Katrin Menzel (Mannheim)
Heike Przybyl (Saarbrücken)
Ekaterina Lapshinova-Koltunski (Hildesheim)

EPIC-UdS – ein mehrsprachiges Korpus als Grundlage für die korpusbasierte Dolmetsch- und Übersetzungswissenschaft

Abstract: In this paper, we present examples of current research foci and results of analyses of the Collaborative Research Centre subproject "Translation as Rational Communication" that focuses on the specific linguistic properties of interpreted and translated texts distinguishing them from original productions. We describe the creation and annotation of EPIC-UdS, a multilingual corpus of simultaneous interpreting for English, German and Spanish. We give an overview of the corpus variants and explore various applications of the corpus. Building on the '*translationese*' hypothesis from translation studies, we investigate whether simultaneous interpreted language resembles written translated language with regard to specific features or whether it carries '*interpretese*' features as a result of a unique language transfer process so that traces of these features tend to occur in all simultaneously interpreted texts and distinguish them from other texts. For instance, with regard to the simplification hypothesis put forward by various translation scholars, we can observe in interpreted texts that there is a tendency towards syntactic simplification. Another analysis shows that interpreted language is characterised by a particular use of discourse particles. EPIC-UdS contains rich metadata and fine-grained linguistic annotations tailored for diverse applications across a broad range of linguistic subfields. This paper provides the first overview in German on the EPIC-UdS corpus with the aim of bringing together results from individual studies, such as Bizzoni/Teich 2019; Karakanta/Vela/Teich 2018; Lapshinova-Koltunski et al. 2021a; Lapshinova-Koltunski/Przybyl/Bizzoni 2021b; Lapshinova-Koltunski/Pollkläsener/Przybyl 2022; Przybyl/Teich 2021; Pollkläsener 2021; Przybyl et al. 2022a, 2022b, and giving a concise and informative summary of applications and impacts of the project.

Keywords: corpus-based interpreting studies, corpus compilation and annotation, spoken and simultaneously interpreted language, multilingual language resource, European Parliament data, translationese, interpretese.

1 EPIC-UdS innerhalb der Familie der European Parliament Interpreting Corpora

In diesem Beitrag stellen wir das an der Universität des Saarlandes (UdS) erstellte Dolmetschkorpus EPIC-UdS (vgl. Przybyl et al. 2022a) und verschiedene Beispiele für seine Anwendungsmöglichkeiten für die korpusbasierte Dolmetsch- und Übersetzungswissenschaft und die forschungsbasierte translationswissenschaftliche Lehre vor. EPIC-UdS mit transkribierten Texten professioneller Simultandolmetscher:innen in verschiedenen Sprachen, die unter realen Praxisbedingungen im Europäischen Parlament entstanden sind, wurde im Rahmen des Saarbrücker Sonderforschungsbereiches 1102[1] im Teilprojekt B7 – „Translation as Rational Communication" mit übersetzungs- und dolmetschwissenschaftlichem Schwerpunkt erstellt. In dem SFB-Teilprojekt, das sich inzwischen in seiner dritten 4-jährigen Förderphase befindet, werden verdolmetschte und übersetzte Texte sowie gesprochene und schriftliche Originaltexte auf spezifische sprachliche Eigenschaften und hinsichtlich ihrer sprachlichen Komplexität mit empirischen Methoden miteinander verglichen (Bizzoni/Teich 2019; Karakanta/Vela/Teich 2018; Lapshinova-Koltunski et al. 2021a; Lapshinova-Koltunski/Przybyl/Bizzoni 2021b; Lapshinova-Koltunski/Pollkläsener/Przybyl 2022; Przybyl/Teich 2021; Pollkläsener 2021; Przybyl et al. 2022a, 2022b).

EPIC-UdS steht in Zusammenhang mit der Familie der EPIC[2]-Parallelkorpora (vgl. Bernardini et al. 2018) und erweitert diese mit dem Sprachenpaar Deutsch-Englisch. Es baut auf den Erfahrungen anderer Europäischer Parlamentskorpora für gesprochene und verdolmetschte Sprache auf und kann beispielsweise in Ergänzung zum Übersetzungskorpus

1 SFB 1102 „Information Density and Linguistic Encoding", Projektnummer bei der Deutschen Forschungsgemeinschaft: 232722074, vgl. auch Crocker/Demberg/Teich (2016).

2 European Parliament Interpreting Corpus

EuroParl_UdS mit schriftlichen Redebeiträgen aus dem Europäischen Parlament und ihren Übersetzungen (vgl. Karakanta/Vela/Teich 2018) oder in Ergänzung der anderen EPIC-Korpora, die in verschiedenen Größen und für unterschiedliche Sprachkombinationen existieren, verwendet werden. Das Saarbrücker EPIC-UdS mit englischen (EN), deutschen (DE) und spanischen (ES) Texten nutzt die im Rahmen des EPIC-Projektes in Bologna (vgl. Bendazzoli/Sandrelli 2005; Russo et al. 2012) und des EPICG-Projektes in Gent (vgl. Defrancq/Plevoets/Magnifico 2015) etablierten Standards und Transkriptionsrichtlinien, um eine Vergleichbarkeit mit diesen Dolmetschkorpora zu gewährleisten, welche englische, italienische und spanische bzw. englische, französische, niederländische und spanische Texte beinhalten. Weitere EPIC-Projekte laufen in Belgrad, Lissabon und Leuven (vgl. Bernardini et al. 2018). ESIC, das EuroParl Simultaneous Interpreting Corpus (vgl. Macháček/Žilinec/Bojar 2021) enthält ebenfalls englische Originaltexte von Redebeiträgen aus dem Europäischen Parlament und neben simultanen Verdolmetschungen ins Deutsche auch Verdolmetschungen ins Tschechische. Das Hungarian European Parliamentary Intermodal Corpus (HEPIC, vgl. Götz 2020) enthält neben englischen und ungarischen Reden aus dem Europäischen Parlament auch ihre Simultanverdolmetschungen ins Ungarische bzw. Englische. Vergleichbare Daten mit Texten aus EU-Plenardebatten, die vor kurzem für das Sprachenpaar Polnisch-Englisch erstellt wurden, finden sich auch im Polish Interpreting Corpus (PINC, Poznań, vgl. Chmiel et. al 2022) und im Dolmetschkorpus EP-Poland (vgl. Bartłomiejczyk/Gumul/Koržinek 2022). An Dolmetschkorpora mit Ausgangs- und Zieltexten, die ebenfalls wie EPIC-UdS Deutsch als wichtige und häufig genutzte Arbeitssprache im Europäischen Parlament im Spektrum ihrer abgedeckten Sprachen beinhalten, sind derzeit lediglich HeiCIC, das Heidelberg Conference Interpreting Corpus (vgl. Kunz/Stoll/Klüber 2021), und das o. g. ESIC zu nennen. HeiCIC enthält Redebeiträge von Wissenschaftler:innen und anderen Fachleuten zu unterschiedlichen Themen, die von Dolmetscher:innen mit unterschiedlichem Erfahrungslevel simultan in acht Sprachen verdolmetscht wurden. Des Weiteren enthalten andere Arten von Korpora teilweise, aber nicht ausschließlich, Verdolmetschungen. Das Translation and Interpreting Corpus (TIC, vgl. Kajzer-Wietrzny 2012, S. 57–70), bei-

spielsweise, enthält neben Übersetzungen von Reden aus dem Europäischen Parlament aus vier Sprachen (Französisch, Spanisch, Niederländisch und Deutsch) ins Englische auch Simultanverdolmetschungen von Parlamentsreden aus diesen vier Sprachen ins Englische. Zudem enthält es sowohl geschriebene, nicht übersetzte als auch gesprochene, nicht-verdolmetschte englische Originalreden als Vergleichstexte. Als monolinguales Vergleichskorpus enthält TIC jedoch keine Ausgangstexte zu den englischen Übersetzungen und Verdolmetschungen. Teile von TIC, und zwar gesprochene englische Originalreden (spoken EN ORG) und Verdolmetschungen aus dem Deutschen ins Englische (SI DE EN), wurden in EPIC-UdS integriert.

Die existierenden Dolmetschkorpora mit Texten aus dem Europäischen Parlament unterscheiden sich hauptsächlich in Bezug auf die abgedeckten Sprachenkombinationen, die Art der Verdolmetschungsprozesse (ausschließlich in die Muttersprache, wie z. B. in EPIC, EPICG und EPIC-UdS, oder auch Retour-Dolmetschen, wobei die Dolmetscher:innen Texte auch von der Muttersprache in die Fremdsprache übertragen, wie z. B. in PINC oder in EP-Poland) und in den Aufbereitungsschritten der Transkriptionen, wie z. B. Time-Alignment in EPICG mit dem EXMARaLDA-System (vgl. Schmidt/Wörner 2014) oder Verfahren zur Satzalignierung wie in EPIC und EPIC-UdS und Dependency Parsing in einer speziellen Version von EPIC-UdS.

Aufnahmen von professionellen Verdolmetschungen und ihren Ausgangstexten sind für zahlreiche Sprachen am ehesten im Bereich Politik verfügbar, während Daten aus anderen Registern häufig schwieriger oder nur sehr eingeschränkt zur Verfügung stehen. Daher enthalten die meisten existierenden Dolmetschkorpora, wie auch EPIC-UdS, Texte zu politischen Themen. Für einige Sprachkombinationen, z. B. Japanisch-Englisch oder Chinesisch-Englisch, existieren auch andere Arten von Dolmetschkorpora mit Texten aus anderen Bereichen, z. B. das Baidu Speech Translation Corpus (BSTC, vgl. Zhang et al. 2021) und das Dolmetschkorpus des Nara Institute of Science and Technology (NAIST, vgl. Doi/Sudoh/Nakamura 2021). Die meisten Dolmetschkorpora enthalten simultan verdolmetschte Texte. Lin/Liang (vgl. 2023) haben neben Simultanverdol-

metschungen aus dem Chinesischen ins Englische auch konsekutiv verdolmetschte Texte in ein Korpus integriert, um die unterschiedlichen Dolmetscharten miteinander zu vergleichen. Da der Aufwand zur Erstellung von qualitativ hochwertigen Dolmetschkorpora wie auch bei anderen gesprochensprachlichen Daten durch die Anforderungen an die meist von Hand erstellten Transkriptionen sehr hoch ist, sind die verfügbaren Dolmetschkorpora im Vergleich zu anderen Korpora typischerweise nicht allzu groß (z. B. EP-Poland mit ca. 150.000 Tokens[3], die aus ca. 20 Stunden Aufnahmen von Ausgangs- und Zieltexten transkribiert wurden). EPIC-UdS ist eines der größten verfügbaren Dolmetschkorpora, s. Abschnitt 2.

EPIC-UdS als eines der wenigen Dolmetschkorpora mit Deutsch und mit seinen über mehrere Jahre weiterentwickelten Korpusversionen eröffnet neue Perspektiven und kann für sich oder in Verbindung mit EuroParl_UdS zur Bearbeitung einer Reihe von Fragen der kontrastiven Linguistik sowie der Übersetzungs- und Dolmetschwissenschaft genutzt werden, beispielsweise für die Analyse von spezifischen Eigenschaften von Verdolmetschungen (‚*Interpretese*') (vgl. Lapshinova-Koltunski et al. 2021a). Darauf aufbauend ließe sich EPIC-UdS auch im Vergleich zu konsekutiv verdolmetschten Texten heranziehen, um zu überprüfen, wie sich die unterschiedlichen Produktionsbedingungen auf die linguistischen Merkmale der auf verschiedene Weise verdolmetschten Texte auswirken (vgl. Lin/Liang 2023). EPIC-UdS ist auch für praktische Übungen in der Übersetzungs- und Dolmetschlehre einsetzbar.

3 Wörter und ggf. Satzzeichen

2 Korpuserstellung, Metadaten, Annotationen und Versionen von EPIC-UdS

EPIC-UdS gehört zu den umfangreichsten verfügbaren Dolmetschkorpora mit ca. 348.626 Tokens (vgl. Przybyl et al. 2022a). Tabelle 1 verdeutlicht die aktuelle Größe der EPIC-UdS-Korpusversion V2 und dessen Subkorpora.

EPIC-UdS enthält aufbereitete Daten aus dem Zeitraum 2008–2013 von englischen, deutschen und spanischen Redebeiträgen aus dem Europäischen Parlament und ihren Verdolmetschungen (EN ↔ DE, ES → EN), s. Tabelle 1. Teile des Korpus basieren auf Verdolmetschungen aus bereits existierenden Korpora. Verdolmetschungen aus dem Deutschen und dem Spanischen ins Englische sowie englischsprachige Originaltexte konnten aus dem o. g. TIC nach einer Überarbeitung im Hinblick auf die für EPIC-UdS verwendeten Transkriptionsrichtlinien übernommen werden. Weitere englischsprachige Ausgangstexte wurden aus dem o. g. EPICG in EPIC-UdS integriert. Die anderen Komponenten für das Deutsche und Spanische wurden für EPIC-UdS neu transkribiert. Die Verwendung einheitlicher Transkriptionsrichtlinien für alle Texte in EPIC-UdS basierend auf EPICG (vgl. Bernardini et al. 2018) gewährleistet die Vergleichbarkeit innerhalb der EPIC-Familie. Die Transkriptionen enthalten typische Merkmale der gesprochenen Sprache.

EPIC-UdS-Subkorpus	Anzahl Tokens	Anzahl der satzwertigen Äußerungseinheiten	Anzahl der Redebeiträge
ORG SP EN (Originalreden Englisch)	67.526	3.622	137
SI EN > DE (Verdolmetschung EN → DE)	57.532	4.076	137
ORG SP DE (Originalreden Deutsch)	56.488	3.409	165

SI DE > EN (Verdolmetschung DE → EN)	58.503	3.623	163
ORG SP ES (Originalreden Spanisch)	53.947	2.537	162
SI ES > EN (Verdolmetschung ES → EN)	54.630	3.076	163
∑	348.626		

Tab. 1: Korpusgröße EPIC-UdS V2[4]

Dazu gehören z. B. unvollständige Äußerungen, wiederholte oder abgebrochene Wörter (z. B. *amoun/* statt *amount)*, Versprecher (*plemary [plenary]*), Pausen (*/*) und Verzögerungssignale (z. B. *euh, hm*). Alle Transkriptionen, Überarbeitungen und Segmentierungen wurden jeweils von einer sprachwissenschaftlichen Hilfskraft durchgeführt und von einer zweiten überprüft. Die Ausgangstexte und ihre Verdolmetschungen wurden in EPIC-UdS zusätzlich (im Unterschied zu EPICG) in satzwertige Äußerungseinheiten segmentiert, die jeweils aus einem einfachen Hauptsatz oder einem Satzgefüge aus Hauptsatz und seinen eventuellen Nebensätzen bestehen. Satzreihen aus mehreren Hauptsätzen wurden in mehrere satzwertige Einheiten segmentiert. In die Transkriptionen wurden keine Satzzeichen eingefügt. Bei der satzalignierten Korpusversion wurden den satzwertigen Segmenten aus dem Ausgangstext die entsprechenden Strukturen im Zieltext zugeordnet.

Detaillierte Textmetadaten wie beispielsweise Sprechgeschwindigkeit, Name des Redners bzw. der Rednerin des Ausgangstextes oder Nationalität ermöglichen gezielte Suchabfragen nach bestimmten außersprachlichen Kriterien. Bei der Korpuserstellung wurde berücksichtigt, dass die in der EU verwendeten Sprachen teilweise plurizentrische Spra-

4 SI = ‚Simultaneous interpreting', ORG = ‚Original', SP = ‚Spoken'

chen, also Sprachen mit mehreren nationalen Zentren und dort kodifizierten Standardvarietäten sind. Die deutschen Ausgangstexte in EPIC-UdS stammen daher nicht ausschließlich von Abgeordneten aus Deutschland, sondern auch von Abgeordneten aus anderen EU-Mitgliedstaaten, in denen Deutsch eine regionale oder nationale Solo- oder Ko-Amtssprache ist. Die Sprecher:innen der deutschen Ausgangstexte haben in den meisten Fällen die deutsche Nationalität, aber auch Beiträge von österreichischen Redner:innen und in geringer Zahl auch von luxemburgischen, italienischen oder belgischen Parlamentsabgeordneten, die sich im Europäischen Parlament auf Deutsch mit muttersprachlichem Niveau äußern, liegen in den Daten vor. Die Sprecher:innen der englischen Ausgangstexte in EPIC-UdS haben in den meisten Fällen die britische Nationalität und in einigen Fällen auch die irische. Es lassen sich daher beispielsweise gezielt Untersuchungen machen zu Texten von deutschen oder britischen Redner:innen oder zu einer breiteren Vielfalt an Texten von Sprecher:innen aus EU-Ländern mit Deutsch bzw. Englisch als Amtssprache, um zu untersuchen, wie Deutsch und Englisch von Muttersprachler:innen in der Sprecher:innengemeinschaft der europäischen Parlamentsabgeordneten insgesamt benutzt werden. Spezifische Daten zur Identität und Nationalität der Dolmetscher:innen liegen nicht vor. Anhand der Aufnahmen wurden die Stimmen der Dolmetscher:innen analysiert, sodass zumindest mit einer gewissen Sicherheit festgestellt werden kann, welche Texte oder Textabschnitte von den gleichen oder unterschiedlichen Dolmetscher:innen verdolmetscht wurden.

Von EPIC-UdS existieren mehrere Korpusvarianten zur Verwendung für unterschiedliche linguistische Fragestellungen. Die Korpusversionen von EPIC-UdS können auch über CQPWeb[5] zu Forschungs- und Lehrzwecken genutzt werden. Ein Beispiel für eine Suche nach Konstruktionen bestehend aus einem als Eigennamen (*Nomen proprium*) und einem als regulären Nomen getaggten Token (z. B. *Lisbon strategy, EU regulation…*) könnte lauten: [xpos="NNP.*"][xpos="NN.*"] oder [upos="PROPN"]

5 https://corpora.clarin-d.uni-saarland.de/cqpweb/

[upos="NOUN"][6]. Die CQP-basierten Abfragen von EPIC-UdS erlauben eine große Vielfalt weiterer spezifischer Analysen.

Die Korpusversionen unterscheiden sich beispielsweise in der Art der Annotationen oder hinsichtlich der Beibehaltung aller oder nur bestimmter gesprochensprachlicher Merkmale, beispielsweise zur Verbesserung des Wortarten-Taggings oder anderer NLP[7]-Anwendungen wie das Parsing. Tabelle 2 enthält eine Übersicht über die in der Version EPIC-UdS V2 vorhandenen Metadaten und Annotationen.

Arten von Metadaten und Annotationen	**Details**
Informationen über Sprecher:in des Ausgangstextes	Name, Staatsangehörigkeit, Geschlecht, Muttersprache
Informationen über Dolmetscher:in	Geschlecht, Muttersprache (= Zielsprache)
Vortragsart	abgelesen, freie Rede, gemischt
Sprechgeschwindigkeit in Wörtern pro Minute	langsam (< 131), mittel (131–160), hoch (> 160)
Textthema (und Fachbereich) der parlamentarischen Plenardebatten	z. B. *Aussprache zum Statut der Europäischen Privatgesellschaft (Recht)* oder *Aussprache zu Roséweinen und zugelassenen önologischen Verfahren (Landwirtschaft)*
Textlänge	Länge der Aufnahme in Sekunden und Länge der Transkription in Tokens
Entstehungszeitraum der Texte	2008–2013

6 Mit *xpos* lassen sich sprachspezifisch getaggte Wortarten abfragen, mit *upos* sprachübergreifend vergleichbare Wortarten.

7 Natural Language Processing

Tokenisierung und Lemmatisierung	spaCy (in EPIC-UdS V3 mit Stanza)
Wortarten-Annotationen	zwei verschiedene POS-Annotationen: sprachspezifische Wortart-Kategorien, z. B. STTS (vgl. Schiller et al. 1999) für das Deutsche wie in EuroParl_UdS, und sprachübergreifende Wortart-Kategorien unter Verwendung des Universal Dependencies POS Tagsets
Parsing	Kopf-Dependent-Beziehungen mit spaCy (in EPIC-UdS V3 mit Stanza)
Segmentierung	satzwertige Äußerungseinheiten

Tab. 2: EPIC-UdS V2: Annotationen und Metadaten

Die Transkripte wurden segmentweise mit spaCy NLP Tools Version 2.3.4 und den entsprechenden Language Models (de_core_news_lg-2.3.0, en_core_web_lg-2.3.1, es_core_news_lg-2.3.1) für die Korpusversion V2 geparst. Die Daten liegen im CoNLL-U-Format vor, sind mit Universal POS-Tags und sprachspezifischen POS-Tags sowie den Tags der Universal Dependencies versehen. Für EPIC-UdS Version V3 konnte das Parsing nach dem Entfernen einiger gesprochensprachlicher Merkmale aus den Transkripten und unter Verwendung von Stanza Language Models (v1.2.3, vgl. Qi et al. 2020) weiter verbessert werden. Dadurch hat sich die Parsinggenauigkeit, gemessen am *Unlabeled Attachment Score* (UAS – Wert für korrekt identifizierte Kopf-Dependent-Beziehungen ohne Berücksichtigung der Bezeichnung) und dem *Labeled Attachment Score* (LAS – Wert für korrekt identifizierte Kopf-Dependent-Beziehungen mit Berücksichtigung der Bezeichnung) von V2 zu V3 deutlich erhöht. Die Parsinggenauigkeit konnte hierdurch von 78,73 (LAS) und 86,42 (UAS) auf 85,94 (LAS) und 89,08 (UAS) für Englisch gesteigert werden. Für das Deutsche wurde sie von 69,74 (LAS) und 76,64 (UAS) auf 85,78 (LAS) und 91,03

(UAS) erhöht (vgl. Przybyl et al. 2022a: 1197). Die spanischen Texte wurden ebenfalls geparst, sie wurden jedoch noch nicht hinsichtlich ihrer Genauigkeit evaluiert.

Die Version *EPIC-UdS (aligned, parsed, surprisal)* verfügt neben Parsingergebnissen und Alignierungen zu den jeweiligen Äußerungseinheiten in der Ausgangs- bzw. Zielsprache auch über Annotationen der *Surprisal*-Werte für jedes Token[8] als informationstheoretisches Maß für die Informationsdichte und Vorhersagbarkeit dieser linguistischen Einheiten in ihrem Kontext.

3 Verdolmetschte vs. übersetzte Reden aus dem Europäischen Parlament

Sowohl EuroParl_UdS als auch EPIC-UdS enthalten muttersprachliche Redebeiträge aus Plenarsitzungen im Europäischen Parlament. Während es sich bei EuroParl_UdS allerdings um offiziell veröffentlichte Reden, d. h. verschriftlichte und redigierte Redeprotokolle und ihre Übersetzungen mit den typischen Merkmalen der Schriftsprache handelt, finden sich in EPIC-UdS transkribierte Redebeiträge und transkribierte Verdolmetschungen für die Sprachenkombinationen Englisch ↔ Deutsch und Spanisch → Englisch, die sich am gesprochenen Wort orientieren, und daher auch Merkmale der gesprochenen Sprache wie Füllwörter, Satzabbrüche, gefüllte Pausen und Verzögerungslaute, gesprächsorganisatorisch eingesetzte Diskurspartikel, abgebrochene und wiederholte Wörter etc. beinhalten.

8 Eine weitere Korpusvariante beinhaltet *Surprisal*-Werte berechnet auf Lemmabasis.

EuroParl_ UdS	Tokens	Satz-anzahl	EPIC-UdS V3	Tokens	Anzahl satzwerti-ger Einheiten
ORG WR EN	8.693.135	372.547	ORG SP EN	68.548	3.623
TR EN > DE	3.100.647	137.813	SI EN > DE	58.218	4.080
TR EN > ES	3.781.376	125.852			
ORG WR DE	7.869.289	427.779	ORG SP DE	57.049	3.408
TR DE > EN	6.260.869	262.904	SI DE > EN	59.100	3.622
ORG WR ES	6.140.211	183.361	ORG SP ES	56.502	2.538
			SI ES > EN	57.765	3.076

Tab. 3: Korpusgrößen von EuroParl_UdS und EPIC-UdS V3[9]

Folgende Korpustextauszüge 1 a–d in Tabelle 4 verdeutlichen einige generelle Unterschiede zwischen den Texten aus EPIC-UdS und EuroParl_UdS:

9 SI = Simultaneous interpreting, ORG = Original, SP = Spoken, WR= Written TR = Translation

Beispiel 1 a: Auszug aus EPIC-UdS, deutscher Ausgangstext[10]	**Beispiel 1 b: Auszug aus EuroParl_UdS für den gleichen deutschen Ausgangstext aus der redigierten und publizierten Version**
wir wollten / Kleinstunternehmen / und dabei sprechen wir von Unternehmen / die extrem klein sind / also mit ganz wenigen Angestellten minimalen Umsatzzahlen minimalen / Gewinnzahlen die im Grunde nur im regionalen Bereich vor Ort im lokalen Bereich tätig sind / der kleine Bäckermeister der kleine Malermeister / die wollen wir von den Bilanzverpflichtungen / befreien /	*Wir wollten Kleinstunternehmen – und dabei sprechen wir von Unternehmen, die extrem klein sind, also mit ganz wenigen Angestellten, mit minimalen Umsatz- und Gewinnzahlen, die im Grunde nur im regionalen Bereich vor Ort, im lokalen Bereich tätig sind, der kleine Bäckermeister, der kleine Malermeister – von den Bilanzverpflichtungen befreien.*
Beispiel 1 c: Auszug aus EPIC-UdS, englische Verdolmetschung des Beispiels 1 a	**Beispiel 1 d: Auszug aus EPIC-UdS EuroParl_UdS, englische Übersetzung des Beispiels 1 b**
we wanted to look at micro-entities and that means / entities which are really very small with very few people working for them / minimum turnover / euh amount / minimum / profit amount / which are very / locally active / j / just a small / hm / baker or painter and decorator we want to reduce the administrative burden on those companies	*We wanted to free micro-entities – and here we are talking about companies that are particularly small, with few employees, minimum turnover and profit figures and which effectively only operate in a regional, local area, say a small baker or painter and decorator – from accounting obligations.*

Tab. 4: Korpustextbeispiele 1 a–d

Während die Beispiele 1 a und c als Segmente aus den gesprochensprachlichen Ausgangs- und Zieltexten zwar auch wie die schriftsprachlichen

10 Die Schrägstriche in den Transkriptionen markieren kurze Redepausen.

Beispiele 1 b und d eine Parenthese beinhalten, wird die Struktur des jeweiligen Hauptsatzes danach auf unterschiedliche Weise wieder aufgenommen. In den transkribierten Texten sind mehr Kohäsionsmittel notwendig, um den Zusammenhang zum Teil vor dem Einschub herzustellen. So wird in 1 a und c am Ende des Einschubs das Subjekt wieder aufgenommen („*wir*" / „*we*"). Auch der Teil des Prädikats, der am Anfang des Satzes genannt wurde, wird in der gesprochenen Sprache wieder aufgenommen bzw. umformuliert und ergänzt („*wollten […] wollen […] befreien*" / „*wanted to look at […] want to reduce […]*"). In den Segmenten aus den editierten, schriftsprachlichen Fassungen der Ausgangs- und Zieltexte 1 b und d wird der jeweilige Hauptsatz nach der Parenthese direkt fortgesetzt („*Wir wollten Kleinstunternehmen – […] – von den Bilanzverpflichtungen befreien.*" / „*We wanted to free micro-entities – […] – from accounting obligations.*").

Auch wenn der lange, den Satz unterbrechende Einschub in 1 b und d einen hohen kognitiven Aufwand beim Lesen bedeutet und es zumindest in der englischen Übersetzung etwas untypisch wirkt, eine einzelne Präpositionalphrase auf einen langen Einschub folgen zu lassen, erlauben der schriftsprachliche Produktionsmodus sowie die angenommenen Texterwartungen und -erfahrungen der Leser:innen als Rezipient:innen die Verwendung eines solchen Einschubs mit direktem Anschluss der noch nicht genannten Satzglieder des Hauptsatzes. In einer gesprochenen Rede wäre dies sehr ungewöhnlich. Zur Vermittlung neuer Informationen und optimalen Anbindung des neu vermittelten Wissens an die vorhandenen Wissensstrukturen dürfen die syntaktischen Strukturen das Arbeitsgedächtnis der Rezipient:innen nicht übermäßig belasten. Neben den schriftsprachlichen/gesprochensprachlichen Unterschieden zwischen EPIC-UdS und Euro-Parl_UdS verdeutlicht Beispiel 1 zudem, dass Verdolmetschungen gegenüber gesprochenen Originaltexten in EPIC-UdS auch gewisse Unterschiede, wie z. B. in Bezug auf inhaltlich vermittelte Informationen und verwendete Kohäsionsmittel aufweisen können. U. a. bedingt durch die generell unterschiedliche Art der Grundwortstellung in den beiden Sprachen und die Dolmetschstrategie der Antizipation von in Letztstellung erwarteten deutschen Hauptverben (vgl. Seeber 2005) wird aus „*wir wollten*

Kleinstunternehmen […] die wollen wir von den Bilanzverpflichtungen befreien“ (1 a) in der Verdolmetschung (1 c) „*we wanted to look at micro entities […] we want to reduce the administrative burden on those companies*“. In der Verdolmetschung wird hier ein zusätzliches Hauptverb (‚*to look at*‘) eingefügt. Das im Original durch das Demonstrativpronomen ‚*die*‘ wiederaufgenommene Objekt nach dem Einschub wird in der Verdolmetschung auch wiederaufgenommen, allerdings durch demonstrative Referenz *(‚those‘)* und durch lexikalische Kohäsion (‚*companies*‘ als Oberbegriff für *‚micro-entities‘*), um expliziter zur Herstellung von Kohärenz beizutragen.

EPIC-UdS und EuroParl_UdS enthalten also Texte, die sich inhaltlich sehr ähnlich sind, sich aber bezüglich ihrer sprachlichen Realisierung aufgrund des schriftlichen oder mündlichen Produktionsmodus deutlich unterscheiden.

4 Anwendungsbeispiele EPIC-UdS

In diesem Abschnitt werden Beispiele für Forschungsschwerpunkte und Analyseergebnisse aus dem entsprechenden SFB-Projekt vorgestellt, in dem EPIC-UdS erstellt wurde. Aufbauend auf der *Translationese*-Hypothese aus der Übersetzungsforschung (vgl. Olohan/Baker 2000; Shlesinger 1995; Teich 2003) gehen wir vor allem der Frage nach, ob Simultanverdolmetschungen Übersetzungen durch bestimmte Merkmale ähneln oder als Ergebnis eines speziellen Sprachtransferprozesses ‚*Interpretese*‘-Merkmale tragen, die sich tendenziell in allen verdolmetschten Texten finden lassen (vgl. Bernardini/Ferraresi/Miličević 2016; Kajzer-Wietrzny 2012, 2015; Shlesinger/Ordan 2012).

In Bezug auf die in der Übersetzungswissenschaft aufgestellte *Simplification*-Hypothese lässt sich auch hinsichtlich der Verdolmetschungen beobachten, dass tendenziell syntaktische Vereinfachungen und Umstrukturierungen beim Sprachtransfer stattfinden. Erste Pilotstudien zu *Dependency-Length*-Analysen zeigen u. a., dass in den Verdolmetschungen in EPIC-UdS kürzere syntaktische Strukturen verwendet werden als

in den Originaltexten und dass es eine Tendenz zur ‚*Dependency length minimisation*' gibt (vgl. Przybyl/Teich 2021).

Zum Vergleich von verschiedenen Wahrscheinlichkeitsverteilungen in Bezug auf Vokabular und grammatische Strukturen in den Korpora EPIC-UdS und EuroParl_UdS sowie deren Subkorpora wurden wortbasierte Unigram-Sprachmodelle für das informationstheoretische Maß der relativen Entropie (Kullback-Leibler-Divergenz, KLD) erstellt (vgl. Przybyl et al. 2022b). Die KLD-Analyse der EPIC-UdS Daten im Vergleich zu EuroParl_UdS misst die Distanz zwischen N-Gramm-Modellen für die jeweiligen Korpora als Unähnlichkeitsmaß für die Wahrscheinlichkeitsverteilungen, hier insbesondere über das Vokabular der Verteilungen. Es wird dabei gemessen, wie gut eine Verteilung die andere approximiert, und es lassen sich Aussagen über den Abstand zwischen den Wahrscheinlichkeitsverteilungen machen. Je größer die Kullback-Leibler-Divergenz ist, desto unterschiedlicher sind beide N-Gramm-Modelle. Durch KLD-Analysen lässt sich u. a. feststellen, dass Übersetzungen offenbar stärkere konzeptionell schriftsprachliche Merkmale als vergleichbare schriftliche Originaltexte aufweisen. Verdolmetschungen hingegen scheinen mehr konzeptionell gesprochensprachliche Merkmale zu enthalten als vergleichbare gesprochene Originaltexte, z. B. kürzere Äußerungseinheiten und mehr gefüllte Pausen und abgebrochene Wörter als in vergleichbaren Originalreden (vgl. Przybyl et al. 2022a, 2022b).

Bei der Untersuchung der Übersetzungen im Vergleich zu den Verdolmetschungen aus dem Deutschen ins Englische konnte weiterhin nachgewiesen werden, dass die Verdolmetschungen mehr distinktive Verzögerungslaute (z. B. *hm*), Diskursmarker (*well*), personaldeiktische und andere deiktische Ausdrücke (*we, there*), Adverbien der Verstärkung und Fokussierung (*really*), Modal- und Hilfsverben und deren Kontraktionen, u. a. bei Verneinungen (*we've, haven't, don't, shouldn't*), beinhalten. Die Übersetzungen hingegen scheinen mehr geprägt zu sein von lexikalischen Verben (*involves, represents*), Adverbien (*instead, currently, firstly*) und einem stärkeren Nominalstil, worauf wir aufgrund der distinktiveren Nutzung von Artikeln (*the*), Demonstrativpronomen (*this*), Präpositionen (*for, of*) und Nomen, die auch im Durchschnitt länger als in den Verdolmetschungen sind (*regulations, democrats*), schließen können (s. Abb. 1).

Abb. 1: KLD-Visualisierung für distinktive Merkmale englischer Übersetzungen (links) und Verdolmetschungen (rechts), markiert durch die jeweilige Größe der Einheiten, wobei die Farbe Aufschluss über deren relative Häufigkeit gibt (Przybyl et al. 2022b: 199)

Ein weiteres Anwendungsbeispiel ist die automatische Textklassifizierung. In Merkmalsanalysen anhand von Ergebnissen aus der automatischen Textklassifizierung lassen sich Texte durch die Unterscheidungsmerkmale, die jeweils für Verdolmetschungen oder Übersetzungen spezifisch sind, klassifizieren. Hierbei haben wir teilweise Ähnlichkeiten bei den Merkmalen, durch die sich Verdolmetschungen und Übersetzungen von vergleichbaren Ausgangstexten unterscheiden, festgestellt. Merkmale, die sowohl für Verdolmetschungen als auch Übersetzungen relevant sind, haben allerdings bei den Verdolmetschungen stets eine höhere Gewichtung (vgl. Lapshinova-Koltunski et al. 2021a). Verdolmetschungen als Ergebnis einer sehr spezifischen Kommunikationsform lassen sich insgesamt von allen anderen Textarten sehr gut mittels Textklassifizierung unterscheiden (s. Abb. 2).

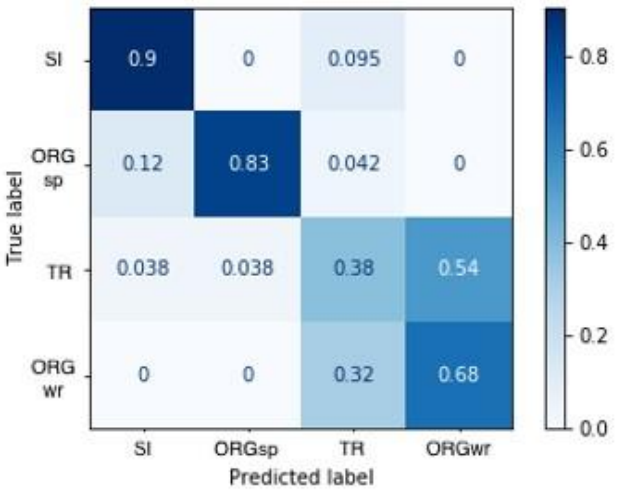

Abb. 2: Konfusionsmatrix für die Korpora Simultanverdolmetschungen (SI), gesprochene Originaltexte (ORGsp), Übersetzungen (TR) und geschriebene Originaltexte (ORGwr)[11] (vgl. Lapshinova-Koltunski et al. 2021a: 87)

Es gibt zudem einige kontextuelle Unterschiede, was das Vorkommen bestimmter Merkmale in Verdolmetschungen und Übersetzungen betrifft (s. Beispiele 2 a und b in Tabelle 5).

11 Der Genauigkeitsgrad wird durch Werte zwischen 0 und 1 verdeutlicht.

Beispiel 2 a:
Auszug aus EPIC-UdS, SI DE > EN
*and euh **obviously** fair trade is the foundation of Europe's prosperity.*
Beispiel 2 b:
Auszug aus EuroParl-UdS TR DE > EN
*If, in this day and age, the very same pesticide attracts 20 % value added tax in one Member State and just 3 % in another, whilst fully complying with the sixth directive on value added tax, something is **obviously** wrong!*

Tab. 5: Korpustextbeispiele 2 a–b

In Beispiel 2 a aus einem verdolmetschten Text wird das Adverb ‚*obviously*' als Satzadverb mit der Funktion eines Diskursmarkers am Anfang der Äußerungseinheit verwendet, während es in dem Auszug aus der Übersetzung 2 b modifizierend vor einem prädikativen Adjektiv verwendet wird. Dieses Beispiel verdeutlicht, dass sich die Kontexte bestimmter Merkmale, welche für beide Textformen relevant sind, unterscheiden, und zwar vermutlich durch die Effekte der geschriebenen bzw. gesprochensprachlichen Form. Eine weitere Analyse zeigt typische Nutzungen von Diskurspartikeln in Verdolmetschungen im Vergleich zu den Originaltexten auf (vgl. Pollkläsener 2021; Lapshinova-Koltunski/Pollkläsener/Przybyl 2022). Einige Verwendungen von ‚*well*' und ‚*now*' beispielsweise waren speziell in verdolmetschten Texten zu finden.

Bizzoni/Teich (vgl. 2019) haben lexikalische Einheiten in den EuroParl_UdS-Übersetzungen und den EPIC-UdS-Verdolmetschungen mit Word-Embedding-Methoden auf der Grundlage neuronaler Netze untersucht. Unterschiede zwischen den jeweiligen semantischen Räumen waren weniger dichte Wort-Cluster bei den Verdolmetschungen als in den Übersetzungen und Unterschiede in Bezug auf die Verteilung von spezifischem und allgemeinerem Vokabular. Lapshinova-Koltunski/Przybyl/Bizzoni (2021b) nutzten ebenfalls spezifische Word-Embeddings und machten prinzipiell ähnliche Beobachtungen wie Bizzoni/Teich (2019). Bei der Untersuchung von Diskursmarkern und Konnektoren in EuroParl_UdS-Übersetzungen und EPIC-UdS-Verdolmetschungen zeigte sich auch, dass in Verdolmetschungen eine geringe Vielfalt an Konnektoren verwendet wird. Diese scheinen in konsistenteren Kontexten verwendet zu werden als

in Übersetzungen. Die Verdolmetschungen waren zudem von mehr Implizierungen gekennzeichnet als Übersetzungen (vgl. Lapshinova-Koltunski/Przybyl/Bizzoni 2021b).

In dem Erklärungsversuch für die *Translationese*-Phänomene liegt ein Schwerpunkt auf Analysen der mit *Surprisal*-Werten annotierten Korpusversion von EPIC-UdS, um Rückschlüsse auf die kognitiven Prozesse beim Dolmetschen gemessen an Informationsdichte/*Surprisal* im Zusammenhang mit den linguistischen Erfahrungen der Dolmetscher:innen in ihrem Gedächtnis zu ziehen (‚*Memory-Surprisal Tradeoff*', vgl. Hahn/ Degen/Futrell 2021). Die *Surprisal*-Werte dienen als Indikator für den kognitiven Aufwand und die Verarbeitungsschwierigkeiten in Bezug auf Ausgangstexte und die Produktion von Zieltexten. Demnach sind Wörter mit hoher Vorhersagbarkeit durch niedrige *Surprisal*-Werte gekennzeichnet und mit geringeren Verarbeitungsschwierigkeiten verbunden. Beim Übersetzen und speziell beim Dolmetschen gibt es bestimmte kognitiv belastende Faktoren und Ressourcenbeschränkungen (z. B. durch Zeitdruck), aber auch spezielle Strategien der Textproduktion, welche sich auf die gewählte Form der Sprache auswirken und der Grund für Implizierungs- und Explizierungsprozesse sein können (vgl. Teich/Martínez-Martínez/Karakanta 2020).

Die Beispiele 3 bis 5 zeigen einige Segmente aus einer Konkordanzsuche in der satzalignierten Korpusversion *EPIC-UdS DE* (*aligned, parsed, surprisal*) für eine Abfrage der unterschiedlichen Formen des Lemmas ‚*müssen*', welches das häufigste Verb in deutschen Originalreden in den Daten ist. Für die jeweiligen Wörter im deutschen Ausgangssegment und im verdolmetschten Zielsegment sind hier auch die *Surprisal*-Werte angegeben.

Beispiel 3 a: ORG_SP_DE_000038:													
und	*da*	*kann*	*ich*	*nur*	*sagen*	*wir*	*müssen*	*EULEX*	*stärken*				
2,52	4,78	5,53	1,61	5,25	0,83	4,30	4,13	12,79	13,56				
Beispiel 3 b: SI_DE_EN_000038:													
and	*I*	*can*	*only*	*say*	*that*	*we*	*have*	*got*	*to*	*try*	*to*	*support*	*EULEX*
2,30	2,35	8,55	3,94	9,52	2,23	3,07	1,00	6,48	1,00	6,88	0,56	7,46	13,68
Beispiel 4 a: ORG_SP_DE_000042:													
das	*muss*	*man*	*unseren*	*Freunden*	*in*	*Israel*	*in*	*aller*	*Klarheit*	*sagen*			
4,12	5,56	3,17	11,42	15,29	5,68	10,32	5,68	11,29	13,88	9,17			
Beispiel 4 b: SI_DE_EN_000042:													
we	*have*	*to*	*tell*	*our*	*friends*	*in*	*Israel*	*that*	*very*	*clearly*			
3,67	2,73	1,20	8,95	8,11	9,95	0,82	0,41	5,55	8,75	7,79			
Beispiel 5 a: ORG_SP_DE_000139 :													
Europa	*muss*	*sich*	*hier*	*aktiv*	*einmischen*								
10,94	2,56	5,46	5,78	9,57	15,30								
Beispiel 5 b: SI_DE_EN_000139:													
we	*have*	*to*	*actively*	*intervene*									
3,67	2,73	1,20	11,76	5,66									

Tab. 6: Korpustextbeispiele 3–5

In den Verdolmetschungen finden sich verschiedene bedeutungsgleiche Entsprechungen für dieses Verb. Oft wird nicht ‚*must*' verwendet, sondern ‚*have (got) to*', aber auch ‚*need to*' oder ‚*should*' mit abschwächendem Aufforderungscharakter. In einigen Fällen ist auch eine Abschwächung durch explizitere Formulierungen zu beobachten, die im Durchschnitt vergleichsweise niedrige *Surprisal*-Werte auszeichnen (z. B. „*...müssen EULEX stärken*", „*...have got to try to support EULEX*"). Das Subjekt ändert sich in zahlreichen Fällen im Zusammenhang mit dieser Art von Modalverben in den Verdolmetschungen hin zu einem personaldeiktischen Ausdruck (*man → we, Europa → we)*. Speziell wenn ein Nomen durch ein Pronomen ersetzt wird, sinkt typischerweise auch der *Surprisal*-Wert für das Subjekt. Dolmetscher:innen produzieren also keine Zieltexte mit dem gleichen Grad an durchschnittlichen *Surprisal*-Werten, wie sie in den Ausgangstexten vorzufinden sind.

Weitere Analysen für zusätzliche Anwendungsgebiete, wie z. B. die maschinelle Übersetzung, sind im Rahmen des mit EPIC-UdS in Verbindung stehenden Projektes durchgeführt worden (s. Bizzoni et al. 2020). Derzeit werden im Projekt detailliert die *Surprisal*-Werte von Konnektoren in Ausgangstexten, Übersetzungen und Verdolmetschungen untersucht. Die Konnektoren aus den Ausgangstexten entsprechen in den Zieltexten Explizierungen, Äquivalenzen oder Implizierungen. Eine Abfrage nach ‚*but*' beispielsweise, die in den alignierten Segmenten einem ‚*aber*' entspricht, wäre ein Fall von Äquivalenz, wenn stärker markierte Formen wie z. B. ‚*jedoch*', ‚*allerdings*', ‚*dennoch*' in der deutschen Übersetzung oder Verdolmetschung im alignierten Segment anstelle von ‚*but*' vorkommen, kann dies als ein Fall von Explizierung gesehen werden. Zu den Distributionen können wir u. a. auch die *Surprisal*-Werte extrahieren und somit beobachten, ob Explizierungen ‚überraschend' oder eher erwartbar für die Rezipient:innen sind, und ob es diesbezüglich ähnliche Tendenzen in Übersetzungen und Verdolmetschungen gibt. Dabei wird das Verhältnis zwischen *Surprisal*-Werten (auf Lemmabasis) in Ausgangs- und Zieltexten untersucht, um die spezifischen Eigenschaften übersetzter und verdolmetschter Texte als Ergebnisse rationaler Kommunikationshandlungen zu erklären.

5 Zusammenfassung

Da EPIC-UdS eines der wenigen Dolmetschkorpora mit Deutsch ist und über mehrere Jahre unter Berücksichtigung hoher Qualitätsstandards entwickelt wurde, eröffnet es neue Perspektiven für kontrastive Studien sowie die Übersetzungs- und Dolmetschwissenschaft und ist auch für praktische Übungen in der Übersetzungs- und Dolmetschlehre einsetzbar. Die Student:innen in sprach- und übersetzungswissenschaftlichen Studiengängen der Universität des Saarlandes nutzen das Korpus als Arbeitsgrundlage für Analysen in Seminaren und zur Beantwortung von eigenen Forschungsfragen im Rahmen von Abschlussarbeiten.

In diesem Artikel wurde ein zusammenfassender Überblick über EPIC-UdS und hierzu bisher veröffentlichte Studien gegeben. Neben Details zur Korpuserstellung, den Metadaten und den Annotationen wurden in diesem Artikel auch die Eigenschaften der unterschiedlichen Korpusversionen von EPIC-UdS beschrieben und verschiedene Anwendungsmöglichkeiten, Analysemethoden und Forschungsfelder aufgezeigt.

Quellenverzeichnis

BARTŁOMIEJCZYK, Magdalena / GUMUL, Ewa / KORŽINEK, Danijel (2022): „EP-Poland: Building a Bilingual Parallel Corpus for Interpreting Research". In: *GEMA, Online Journal of Language Studies* 22/1, S. 110–126.

BENDAZZOLI, Claudio / SANDRELLI, Annalisa (2005): „An Approach to Corpus-based Interpreting Studies: Developing EPIC (European Parliament Interpreting Corpus)". In: GERZYMISCH-ARBOGAST, Heidrun / NAUERT, Sandra (Hrsg.): *Mutra2005 – Challenges of Multidimensional Translation. Proceedings of the Marie Curie Euroconferences. Saarbrücken. Mai 2005.* URL: <https://www.euroconferences.info/ proceedings/2005_Proceedings/ 2005_proceedings.html> (26.05.2024).

BERNARDINI, Silvia / FERRARESI, Adriano / MILIČEVIĆ, Maja (2016): „From EPIC to EPTIC – Exploring Simplification in Interpreting and Translation from an Intermodal Perspective". In: *Target* 28, S. 61–86.

BERNARDINI, Silvia / FERRARESI, Adriano / RUSSO, Mariachiara / COLLARD, Camille / DEFRANCQ, Bart (2018): „Building Interpreting and Intermodal Corpora: A How-to for a Formidable Task". In: RUSSO, Mariachiara / BENDAZZOLI, Claudio / DEFRANCQ, Bart (Hrsg.): *Making Way in Corpus-based Interpreting Studies*. Singapur: Springer Nature, S. 21–42.

BIZZONI, Yuri / TEICH, Elke (2019): „Analyzing Variation in Translation through Neural Semantic Spaces". In: SHAROFF, Serge / ZWEIGENBAUM, Pierre / RAPP, Reinhard (Hrsg.): *Proceedings of the 12th Workshop on Building and Using Comparable Corpora (BUCC) at RANLP- 2019, Varna, Bulgaria*. Association for Computational Linguistics, S. 1–8.

BIZZONI, Yuri / JUZEK, Tom S. / ESPAÑA-BONET, Cristina / CHOWDHURY, Koel Dutta / VAN GENABITH, Josef / TEICH, Elke (2020): „How Human is Machine Translationese? Comparing Human and Machine Translations of Text and Speech". In: FEDERICO, Marcello / WAIBEL, Alex / KNIGHT, Kevin / NAKAMURA, Satoshi / NEY, Hermann / NIEHUES, Jan / STÜKER, Sebastian / WU, Dekai / MARIANI, Joseph / YVON, Francois (Hrsg.): *Proceedings of the 17th International Conference on Spoken Language Translation.* Association for Computational Linguistics, S. 280–290.

CHMIEL, Agnieszka / KORŽINEK, Danijel / KAJZER-WIETRZNY, Marta / JANIKOWSKI, Przemysław / JAKUBOWSKI, Dariusz / POLAKOWSKA, Dominika (2022): „Fluency Parameters in the Polish Interpreting Corpus (PINC)". In: KAJZER-WIETRZNY, Marta / FERRARESI, Adriano / IVASKA, Ilmari / BERNARDINI, Silvia (Hrsg.): *Mediated discourse at the European Parliament: Empirical Investigations.* Berlin: Language Science Press, S. 63–91.

CROCKER, Matthew W. / DEMBERG, Vera / TEICH, Elke (2016): „Information Density and Linguistic Encoding (IDeaL)". In: *Künstliche Intelligenz* 30, S. 77–81.

DEFRANCQ, Bart / PLEVOETS, Koen / MAGNIFICO, Cédric (2015): „Connective Items in Interpreting and Translation: Where do they come from?" In: ROMERO-TRILLO, Jesús (Hrsg.): *Yearbook of Corpus Linguistics and Pragmatics 2015: Current Approaches to Discourse and Translation Studies.* Cham: Springer International Publishing, S. 195–222.

DOI, Kosuke / SUDOH, Katsuhito / NAKAMURA, Satoshi (2021): „Large-Scale English-Japanese Simultaneous Interpretation Corpus: Construction and Analyses with Sentence-Aligned Data". In: FEDERICO, Marcello / WAIBEL, Alex / COSTA-JUSSÀ, Marta R. / NIEHUES, Jan / STÜKER, Sebastian / SALESKY, Elizabeth (Hrsg.): *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021), Bangkok.* Association for Computational Linguistics, S. 226–235.

GÖTZ, Andrea (2020): „Discourse Markers and Connectives in Interpreted Hungarian Discourse: A Corpus-based Investigation of Discourse Properties and their Interdependence". In: *Beszédtudomány – Speech Science* 1, S. 259–284.

HAHN, Michael / DEGEN, Judith / FUTRELL, Richard (2021): „Modeling Word and Morpheme Order in Natural Language as an Efficient Trade-off of Memory and Surprisal". In: *Psychological Review* 128/4, S. 726–756.

KAJZER-WIETRZNY, Marta (2012): *Interpreting Universals and Interpreting Style.* Dissertation, Adam-Mickiewicz-Universität Poznań.

KAJZER-WIETRZNY, Marta (2015): „Simplification in Interpreting and Translation". In: *Across Languages and Cultures* 16/2, S. 233–255.

KARAKANTA, Alina / VELA, Mihaela / TEICH, Elke (2018): „EuroParl_UdS: Preserving and Extending Metadata in Parliamentary Debates". In: FIŠER, Darja / ESKEVICH, Maria / DE JONG, Franciska (Hrsg.): *Proceedings of ParlaCLARIN workshop, LREC2018, Miyazaki, Japan, 2018.* URL: <http://lrec-conf.org/workshops/lrec2018/W2/summaries/10_W2.html> (26.05.2024), Korpus verfügbar unter: <https://fedora.clarin-d.uni-saarland.de/europarl-uds/>; <https://corpora.clarin-d.uni-saarland.de/cqpweb/> (26.05.2024).

KUNZ, Kerstin / STOLL, Christoph / KLÜBER, Eva (2021): „HeiCiC: A Simultaneous Interpreting Corpus Combining Product and Pre-process Data". In: BIZZONI, Yuri / TEICH, Elke / ESPAÑA-BONET, Cristina / VAN GENABITH, Josef (Hrsg.): *Proceed-ings*

of the Workshop on Modelling Translation: Translatology in the Digital Age (MoTra21). *NoDaLiDa*, *31. Mai 2021*, S. 8–14.

LAPSHINOVA-KOLTUNSKI, Ekaterina / BIZZONI, Yuri / PRZYBYL, Heike / TEICH, Elke (2021a): „Found in Translation / Interpreting: Combining Data-driven and Supervised Methods to Analyse Cross-linguistically Mediated Communication". In: BIZZONI, Yuri / TEICH, Elke / ESPAÑA-BONET, Cristina / VAN GENABITH, Josef (Hrsg.): *Proceedings of the Workshop on Modelling Translation: Translatology in the Digital Age (MoTra21). NoDaLiDa, 31. Mai 2021*, S. 82–90.

LAPSHINOVA-KOLTUNSKI, Ekaterina / PRZYBYL, Heike / BIZZONI, Yuri (2021b): „Tracing Variation in Discourse Connectives in Translation and Interpreting through Neural Semantic Spaces". In: BRAUD, Chloé / HARDMEIER, Christian / LI, Junyi Jessy / LOUIS, Annie / STRUBE, Michael / ZELDES, Amir (Hrsg.): *Proceedings of the 2nd Workshop on Computational Approaches to Discourse, Punta Cana*. Association for Computational Linguistics, S. 134–142.

LAPSHINOVA-KOLTUNSKI, Ekaterina / POLLKLÄSENER, Christina / PRZYBYL, Heike (2022): „Exploring Explicitation and Implicitation in Parallel Interpreting and Translation Corpora". In: *The Prague Bulletin of Mathematical Linguistics* 119, S. 5–22.

LIN, Yumeng / LIANG, Junying (2023): „Informativeness across Interpreting Types: Implications for Language Shifts under Cognitive Load". In: *Entropy* 25/2, S. 243.

MACHÁČEK, Dominik / ŽILINEC, Matúš / BOJAR, Ondřej (2021): ESIC 1.0 – Europarl Simultaneous Interpreting Corpus, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. URL: <http://hdl.handle.net/11234/1-3719> (26.05.2024).

OLOHAN, Maeve / BAKER, Mona (2000): „Reporting *that* in Translated English: Evidence for Subconscious Processes of Explicitation?" In: *Across Languages and Cultures* 1, S. 141–158.

POLLKLÄSENER, Christina (2021): „A Comparison of Discourse Particles in English Original and Simultaneous Interpreted Speeches". In: *Book of Abstracts of the 3rd International Conference on Translation, Interpreting & Cognition, Forlì, 2.–5. November 2021*, S. 76.

PRZYBYL, Heike / TEICH, Elke (2021): „Dependency Length Minimization in Simultaneous interpreting". In: *Book of Abstracts of the 3rd International Conference on Translation, Interpreting & Cognition, Forlì, 2.–5. November 2021*, S. 45.

PRZYBYL, Heike / LAPSHINOVA-KOLTUNSKI, Ekaterina / MENZEL, Katrin / FISCHER, Stefan / TEICH, Elke (2022a): „EPIC-UdS – Creation and Applications of a Simultaneous Interpreting Corpus". In: CALZOLARI, Nicoletta / BÉCHET, Frédéric / BLACHE, Philippe / CHOUKRI, Khalid / CIERI, Christopher / DECLERCK, Thierry / GOGGI, Sara / ISAHARA, Hitoshi / MAEGAARD, Bente / MARIANI, Joseph / MAZO, Hélène / ODIJK, Jan / PIPERIDIS, Stelios (Hrsg.): *Proceedings of the 13th Confer-ence*

on Language Resources and Evaluation (LREC 2022), Marseille, Frankreich, 20.–25. Juni 2022, S. 1193–1200. Korpus verfügbar unter: <https://fedora.clarin-d.uni-saarland.de/epic-uds/index.html; https://corpora.clarin-d.uni-saarland.de/cqpweb/> (26.05.2024).

PRZYBYL, Heike / MENZEL, Katrin / KARAKANTA, Alina / TEICH, Elke / FISCHER, Stefan (2022b): „Exploring Linguistic Variation in Mediated Discourse: Translation vs. Interpreting". In: KAJZER-WIETRZNY, Marta / FERRARESI, Adriano / IVASKA, Ilmari / BERNARDINI, Silvia (Hrsg.): *Mediated discourse at the European Parliament: Empirical Investigations.* Berlin: Language Science Press, S. 191–218.

QI, Peng / ZHANG, Yuhao / ZHANG, Yuhui / BOLTON, Jason / MANNING, Christopher D. (2020): „Stanza: A Python Natural Language Processing Toolkit for Many Human Languages". In: CELIKYILMAZ, Asli / WEN, Tsung-Hsien (Hrsg.): *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* Association for Computational Linguistics, S. 101–108.

RUSSO, Mariachiara / BENDAZZOLI, Claudio / SANDRELLI Annalisa / SPINOLO, Nicoletta (2012): „The European Parliament Interpreting Corpus (EPIC): Implementation and Developments". In: STRANIERO SERGIO, Francesco / FALBO, Caterina (Hrsg.): *Breaking Ground in Corpus-Based Interpreting Studies*. Bern: Peter Lang, S. 53–90.

SCHILLER, Anne / TEUFEL, Simone / STÖCKERT, Christine / THIELEN, Christine (1999): „Guidelines für das Tagging deutscher Textcorpora mit STTS. (Kleines und großes Tagset)". Universität Stuttgart, Institut für maschinelle Sprachverarbeitung; Universität Tübingen, Seminar für Sprachwissenschaft. URL: <http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf> (26.05.2024).

SCHMIDT, Thomas / WÖRNER, Kai (2014): „EXMARaLDA". In: DURAND, Jacques / GUT, Ulrike / KRISTOFFERSEN, Gjert (Hrsg.): *The Oxford Handbook of Corpus Phonology.* Oxford: OUP, S. 402–419.

SEEBER, Kilian G. (2005): „Temporale Aspekte der Antizipation beim Simultandolmetschen von SOV-Strukturen aus dem Deutschen". In: *Vereinigung für angewandte Linguistik in der Schweiz (VALS-ASLA)* 81, S. 123–140.

SHLESINGER, Miriam (1995): „Shifts in Cohesion in Simultaneous Interpreting". In: *The Translator* 1, S. 193–214.

SHLESINGER, Miriam / ORDAN, Noam (2012): „More spoken or more translated? Exploring a known unknown of simultaneous interpreting". In: *Target* 24, S. 43–60.

TEICH, Elke (2003): *Cross-Linguistic Variation in System and Text. A Methodology for the Investigation of Translations and Comparable Texts*. Berlin: Mouton de Gruyter.

TEICH, Elke / MARTÍNEZ-MARTÍNEZ, José / KARAKANTA, Alina (2020): „Translation, information theory and cognition". In: ALVES, Fabio / JAKOBSEN, Arnt Lykke (Hrsg.): *The Routledge Handbook of Translation and Cognition.* London: Routledge, S. 360–375.

UNIVERSAL POS TAGS (o. D.): URL: <https://universaldependencies.org/u/pos/ (26.05.2024)>.

ZHANG, Ruiqing / WANG, Xiyang / ZHANG, Chuanqiang / HE, Zhongjun / WU, Hua / LI, Zhi / WANG, Haifeng / CHEN, Ying / LI, Qinfei (2021): „BSTC: A Large-Scale Chinese-English Speech Translation Dataset“. In: WU, Hua / CHERRY, Colin / HUANG, Liang / HE, Zhongjun / LIU, Qun / ELBAYAD, Maha / LIBERMAN, Mark / WANG, Haifeng / MA, Mingbo / ZHANG, Ruiqing (Hrsg.): *Proceedings of the Second Workshop on Automatic Simultaneous Translation*. Association for Computational Linguistics, S. 28–35.