HeiCIC: A simultaneous interpreting corpus combining product and pre-process data

Kerstin Kunz
Ruprecht-Karls-Universität Heidelberg
kerstin.kunz@iued.uni-heidelberg.de

Christoph Stoll
Ruprecht-Karls-Universität Heidelberg
christoph.stoll@iued.uni-heidelberg.de

Eva Klüber
Ruprecht-Karls-Universität Heidelberg
eva.klueber@iued.uni-heidelberg.de

Abstract

This paper presents HeiCIC, a simultaneous interpreting corpus that comprises audio files, time-aligned transcripts and corresponding preparation material complemented by annotation layers. The corpus serves the pursuit of a range of research questions focusing on strategic cognitive load management and its effects on the interpreting output. One research objective is the analysis of semantic transfer as a function of problem triggers in the source text which represent potential cognitive load peaks. Another research approach correlates problem triggers with solution cues in the visual support material used by interpreters in the booth. Interpreting strategies based on this priming reduce cognitive load during SI.

1 Motivation

The aim of this paper is twofold: We present the architecture and on-going collation of a series of simultaneous interpreting (SI) subcorpora, integrated in the Heidelberg Conference Interpreting Corpus (HeiCIC): HeiCIC contains authentic speeches from LSP domains with simultaneous interpretations by learners and professionals in eight languages. The English-German core corpus is aligned with pre-process data that visualize the established conference interpreting workflow.

The pre-process data we are interested in is the visual support material which is used by interpreters to cope with expected problem triggers (PT) in a speech and to avoid peaks in cognitive processing.

Conference Interpreters are trained to condense the logical structures and PT of source texts as cues to target text solutions using a special note-taking technique for consecutive interpreting. The result of the pre-process PT analysis for simultaneous interpreting is noted in a similar fashion: as an amalgamation of source text logic, terminology and cues for cognitive load relief in a visio-spatial structure or 'map' of the thought processes (Gile, 1995; Stoll, 2009).

More precisely, this map combines expected source language macrostructures, conceptual relations and terminology with cues to trigger target language structures with cognitive load-relieving interpreting strategies. These include memory relief, listening analysis and comprehension relief, patterns for target text production and strategies for easier output monitoring using top-down and bottom-up plausibility checks (Gile, 1995; Stoll, 2009).

Furthermore, we introduce the research in progress to be done on the core corpus: Our on-going research has two objectives: a) analysing semantic transfer from source to target text in relation to expected problem triggers in the source text and b) correlating semantic transfer with pre-process data to determine which features reflect high performance SI strategies. In this way, our empirical research combines product- and process-related studies.

There are several aspects that set the corpus apart from other SI corpora: To date, no large, comparative learner/professional LSP corpus exists for SI, least for the language combinations in focus here. There are some learner corpora for Chinese <-> English, such as the learner corpus from Leung and Yip containing interpretations of nine trainees (Bendazzoli, 2018; Leung and Yip, 2013; Zhang, 2017), which are however rather limited in size. Existing professional interpreter corpora are larger but differ in terms of metadata: For instance, EPIC, EPTIC and EPICG (Bernardini et al., 2018) focus on interpreting in the institutional setting of the European Union and therefore are rather heterogeneous in terms of topic, register and level of technicality. NAIST (Japanese - English) (Neubig et al., 2018), (387,000 word and comparable to HeiCIC in size) reflects interpreting environments for a general/non-expert audience. Other SI corpora incorporate other forms of interpreting such as SIREN, which includes simultaneous interpreting with text and television interpreting in English and Russian in its 33.55h (235,040 words) of records (Dayter, 2018).

HeiCIC is designed to map authentic professional settings, where the highly technical nature of LSP and scientific conferences requires a structured, partially automated workflow for knowledge acquisition, content organization and terminology management. Our corpus design is unique in that it aligns this pre-process data with both original speeches and interpreting output. This permits insights into advanced interpreting strategies used in LSP settings and thus process-related phenomena, while other corpora typically focus on product data (Gile, 2002; Díaz Galaz, 2015).

2 Data collection and corpus design

HeiCIC is collated mainly at the Heidelberg Conferences: scientists and experts present their research in a variety of LSP domains and send preparation material, which is used by interpreters with different levels of expertise (students at MA level from the second to the final semester, young and seasoned professionals) to prepare and then interpret from, into and between German, English, French, Italian, Spanish, Portuguese, Russian and Japanese. Subcorpora differ in terms of formats available, languages included, LSP domains covered and level of interpreter expertise.

The core corpus is a homogeneous subpart containing several parallel interpretations by students, professionals with different levels of interpreter expertise, and transcripts (English <-> German) in selected LSP domains such as electrical engineering in car manufacturing, astronomy, investor relations and annual general meetings (AGMs) of international corporations. It currently contains recorded speeches and interpretations of around 83 hours with transcripts comprising around 400,000 tokens and is constantly expanded as new recordings, transcripts and annotation layers are added.

We seek to follow basic principles of corpus compilation (Bernardini et al., 2018; Hansen-Schirra et al., 2012). Metadata are stored in a separate file for each transcript. They are structured as follows: information about speaker (e.g. gender, role, native language and language variety), interpreter (e.g. gender, level of expertise, native language and language combination) and text (e.g. setting, language, register, topic and mode, text length in seconds and tokens) and allow for filtering according to these criteria.

In addition, transcripts, recordings and annotation layers are aligned with strategic pre-process data of interpreters. Pre-process data, which includes visual preparation material created by interpreters, is available in an electronic format and attributed to the individual interpreter, target and source text combination.

2.1 Transcription

The transcription process used to provide the transcripts as a basis for analysis includes several steps and is partially automated. Transcripts are generated automatically using automatic speech recognition and corrected by manual revision.

We apply transcription guidelines which are a slightly modified version of those for the GECCo Corpus (Kunz et al., 2011; Lapshinova et al., 2012). They include tags accounting for spoken language features (such as non-standard language, truncated or repeated words), tags related to cognitive load in general (such as filled and silent pauses), and tags related to SI in particular (such as interpreter turns, incomplete sentences and grammatical errors), (Plevoets and Defrancq, 2016). For instance, Example 1 shows tags for truncated words and phrases and fillers.

> […] In case of a mosquito bite, [t=or a malaria] malaria [t=is] [ehm] [t=can] is supposed to be the case. […]

Example 1. Transcription tags for spoken language features.

Revised transcripts are automatically time-aligned with the audio signal using WebMAUS (Kisler et al., 2017). The resulting files are further processed with EXMARaLDA. In combination with the time-aligned transcripts, this allows for alignment of several interpretations with one original speech (Schmidt and Wörner, 2014).

2.2 Annotation and alignment

The core part of the corpus contains automatic basic level annotations, such as tokenization, lemmatization and POS tagging. The performance of the latter is improved via additional renderings during transcription (see above example). In addition, semi-automatic and manual annotation layers are added in alignment with the current research objectives (see more details below). Main annotations include information on problem triggers in the source text and on semantic transfer between source and target text. Manual annotation steps are performed by several annotators. Each source text is currently annotated by two skilled student annotators. We regularly evaluate annotator agreement to ensure high annotation quality and to improve detailed annotation guidelines.

In order to analyse correlations between process and product data, we include several alignments: Problem triggers in the source text are aligned with respective renderings in the visual support material and corresponding expressions in the target texts. Moreover, solution cues marked in the visual support material are related to indicators of interpreting strategies in the target text.

3 Problem triggers

In a first step we annotate source texts for problem triggers representing potential cognitive load peaks in original texts (Gile, 2009). We focus on "problem triggers pertaining to the message", as classified by Mankauskienė (2016: 146). This type is structured further into categories such as numbers, proper nouns, collocations, terminology and complex phrases. Sender-related problem triggers (e.g. accent) or technical problem triggers can be integrated at later stages of the project. We currently implement procedures for semi-automatic extraction and manual post-correction for some of these categories (e.g. terminology and numbers). Other categories, e.g. complex phrases are annotated manually. Double annotation is possible, meaning that one source text element can incorporate several problem triggers.

4 Semantic transfer

In a second step, (non-)renderings corresponding to problem triggers are identified in the respective target texts and grouped into transfer categories specifying their relation to the source text problem trigger. Transfer categories focus on semantic relations with category options determined by the problem trigger category.

This serves as a basis for the analysis of semantic transfer from source to target text, i.e. the reproduction of a message uttered in one language into another (Schjoldager, 1995). Problem trigger renderings are not analysed in isolation, but within the units of meaning in which they occur to allow for a more comprehensive analysis of semantic transfer from source to target text. For this purpose,

interpreting units are identified in both source and target text based on functional, semantic and syntactical information (Alves et al., 2019; Christoffels and de Groot, 2005).

Semantic transfer is defined as the relation between source and target interpreting units on a scale from omission and implicitation to explicitation and addition and analysed by assessing features contained in the interpreting units in terms of their structure and their semantic content (Becher, 2011; Hansen-Schirra et al., 2012). The semantic content is categorised in terms of explicitness: Words (or expressions) that can potentially encode a higher semantic range than others are classified as less explicit than words (or expressions) that have a narrower semantic range (Gumul, 2017). Semantic transfer may be encoded using different means, for example substitution such as pronouns or hyponyms or hypernyms in the target text in relation to the source text segment. Examples 2 and 3 show instances of the semantic transfer categories implicitation by substitution and omission of part of a segment.

	source text	target text	semantic transfer
47	Why is this tiredness warning system useful?	Wieso ist dies hilfreich?	implicitation

Example 2. Semantic transfer: implicitation.

	source text	target text	semantic transfer
43	In other words, you can remain in the navigation system or rate your list view and still change the driving mode for the car at the push of a button.	Man kann beispielsweise während des Navigationsmodus den Effizienzmodus einschalten auf Tasterdruck.	omission

Example 3. Semantic transfer: omission.

The focus of analysis lies on interpreting units that contain problem triggers as they potentially provide insights into the effect of cognitive load peaks on semantic transfer (Mankauskienė, 2016). Shifts in the position of interpreting units within sentence and text structures are analysed as well.

Previous studies on SI have focused either on individual transfer phenomena such as explicitation or on linguistic features such as cohesion markers (Kajzer-Wietrzny, 2012). To our knowledge, a comprehensive analysis of the semantic content of interpreting units and transfer categories in combination with the analysis of information structure has not been attempted so far.

5 Visual support material

In a third step, the properties of the interpretation output are correlated with pre-process data: visual support materials prepared by interpreters as a substantial part of the interpretation workflow.

As widely agreed in research on simultaneous interpreting, conference preparation goes beyond the bilingual organization of terminology and glossaries, notably in alphabetical order (Rütten, 2007; Will, 2009). Visual support material ideally combines information on expected content with organizations of concepts and terminology (Stoll, 2009 and 2019). It contains chronological renderings of expected macrotopics reflecting textual function and skopos. Macrotopics are complemented on the microlevel as ontological representations of concepts (i.e. semantic relations and semantic roles) and mapped onto terminological expressions.

Furthermore, these visio-spatial maps integrate simultaneous interpreting strategies, i.e. strategy cues relating predictions of source language problem triggers such as cognitive load conflicts and overruns (Seeber, 2011-17) to efficient target language solutions (Stoll, 2019). Some examples are structures related to listening comprehension enhancing anticipation/priming of collocations, complex syntactic structures and terminology.

For instance, the source text cue revenue in an earnings release event semantically primes the hypernym, KPI (key performance indicators for corporations) and other co-hyponyms such as earnings and profit. The target language solutions (“Umsatz, Absatz, Ertrag”) are directly

linked to the semantic priming by the cue 'revenue' in the visual support material (cue map). Shortcuts from consecutive note taking are used to indicate such semantic relations.

Speech production and monitoring effort relief strategies in the visual map use domain specific jargon compression, e.g. Luftwiderstands-Beiwert (“aerodynamic drag coefficient”) is rendered as “drag”. Other strategies replace complex syntactic structures by prosodic and cohesive elements.

These electronic maps of pre-process thoughts are mind-map-like multidimensional structures that tap into the interpreter skillset: layout patterns and symbols from consecutive note-taking in relational databases, xml structures, spread sheets, and multi-layered documents bear tangible and - correlatable testimony to the categories of cognition moved upstream in the interpreting workflow in several dimensions: In keeping with professional practice, conceptual and terminological information is combined into a single structure with different views for pre- and in-process phases (Stoll, 2009; Fantinuoli, 2012): While the pre-process view shapes terminology and expert knowledge into an ontological hierarchy (Rütten, 2007; Will, 2009), the in-process view lists macrotopics, semantic relations, terminology and strategy cues in chronological order. Thus, visual support material used in the booth is a condensed in-process version of the pre-process map (Stoll, 2009). The level of condensation may vary, depending on the level of expertise and familiarity with the topic and register.

6 Correlating product and process data

Our approach aims to determine which features in visual support materials used in the booth can be identified as solution cues and therefore indicators of deliberate high-performance SI strategies as they correlate with the interpreter’s output, thus proving process in product features. Correlating problem triggers in the source text with semantic transfer categories and thus interpreting output on the one hand, and with entries in the support material on the other hand, should yield information as to how predictions of source language problem triggers are marked and strategically related to efficient target language solution cues. They may then be assigned to individual types of cognitive load, as mentioned above. Moreover, our analyses may reveal whether and how these entries in the visual support material relate to solutions in the interpretation output. In this, we invert the traditional errors-and-omissions-based approach to establish an evidence-based, hierarchical typology of verifiable strategies of semantic, conceptual, lexical and strategic priming.

Insights obtained may serve to optimize the organization of electronic visual support material in general and improve CAI tools for in-process use, contributing to augmented interpretation.

We plan to make our corpus accessible for corpus-querying via a web interface such as CQPWeb for independent validation, validity and reliability of our research. The corpus is well documented to permit research beyond our current focus in the future.

References

Fabio Alves, Adriana Pagano, Stella Neumann, Erich Steiner and Sylvia Hansen-Schirra. 2019. Translation units and grammatical shifts. David B. Sawyer, Frank Austermühl and Vanessa Enríquez Raído. (Eds.) *The Evolving Curriculum in Interpreter and Translator Education, American Translators Association Scholarly Monograph* Series, XV. John Benjamins Publishing Company, Amsterdam. 109–142.

Viktor Becher. 2011. *Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts.* Hamburg, Universität Hamburg.

Claudio Bendazzoli. 2018. Corpus-based Interpreting Studies: Past, Present and Future Developments of a (Wired) Cottage Industry. Mariachiara, Claudio Bendazzoli, Bart Defrancq (Eds.) *Making way in corpus-based interpreting studies. New Frontiers in Translation Studies*. Springer, Singapore. 1–20.

Silvia Bernardini, Adriano Ferraresi, Mariachiara Russo, Camille Collard and Bart Defrancq. 2018. Building Interpreting and Intermodal Corpora: A How-to for a Formidable Task. Mariachiara Russo,

Claudio Bendalozzi, and Bart Defrancq. (Ed.) *Making Way in Corpus-based Interpreting Studies*. Singapore, Springer.

Ingrid K. Christoffels and Annette M. B. de Groot. 2005. Simultaneous interpreting: A cognitive perspective. Judith F. and Annette M. B. de Groot. (Ed.). *Handbook of Bilingualism: Psycholinguistic Approaches*. New York, Oxford University Press. 454–479.

Daria Dayter. 2018. Describing lexical patterns in simultaneously interpreted discourse in a parallel aligned corpus of Russian-English interpreting (SIREN). *FORUM* 16:2. 241–264.

Stephanie Díaz-Galaz. 2015. *La influencia del conocimiento previo en la interpretación simultánea de discursos especializados: Un estudio empírico*. PhD thesis, Universidad de Granada.

Claudio Fantinuoli. 2012. *InterpretBank - Design and Implementation of a Terminology and Knowledge Management Software for Conference Interpreters*. Berlin, epubli GmbH.

Daniel Gile. 2002. The Interpreter's Preparation for Technical Conferences: Methodological Questions in Investigating the Topic. *Conference Interpretation and Translation* 4:2. 7-27.

Ewa Gumul. 2017. Explicitation and directionality in simultaneous interpreting. *Linguistica Silesiana*, 2017. 311-329.

Sylvia Hansen-Schirra, Stella Neumann and Erich Steiner. 2012. *Cross-linguistic corpora for the Study of Translation: Insights from the Language-Pair English German*. Berlin, de Gruyter.

Marta Kajzer-Wietrzny. 2012. *Interpreting universals and interpreting style*. PhD dissertation, Adam Mickiewicz University.

Cynthia Jane Mary Kellett Bidoli. 2016. Methodological challenges in Consecutive Interpreting Research: Corpus analysis of notes. Claudio Bendazzoli and Claudia Monacelli. (Ed.) *Addressing methodological challenges in Interpreting Studies Research*. Newcastle upon Tyne, Cambridge Scholars. 141-169.

Thomas Kisler, Uwe Reichel and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45: 326–347.

Ekaterina Lapshinova-Koltunski. Kerstin Kunz and Marilisa Amoia. 2012. Compiling a Multilingual Spoken Corpus. *Spoken Corpora and Annotation; Speech Technology and Data Bases. Proceedings of the VIIth GSCP International Conference*. Firence, Firence University Press.

Leung, S.M. Ester and Leonard Yip. 2013. *A bilingual corpus of interpreting students' performance*. http://arts.hkbu.edu.hk/~engester/main.html.

Dalia Mankauskienė. 2016. Problem trigger classification and its applications for empirical research. *Procedia - Social and Behavioral Sciences* 231. 143 – 148

Graham Neubig, Hiroaki Shimizu, Sakriani Sakti, Satoshi Nakamura and Tomoki Toda. 2018. The NAIST Simultaneous Translation Corpus. Mariachiara Russo, Claudio Bendalozzi and Bart Defrancq. (Ed.) *Making Way in Corpus-based Interpreting Studies*. Singapore, Springer.

Koen Plevoets and Bart Defrancq. 2016. The effect of informational load on disfluencies in interpreting: A corpus-based regression analysis. *Translation and Interpreting Studies*, 11 (2): 202-224.

Anja Rütten. 2007. *Information and Knowledge Management in Conference Interpreting (in German)*. Frankfurt, Lang.

Anne Schjoldager. 1995. An Exploratory Study of Translational Norms in Simultaneous Interpreting: Methodological Reflections. *Hermes, Journal of Linguistics*, 8 (14): 65-88.

Thomas Schmidt and Kai Wörner. 2014. EXMARaLDA. *Handbook on Corpus Phonology*. Oxford, Oxford University Press. 402-419.

Kilian Seeber. 2011. Cognitive load in simultaneous interpreting. Existing theories – new models. *Interpreting*, 13 (2): 176-204.

Kilian Seeber. 2013. Cognitive load in simultaneous interpreting: Measures and methods. *Target*, 25 (1): 18-32.

Kilian Seeber. 2017. Multimodal processing in simultaneous interpreting. John W. Schwieter and Aline Ferreira. (Ed.) *The Handbook of translation and cognition*. New Jersey, Wiley Blackwell.

Christoph Stoll. 2002. Dolmetschen und neue Technologien. Joanna Best and Sylvia Kalina. (Ed.) *Übersetzen und Dolmetschen. Eine Orientierungshilfe*. Tübingen, Francke. 307-312.

Christoph Stoll. 2009. *Jenseits simultanfähiger Terminologiesysteme. Methoden der Vorverlagerung von Kognition im Arbeitsverlauf professioneller Konferenzdolmetscher*. Trier, WVT.

Christoph Stoll. 2019. Terminology Systems and Workflow Automation for Simultaneous Interpreters: CAI tools and Research within the HeiCiC Corpus (in German). *edition*, 1/2019. 25-33.

Martin Will. 2009. *Interpreting-Oriented Terminology Work (in German).* Tübingen, Narr.

Wei Zhang. 2017. Chinese interpreting learner corpus construction and research: Theory and practice (in Chinese). *Chinese Translators Journal*, 38 (1): 53-60