedoch keine mit der Charakterkette *liarse* (Infinitiv mit enklitischem Pronomen *se*). Letztere Ergebnisse erhält man ausschließlich mit folgender Suchanfrage:

[SS] liarse a~0

Diese Tatsache muss man im Hinterkopf behalten, wenn man nach Belegen von weiteren spanischen Verbalperiphrasen sucht, deren Hilfsverb ein enklitisches Pronomen enthält (z. B. *ponerse aecharse a* + *Infinitiv*, usw.), und umfangreichere Suchen nach folgendem Muster formulieren:

[SS] liar a~0 OR liarse a~0

[14] In: Christian Oehlschläger, «Sommernacht». *Auf trügerischer Spur. Jagd- und Kriminalgeschichten*, Neumann-Neudamm. Meligen, 2018, S. 158-192. Hervorhebung der Autorinnen.

Weiter ist anzumerken, dass, wenn der DO einer Charakterkette nachgestellt ist, die am Anfang den WUO enthält, der Befehl *[SS]* nicht mitgeschrieben werden darf, wie folgende Suchanfragenbeispiele nachweisen:

seguir *ndo~0 => z. Z. 4202 Ergebnisse in Originaltexten

[SS] seguir *ndo~0 => keine Ergebnisse

empezar a *er~0 => z. Z. 1189 Ergebnisse in Originaltexten

[SS] empezar a *er~0 => keine Ergebnisse

4. SCHLUSSBEMERKUNGEN

Wie aus den Ausführungen unter Punkt 3. zu entnehmen ist, weist die aktuelle Version von *PaGeS* einige Beschränkungen auf.

Die Wortartinformation (*Part-of-Speech-Tagging*) kann noch nicht in die Suche miteinbezogen werden, um komplexere Suchanfragen zu bilden, in denen z. B. zwischen Homonymen wie *Urteilen* (Substantiv im Dativ Plural) und *urteilen* (Verb) unterschieden werden muss. Satzzeichen sind z. Z. als Tokens nicht abfragbar. Ebenfalls erlaubt das System es z. Z. noch nicht, Kollokationen zu erstellen, die Ergebnisse zu sortieren oder Frequenzlisten zusammenzustellen.

PaGeS hat aber auch einige Stärken, die es als Werkzeug für Lehre und Forschung besonders nützlich machen.

PaGeS ist ein genuin zweisprachiges, bidirektionales Korpus. In den Belegen ist jederzeit erkennbar, ob Spanisch oder Deutsch dabei die Ausgangs- bzw. Zielsprache ist, und die Textressourcen sind so indexiert, dass die Benutzer bestimmen können, ob sie Belege in originalem Spanisch, originalem Deutsch, übersetztem Spanisch oder übersetztem Deutsch suchen möchten. Bei der Suche von übersetzten Belegen können Benutzer auch bestimmen, ob die Originalsprache Spanisch, Deutsch bzw. eine dritte Sprache ist (s. 3.1). Dies erlaubt es, *PaGeS* auch in bestimmten Fällen als übersetzerisches Vergleichskorpus einzusetzen, indem man ein Phänomen in Belegen auf Originalspanisch, in Belegen auf übersetztem Spanisch aus dem Deutschen und weiter in Belegen auf übersetztem Spanisch aus einer dritten Sprache analysiert.

Die zwei Schnittstellen auf der *PaGeS* -Webseite (Einfache Suche und Erweiterte Suche) sind so konzipiert, dass Suchanfragen jeder Art und Absicht möglich sind – sowohl schnelle Anfragen als auch sorgfältig geplante, auf einen sehr konkreten Teil der textuellen Basis abzielende Anfragen. Die Abstraktheit der SolR-basierten, formellen Suchsprache ist auf ein Minimum reduziert und beschränkt sich auf Operatoren (WUO, CEO, BO,

DO, s. Sektion 3), die auch in anderen digitalen Kontexten gängig sind und somit Benutzern bekannt sein dürften, was die «Lernkurve» schnell steigen lässt.

Als zukünftige Schritte sind eine Erweiterung der Textbasis mit TED-Talks geplant, sowie neue Indexierungen, die die Wortartinformation und die Information zur Alignierung auf Wortartniveau enthalten. Weiter ist für die nächsten Monate die Veröffentlichung von *PaEnS* geplant, des Schwester-Projekts von *PaGeS,* mit Englisch und Spanisch als Suchsprachen.

LITERATURVERZEICHNIS

Doval, Irene, «POS-tagging a Bilingual Parallel Corpus: Methods and Challenges», *Research in Corpus Linguistics*, 5 (2017), S. 35-46.

Doval, Irene, «Das *PaGeS*-Korpus, ein Parallelkorpus der deutschen und spanischen Gegenwartssprache», *Revista de Filología Alemana* 26 (2018), S. 181-197. http://dx.doi.org/10.5209/RFAL.60148 [zuletzt abgerufen: 20.04.2021].

Doval, Irene; Fernández Lanza, Santiago; Jiménez Juliá, Tomás; Liste Lamas, Elsa und Lübke, Barbara, «Corpus *PaGeS*: A multifunctional resource for language learning, translation and cross-linguistic research», in Irene Doval und María Teresa Sánchez Nieto, Hrsg., *Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications*, Amsterdam, John Benjamins, 2019, S. 103-121.

Gale, William A.; Church, Kenneth W., «A program for aligning sentences in bilingual corpora», *Computational Linguistics,* 19/1 (1993), S. 75–102.

Graën, J., Batinić, D., und Volk, M., «Cleaning the Europarl corpus for linguistic applications», in *Actas de The 12th KONVENS (Konferenz zur Verarbeitung Natürlicher Sprache), Hildesheim, 8 Oktober 2014 - 10 Oktober 2014*.https://www.zora.uzh.ch/id/eprint/99005/1/Cleaning_the_Europarl_Corpus_for_Linguistic_Applications.pdf [zuletzt abgerufen: 20.04.2021].

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

Koehn, Philipp, «EuroParl. A parallel corpus for statistical machine translation». *Proceedings of the machine translation summit*, Phuket, Thailand, 2005, S. 79–86, http://www.statmt.org/europarl/ [zuletzt abgerufen: 20.04.2021].

Lemnitzer, Lothar und Zinsmeister, Heike, *Korpuslinguistik. Eine Einführung,* Tübingen, Narr, 2015.

Sánchez-Nieto, María Teresa, «La doble interpretación aspectual de predicados en la traducción alemán-español de secuencias narrativas: análisis de un corpus de traducciones estudiantiles», *TRANS. Revista de Traductología,* 16 (2012), S. 79–99, http://uvadoc.uva.es/handle/10324/16929 [zuletzt abgerufen: 20.04.2021].

Schmid, Helmut, «Improvements in Part-of-Speech Tagging with an Application to German», in Armstrong Susan et al., Hrsg., Natural Language Processing Using Very Large Corpora. Text,

Speech and Language Technology, vol 11. Springer, Dordrecht, 1999, S. 13-25. https://doi.org/10.1007/978-94-017-2390-9_2 [zuletzt abgerufen: 20.04.2021].

Schmid, Helmut, «Probabilistic part-of-speech tagging using decision trees», in *International Conference on New Methods in Language Processing*, Manchester, UK, 1994, S. 44–49.

Tiedemann, Jörg, *Bitext Alignment*, Toronto, Morgan & Claypool, 2011.

Varga, Dániel et al., «Parallel corpora for medium density languages», *Proceedings of the RANLP,* 2007, S. 590–596.

Volk, Martin; Graën, Johannes und Callegaro, Elena, «Innovations in parallel corpus search tools», in *Ninth International Conference on Language Resources and Evaluation* (LREC'14), Reykjavik, 2014, http://www.lrec-conf.org/ proceedings /lrec2014/ pdf/504_Paper.pdf. [zuletzt abgerufen: 20.04.2021].

Zinsmeister, Heike, «Corpora», in Carstensen, Kai-Uwe et al., Hrsg., *Computerlinguistik und Sprachtechnologie: Eine Einführung,* Heidelberg, Spektrum, Akad. Verl., 3. Aufl., 2010, S. 481-492.