Interpreting and technology

Edited by

Claudio Fantinuoli

Translation and Multilingual Natural Language Processing 42

language science press

Translation and Multilingual Natural Language Processing

Editors: Oliver Czulo (Universität Leipzig), Silvia Hansen-Schirra (Johannes Gutenberg-Universität Mainz), Reinhard Rapp (Johannes Gutenberg-Universität Mainz)

In this series:

1. Fantinuoli, Claudio & Federico Zanettin (eds.). New directions in corpus-based translation studies.
2. Hansen-Schirra, Silvia & Sambor Grucza (eds.). Eyetracking and Applied Linguistics.
3. Neumann, Stella, Oliver Čulo & Silvia Hansen-Schirra (eds.). Annotation, exploitation and evaluation of parallel corpora: TC3 I.
4. Czulo, Oliver & Silvia Hansen-Schirra (eds.). Crossroads between Contrastive Linguistics, Translation Studies and Machine Translation: TC3 II.
5. Rehm, Georg, Felix Sasaki, Daniel Stein & Andreas Witt (eds.). Language technologies for a multilingual Europe: TC3 III.
6. Menzel, Katrin, Ekaterina Lapshinova-Koltunski & Kerstin Anna Kunz (eds.). New perspectives on cohesion and coherence: Implications for translation.
7. Hansen-Schirra, Silvia, Oliver Czulo & Sascha Hofmann (eds). Empirical modelling of translation and interpreting.
8. Svoboda, Tomáš, Łucja Biel & Krzysztof Łoboda (eds.). Quality aspects in institutional translation.

ISSN: 2364-8899

Interpreting and technology

Edited by

Claudio Fantinuoli

language
science
press

Claudio Fantinuoli (ed.). 2018. *Interpreting and technology* (Translation and Multilingual Natural Language Processing 42). Berlin: Language Science Press.

This title can be downloaded at:
http://langsci-press.org/catalog/book/209

ISBN: no digital ISBN
no print ISBNs!
ISSN: 2364-8899
no DOI
Source code available from www.github.com/langsci/209
Collaborative reading: paperhive.org/documents/remote?type=langsci&id=209

Cover and concept of design: Ulrike Harbort
Typesetting: Change typesetter in localmetadata.tex
Fonts: Linux Libertine, Libertinus Math, Arimo, DejaVu Sans Mono
Typesetting software: XeLaTeX

Language Science Press
Unter den Linden 6
10099 Berlin, Germany
langsci-press.org

Storage and cataloguing done by FU Berlin

Contents

Chapter 1

An exploratory study on CAI tools in simultaneous interpreting: Theoretical framework and stimulus validation

Bianca Prandi
University of Mainz

The acquisition of terminology and specialized knowledge prior to a technical conference represents a fundamental phase in the interpreter's workflow, but quick and easy access to terminological information during the interpreting task is equally important to support the interpreter in the rendition of terminology and to ensure a high-quality interpreting performance. The processing of terminology during simultaneous interpreting represents a sub-task that can be a cause of saturation and lead to errors or omissions. In order to facilitate the processing of terminology and avoid cognitive overload, interpreters resort to various terminology management solutions.

Over the past few years, terminology management tools have been developed specifically for interpreters. Known as Computer-Aided Interpreting (CAI) tools, they are designed to support interpreters in their entire workflow while keeping the additional cognitive load deriving from the terminology search during interpreting as low as possible. The role of CAI tools is now being acknowledged by practitioners and interpreting scholars, but the impact of such tools on the cognitive processes involved in simultaneous interpreting is still unclear. A project underway at the University of Mainz/Germersheim aims at bridging this knowledge gap by adopting an empirical approach.

To this end, an exploratory study was conducted to evaluate the appropriateness of the stimuli adopted for data collection and to verify whether the use of CAI tools causes saturation or, on the contrary, helps prevent it by reducing the local cognitive load during terminology search and delivery of the target text. In particular, the terminology search is expected to be quicker and more effective than for other solutions and the added cognitive load to be lower. In this paper, after describing

the theoretical framework adopted in the study and our hypotheses, we present the structure of the pilot study and the stimuli used. We then validate the stimuli used and present the first results derived from the analysis of the test subjects' renditions.

1 Introduction

Computer-Assisted Interpreting emerged around 10 years ago to provide interpreters with tools to prepare for specialized events and to support them along the individual phases of their workflow, from preparation, to interpretation proper, to follow-up work after the assignment. CAI tools thus rationalize the interpreter's terminology work by making preparation more efficient and ultimately aim at improving the quality of the interpreter's output, at least in terms of terminological precision and adequacy. **Rütten2007** and **Will2009** developed a theoretical model of the interpreter's preparation work and laid the foundations for how a CAI tool should be structured in order to address the specific needs of conference interpreters, which are mainly linked to the online nature of interpreting and the time constraints it entails.

To date, the number of CAI tools available to interpreters is limited and their functionalities do not always cover all the phases of the interpreting process. **Fantinuoli2018** distinguishes between first-generation and second-generation CAI tools. The first (e.g. Interplex[1] and Terminus[2]) are "designed to manage multilingual glossaries in an interpreter-friendly manner" (**Fantinuoli2018**), but do not offer an advanced search algorithm. The latter "offer advanced functionalities that go beyond basic terminology management, such as features to organize textual material, retrieve information from corpora or other resources, learn conceptualized domains, and advanced search functions" (**Fantinuoli2018**) and include Intragloss[3] and InterpretBank[4]. Interpreter's Help[5] can also be considered a second-generation CAI tool, as it implements and advanced search function through its companion tool Boothmate[6].

Given the relatively recent introduction of these tools on the market and their varied nature, first attempts at an evaluation have been made. Two main trends can be identified in this respect. The most recent one focuses on developing a set

[1] http://www.fourwillows.com/interplex.html
[2] http://www.wintringham.ch/cgi/ayawp.pl?T=terminus
[3] http://intragloss.com
[4] http://www.interpretbank.com
[5] https://interpretershelp.com
[6] https://interpretershelp.com/boothmate

of criteria against which the tools can be evaluated (**Costa2016**; **Will2015**). This approach is certainly ambitious, but it remains somewhat arbitrary. The evaluation criteria mainly reflect the features offered by the tools, but do not consider how they influence the product of the interpreting process in terms of terminological quality and whether they optimize the interpreters' preparation and facilitate their work in the booth, by making the online retrieval of terminological units easier and improving the terminological quality. While the number and type of features of CAI tools certainly is of interest for practitioners, we would argue that the main reason for choosing to use CAI tools, and to prefer one tool to the other ones available, should be the ability of such tools to positively influence the interpreter's work in terms of cognitive capacity and, ultimately, quality.

This is where the second trend in the evaluation of CAI tools comes into play. Soon after the development of the first CAI tools, the first studies on the topic appeared. Apart from a few Master's theses, which are limited in scope and often take a descriptive approach, rather than an investigative one (see for example **DeMerulis2013**), a rather small number of publications can be found which mostly deal with the application of CAI tools to the preparation phase (**Xu2015**; **Fantinuoli2017a**). When it comes to the use of CAI tools in the booth, the number of studies is very limited, as is their scope. First attempts at an empirical analysis of the use of CAI tools during SI can be identified in **Prandi2015a**; **Prandi2015b** and Biagini **Biagini2015**. These initial investigations of the issue speak in favor of the usability of CAI tools and seem to suggest that they do improve terminological quality of simultaneous interpreting (SI). Both experiments were based on a product-oriented analysis of the test subjects' deliveries. Biagini also included a statistical analysis of transcription data. Apart from these initial analyses, no empirical methodology has been tested in a wide-ranging experiment which implements psycho-physiological, process-oriented methods in addition to product-based analysis. In addition, **Fantinuoli2017b** recently addressed the topic of the integration of automatic speech recognition (ASR) in CAI tools for the use in the booth.

A PhD research project underway at the Johannes Gutenberg University of Mainz/Germersheim (**Prandi2016**; **Prandi2017a**; **Prandi2017b**) aims at bridging this research gap. By triangulating pupillometry data, eye-tracking data and the analysis of the test subjects' transcriptions, we aim at providing a picture not only of the usability of CAI tools during simultaneous interpreting, but also of the local variations in Cognitive Load (CL) and of the terminological quality of simultaneous interpretation performed with the support of a CAI tool when compared to more traditional terminology management solutions. Through our study, we

hope to develop a research methodology that can be used to evaluate CAI tools and provide the much-needed empirical data that will be helpful not only to practitioners in choosing the best tool, but also to software developers, by highlighting potential shortcomings.

After discussing the theoretical framework of our analysis (§2), we present our research desiderata and the structure of an exploratory study conducted to test our research methodology (§3). §4 describes the rationale behind the stimuli used in the experiment and the features of the speeches used. We then present the results of the analysis of the transcriptions (§5), which we use to evaluate the appropriateness of our stimuli. In the conclusions, we address future work and provide suggestions for further research.

2 Theoretical framework

In our investigation of simultaneous interpreting performed with the support of CAI tools, our aim is not only to look at the product of such activity, but also at the process that lies behind it. For this reason, in establishing a theoretical framework for our analysis, we took into consideration the two theoretical models that set out to describe interpreting from a procedural point of view and that address the allocation of cognitive sources during this very complex mental activity: Gile's Effort Model (EM) of SI and Seeber's Cognitive Load Model (CLM) of Simultaneous Interpreting. In this section, we discuss why Seeber's approach is more suited to the operationalization of our hypotheses.

2.1 Gile's Effort Model and Seeber's Cognitive Load Model of Simultaneous Interpreting

The main point of divergence between Gile's Effort Model of Simultaneous Interpreting (**Gile1988**; **Gile1997**; **Gile1999**) and Seeber's Cognitive Load Model of SI lies in the theoretical assumptions they stem from. Gile draws from Kahneman's single resource theory (**Kahneman1973**), which does not find much validation in scientific literature. If there is one single pool of resources interpreters can adopt, how can some interpreters perform a terminological search on the Internet, while at the same time delivering a perfectly acceptable rendition of the original speech? This kind of multi-tasking might seem impossible to a first-year interpreter trainee, but is commonly observed among experienced professional interpreters. The second controversial assumption is that interpreters work close to saturation level most of the time (**Gile1999**'s "tightrope hypothesis", **Gile1999**).

While this might be true in some cases, for instance when the source speech is particularly dense, fast, or pronounced with a non-native accent, there might very well be cases in which the interpreter has enough spare cognitive resources to do something else while interpreting.

In his Cognitive Load Model of Simultaneous Interpreting, Seeber takes an opposite approach to Gile's, basing his model on Wickens's multiple resource theory and on his Cognitive Load Model (**Wickens1984**; **Wickens2002**). Wickens developed his model to account for the fact that qualitative differences in tasks being performed at the same time lead to "differences in time-sharing efficiency" (**Wickens2002**), as shown by **Kantowitz1976** and Wickens himself (**Wickens1976**). According to Wickens, different kinds of tasks require resources that are managed by discrete structures. When two or more tasks are performed simultaneously and "all other things [are] equal (i.e. equal resource demand or single task difficulty), two tasks that both demand one level of a given dimension (e.g. two tasks demanding visual perception) will interfere with each other more than two tasks that demand separate levels on the dimension (e.g. one visual, one auditory task)" (**Wickens2002**). In other words, performing a visual and an auditory task simultaneously will be "easier" (i.e. more efficient), because the underlying structures are not shared, than performing two visual tasks, as they share the same structures. In his model, Wickens identifies four dimensions, each made up of two "levels":

- processing stages (perception & cognition[7] / responding)
- processing codes (spatial / verbal)
- processing modalities (visual / auditory)
- visual processing (ambient / focal)

Not shown in the graphic representation of the model (Figure 1), but postulated by Wickens, is an additional pool of general capacity, which is always available to all tasks. In his adaptation of Wickens's model to simultaneous interpreting, Seeber takes a step further, simplifying the graphical representation of Wickens's model by turning it into a 2D model (see Figure 2). This has two main advantages. First, it allows seeing all "sides" of the cube (i.e. all dimensions) at once. Second, it graphically introduces the general capacity left out in Wickens's "cube".

[7] Perception and cognition are considered as one dimension, as one cannot take place without the other.

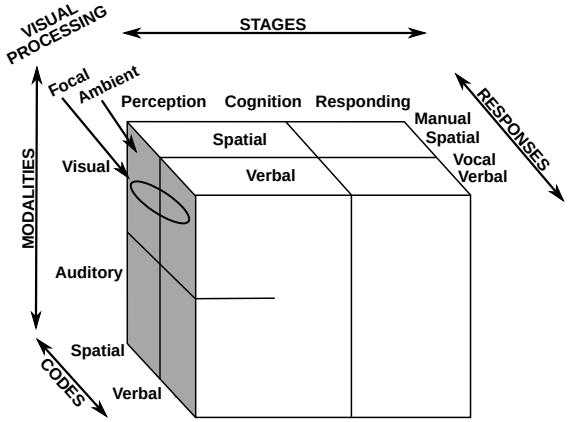

Figure 1: Cognitive Load Model (adapted from **Wickens2002**)

The result of this adaptation is a Cognitive Resource Footprint (CRF), which Seeber also develops for shadowing and sight-translation (**Seeber2007**). Simultaneous interpreting is the combination of two main tasks: the listening and comprehension task on the one hand, and the production and monitoring task on the other. As shown by 2, the first task mobilizes auditory-verbal and cognitive verbal resources at the perceptual-cognitive stage (interpreters receive the aural stimulus, i.e. the words pronounced by the speaker, and analyze the verbal message). The second task requires the same kind of resources at the perceptual-cognitive stage and additional vocal-verbal resources at the response stage (interpreters verbally "respond" to what they have heard by delivering the message in the target language, but also listen to and monitor their own rendition).

The footprint is integrated by a Conflict Matrix, which shows the degree of interference between two co-occurring tasks as the sum of the demand vectors for each sub-task and of the individual conflict coefficients between sub-tasks (see Figure 3).

The demand vectors indicate the degree to which each sub-task recruits a certain type of resource. Seeber postulates a demand vector of 1 for each sub-task. Conflict coefficients instead show to which degree the single sub-tasks compete for the same resources. When two sub-tasks share resources that are governed by the same structures, their level of conflict is higher than for two sub-tasks that do not share resources (and time-sharing between them is not as efficient). The sum of demand vectors and conflict coefficients gives a value of 9 for simultaneous interpreting.

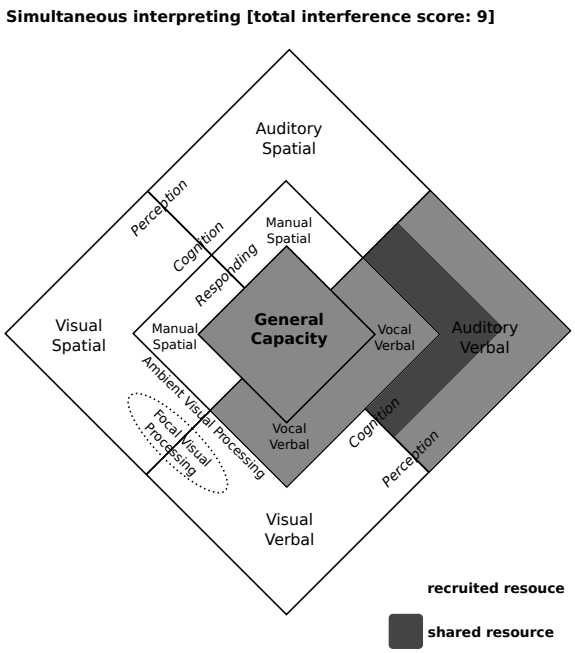

Figure 2: Cognitive resource footprint for simultaneous interpreting (adapted from **Seeber2007**)

The possibility to "quantify" the degree of interference between co-occurring tasks and to explain multi-tasking makes Seeber's Cognitive Load Model more suited than Gile's EM to formulate hypotheses on Simultaneous Interpreting with CAI tools, as discussed below. For this reason, we chose the Cognitive Load Model for simultaneous interpreting as our theoretical framework.

Figure 3: Conflict matrix for simultaneous interpreting (adapted from **Seeber2007**)

2.2 Hypotheses on SI with CAI

Seeber uses his model to represent the allocation of cognitive resources during "standard" simultaneous interpreting, without indicating any specific conditions under which this activity is performed. What happens when, during SI, the interpreter can query a terminological database? What kind of cognitive resources are recruited, and at which stage? And how much do they interfere with each other?

In addition to the operations traditionally performed during simultaneous interpreting, when working with a CAI tool, or with another terminology management solution - such as an electronic glossary – the interpreter has to type a term or part thereof in order to query the database. This action can be considered as a response to the auditory stimulus, a reaction that precedes the vocal-verbal response (i.e. the interpreter's delivery of the term in question). During the look-up process, manual-spatial resources are therefore recruited at the response stage. After the query has been completed, the interpreter is typically presented with a list of terminological pairs (the term and its equivalent(s) in the target language). He or she will therefore need to visually identify on the screen the term needed, an operation that requires visual-spatial resources at the perceptive-cognitive stage. Once the term has been identified, it is also read, making use of visual-verbal resources in the same stage of the process. As illustrated by 4 and Figure 5, the Cognitive Resource Footprint for simultaneous interpreting during which a terminological query is performed using a CAI tool or an electronic glossary recruits more resources than "standard" SI.

It goes without saying that the CRF shown in Figure 5 applies only to those moments when the interpreter is performing a query, and should not be seen as representative of the whole interpreting process. Cognitive load is not static, but varies constantly during the interpreting process, as a function of the cognitive resources recruited. We hypothesize that cognitive load is higher while the query is performed, since more cognitive resources are recruited (as shown by the CRF). In some cases, it might even lead to cognitive overload. If the term retrieval is successful, however, we expect cognitive load to go back to normal levels during production. Cognitive load might even be lower than for "standard" simultaneous interpreting, as the search for the appropriate term in the interpreter's memory would be replaced by a query in the glossary.

If we only took into consideration the Cognitive Resource Footprint, we would not, however, be able to formulate hypotheses on the differences in Cognitive Load experienced while working with a CAI tool or with less advanced terminology management solutions, such as electronic glossaries in the form of a Word or Excel table. These differences can be explored by assigning a higher demand vector to the various terminology management solutions. The Conflict Matrices can thus help visually represent the different levels of recruitment of cognitive resources. If the glossary is the same, what varies among the tools are the user interface and the search algorithm. The most advanced CAI tools, and InterpretBank[8] in particular, which we adopt in our study, are designed to yield the most

[8] http://www.interpretbank.com

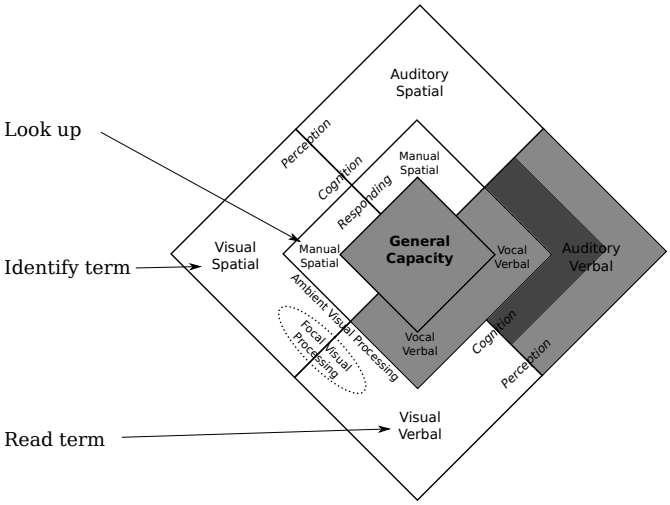

Figure 4: Additional cognitive resources recruited during SI with a CAI tool/electronic glossary

accurate results and to facilitate the user in identifying the term needed on the screen. We therefore expect the tools to require a lower level of manual-spatial resources (to look up the term) and of visual-spatial resources (to locate the term on the screen), when compared, for instance, to an Excel spreadsheet. As shown by Figures 6 and 7, we can therefore assign a demand vector of 1 to each of these resources in the case of CAI tools, and a demand vector of 2 in the case of an Excel spreadsheet. The total interference score for SI performed during the use of a CAI tool will therefore be equal to 14.8, while for SI with the use of an Excel spreadsheet it will be higher (16.8).

The integration of automatic speech recognition in a CAI tool (see **Fantinuoli2017b**) would require no manual-spatial resources, thus lowering the total interference score to at least 13.2.

3 Designing a pilot study on the use of CAI tools in the booth

3.1 Introduction

The debate around how CAI tools influence the process and the quality of interpretation is in large measure not based on empirical data, which are still very scarce and limited to a few small experiments, but is rather the result of personal beliefs

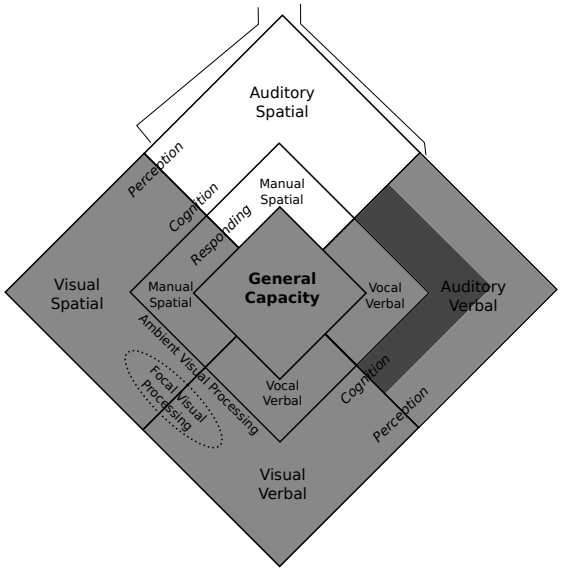

Figure 5: Cognitive Resource Footprint for SI with a CAI tool/electronic glossary

and assumptions which have not been proven empirically. A research project currently underway at the University of Mainz/Germersheim (**Prandi2016**; **Prandi2017a**; **Prandi2017b**) aims at bridging this research gap by providing data that can substantiate arguments in favor and against CAI tools. One source of difficulty in the investigation of CAI tools lies in the fact that no research methodology for the combined collection of data both on the process and on the product of SI with CAI has been developed and tested yet. In order to provide a first solution to this issue, we therefore conducted an exploratory study with the aim of evaluating the appropriateness of the stimuli used for data collection. In the following sections we will present our research questions, describe the structure of the exploratory study and illustrate the stimuli used. The analysis of the participants' renditions is the subject of the remainder of the paper.

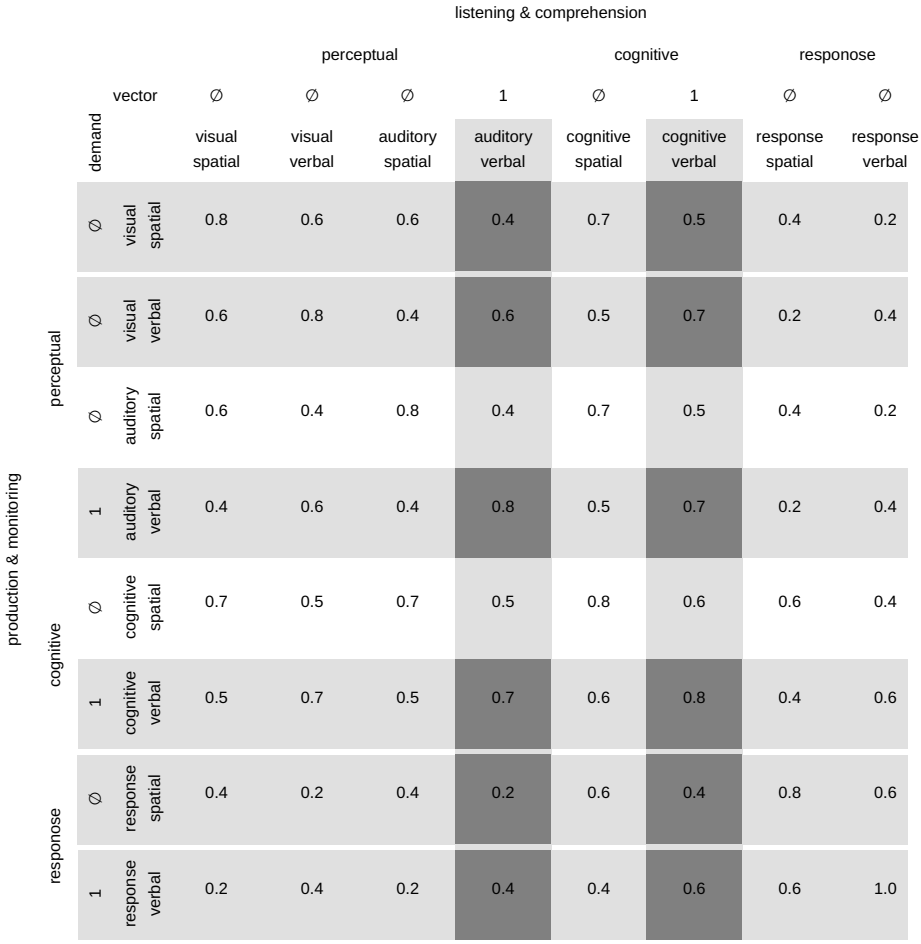

Figure 6: Conflict matrix for SI with spreadsheet (Excel) TIC = 16.8

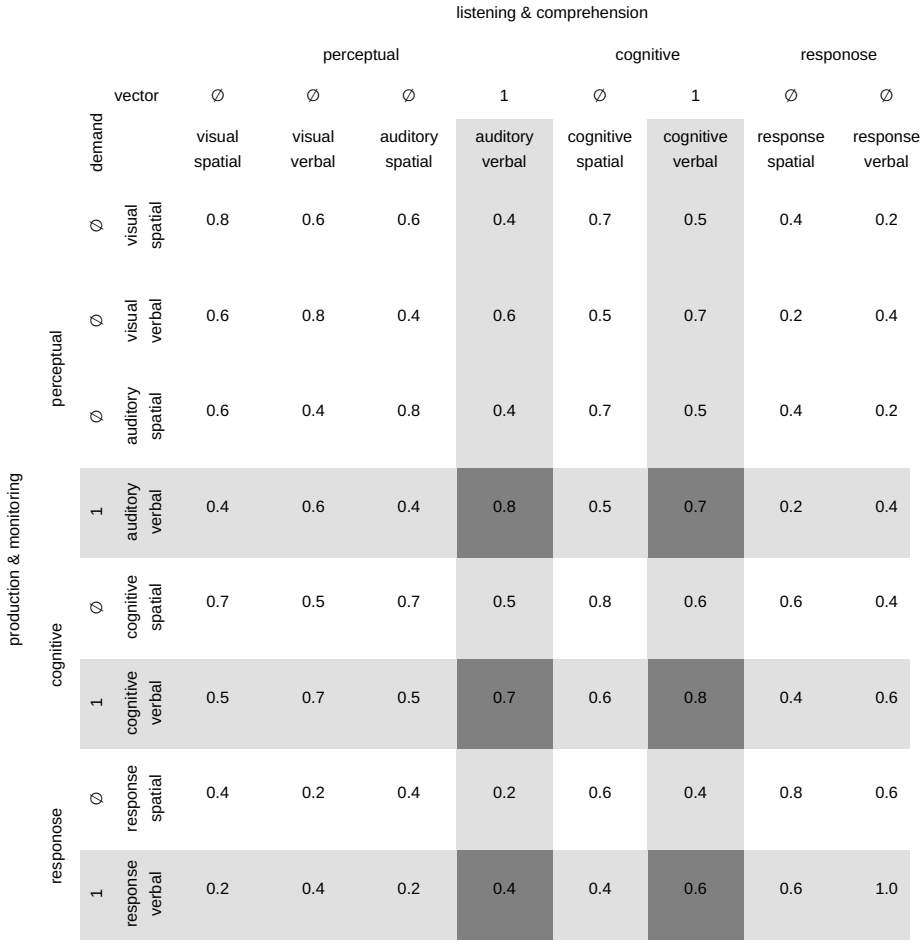

Figure 7: Conflict matrix for SI with CAI (InterpretBank) TIC = 14.8

3.2 Research questions

Our research projects aims at answering three fundamental questions:

- Do CAI tools help improve the terminological quality of the interpretation when compared to traditional electronic glossaries?
- Does a query performed with a CAI tool during SI lead to lower additional local cognitive load when compared to traditional glossary prepared with Word or Excel? Does looking up terminology lead to cognitive overload and if so, does this also happen when CAI tools are used?
- Can a combination of eye-tracking measures, pupillometry and transcription analysis be used to acquire data on the interpreting process, the terminological quality of the product and the usability of CAI tools?

In order to collect first data to help answer these questions, an exploratory study was conducted between May and July 2017 at the University of Mainz/Germersheim. For the scope of this paper, we will report on the observations made during the analysis of the product, while further work will be required to address the issues related to the process of simultaneous interpreting with CAI tools.

3.3 Structure of the study: sample, duration, training and data collection

The pilot study involved 6 advanced students of the Master's degree in conference interpreting of the University of Mainz/Germersheim. Prior to the study, all students had had at least 3 semesters of practice in simultaneous and consecutive interpreting and had English in their combination, as B or C language. Half of the sample was made up of German natives (one male and two females), half of Italian natives (one female and two males). The test subjects were recruited by e-mail and their participation in the experiment was voluntary. No monetary compensation was offered, but the participation in the study gave the trainees the opportunity to learn about a new tool, InterpretBank, and to practice in the booth with a laptop, something they rarely do in class.

The trainees attended one preliminary meeting during which the basics of terminology management for conference interpreters were covered. The presentation was centered on practice rather than theory, since a previous study confirmed this was more beneficial to achieve a good level of expertise (**Prandi2015a**;

Prandi2015b). The search functions in Word, Excel and InterpretBank were described in detail. For the purpose of the study, participants could visualize all the results of a query when working with Word[9], while they had to skip to the next occurrence when using Excel. In our presentation we made sure to choose a neutral approach to the different tools, so as not to favor the CAI tool chosen.

After covering the basics, 5 practice sessions followed in the subsequent weeks, with around 1 session per week. During each training session, the students interpreted 3 short speeches from English into their mother tongue (either German or Italian). They could use a glossary provided by us, for both language combinations, which they could access in all three formats (Word, Excel or InterpretBank). During each session, they used a different tool for each speech, so equal practice time was dedicated to each tool. The first few speeches had been prepared ad-hoc by the author for a previous study (**Prandi2015a**; **Prandi2015b**), while the last few speeches were authentic speeches selected by us, so as to ensure a certain progression in the practice material. The topics covered during the practice sessions were medicine and biology. After the last session, the students took a short test to verify their proficiency in the use of the tools. All students passed the test and were deemed ready for data collection.

Data collection took place in the *Translation and Cognition Centre* of the University of Mainz/Germersheim. The test subjects were briefed about the structure of the study and were informed that they were going to interpret 3 speeches from English into their mother tongue. They were told the topic of the speeches (renewables and other sources of energy) right before data collection started. While this does not reflect professional practice, which requires thorough preparation before interpretation proper, the students were not given the chance to prepare in advance since this would have introduced an additional variable in the study. The methods of preparation and the time dedicated to this fundamental phase of interpreting are very personal and would have been very difficult to standardize and to verify. We therefore decided to sacrifice some ecological validity to limit the number of independent variables.

Every test subject interpreted three speeches, each about 12 minutes long and with an average speed of 122.26 words per minute. This speed was chosen to make sure that looking up terminology during interpreting was challenging, but not impossible. All three speeches had been prepared ad-hoc for the study and previously recorded by a native speaker of British English. One glossary of 421 terms was prepared by us. It contained the same terms for both language combinations and had a simple tabular structure – one column for the source language

[9] The results are displayed in a column on the left-hand side of the window.

and one for the target language. The glossary was prepared with InterpretBank and then exported as an Excel spreadsheet, which was then also converted in a Word table. The glossaries were not shown to the test subjects before the interpreting task started. During the interpreting task, the screen was divided in two areas. On the left-hand side, the test subjects could see the video of the speaker, which served as a fixation cross when no term query was being performed. The glossary window was placed on the right-hand side.

The test subjects' deliveries were recorded with Audacity, while an SMI Red250m eye-tracker was used also to record pupillometry data. A LOG file, automatically created by InterpretBank, served as a reference to check what terms had been looked up by the test subjects. The same was done manually for the trials in which Word and Excel were adopted, using the Gaze Replay recordings. The interpretations were then transcribed using Partitur Editor, the transcription tool of the Exmaralda suite, and then analyzed. Before presenting the method used for this analysis and its results in section 4, we will describe the main features of the speeches used for data collection, with a focus on stimuli distribution and morphological complexity.

4 Stimuli features and distribution

While asking the test subjects to interpret single terms would have eliminated the time constraint typical of simultaneous interpreting, working with authentic, unedited speeches would have introduced too many variables in our experiment. For this reason, we decided to adapt **Seeber2011a**'s methodology (**Seeber2011a**) by creating ad-hoc speeches made up of sentence clusters. This method presents three main advantages. First, it enables us to focus our investigation on the target sentences (i.e. the ones which include the stimulus). Second, it makes it easier to work with comparable speeches, as they have the same structure. Third, it gives the test subjects the impression that they are interpreting a speech, rather than disconnected sentences, thus helping us retain a certain degree of ecological validity. Each sentence cluster is composed of a general, introductory sentence, followed by the target sentence containing the stimulus, followed by a third sentence which, like the first one, does not contain specialized terminology. The structure is repeated throughout the speech, so that each stimulus is separated from the next one by two sentences. Here is an example from speech n.1:

(1) So we need to change this basic trend and this is why the urgency is there.

> In our policies, we should definitely address the need to improve **vehicle efficiency**.
> But there is still much more we can do, in many other areas, as you are aware.
> At the EU level, there is another policy option that can help us.
> By focusing, for instance, on **woody biomass fuels**, we can truly make a difference.
> They have the potential to help us respond to the challenges we're facing.

Each speech prepared for data collection contained 36 terms, 12 of which are unigrams (e.g. "bioenergy"), 12 bigrams (e.g. "energy poverty") and 12 trigrams ("pressurized water reactor"). This variable was introduced because the length of the stimuli is expected to also play a role in the usability of the tools. We expect there to be differences between tools when a more morphologically complex term is looked up – it should be more difficult to find a trigram when using a Word or an Excel glossary than when working with a CAI tool.

Of each group of stimuli, 6 are placed at the end of the sentence and 6 in the middle of the sentence. This was done to verify whether the stimulus position has an impact on cognitive load and on the test subjects' behavior in querying the glossary. We expect the stimuli placed at the end of the sentence to lead to a lower increase in cognitive load and to be looked up more often, thanks to anticipation.

Of the 6 terms placed at the end of the sentence, 3 should require a query in the glossary, because they are less frequent and thus probably unknown to the participants, and 3 should not.[10] The same is true for the terms placed in the middle of the sentence. Half of the stimuli in each speech should therefore require a query and half should not. This variable was introduced because we want to verify whether CAI tools, which are usually deemed to be user-friendlier and to take up fewer cognitive resources, allow participants to perform more queries without leading to a higher number of errors or omissions. Figure 8 sums up the features of the stimuli and their distribution in each speech. For each speech, the stimuli can thus be classified according to their features, for future analysis. Table 1 shows an example of this classification for the stimuli in speech nr. 1.

[10] This classification was based on the frequency of the terms as per the 2015 news corpus, the 2012 web corpus (UK) and the 2016 Wikipedia corpus for the English language (Projekt Deutscher Wortschatz, http://wortschatz.uni-leipzig.de).

Table 1: stimuli classification - speech 1

Stimulus	Position	Morphological complexity	Glossary search needed (GS)
bioenergy	E	1	
security of supply	M	3	
gasoline	M	1	
conventional fossil fuels	M	3	
vehicle efficiency	E	2	X
woody biomass fuels	M	3	X
liquid biofuels	E	2	
rapeseed methyl ester	E	3	X
transesterification	E	1	X
short-rotation coppice	E	3	X
black liquor	E	2	X
corn stover	E	2	X
lignocellulosic solid biomass	E	3	X
gasification	M	1	X
gasifier	E	1	
green charcoal	M	2	X
briquettes	M	1	X
biofuels sector	M	2	
soil protection	E	2	
petroleum	M	1	
greenhouse gas emissions	E	3	
EU biofuels directive	M	3	X
indicative targets	M	2	X
incentives	E	1	
set-aside land	M	3	X
arable land	M	2	
solar power	M	2	
second-generation biofuels	E	3	
switchgrass	E	1	X
first-generation biofuels	E	3	
residue cake	M	2	X
milling	E	1	X
malting	M	1	X
overall energy demand	M	3	
renewables	M	1	
energy mix	E	2	

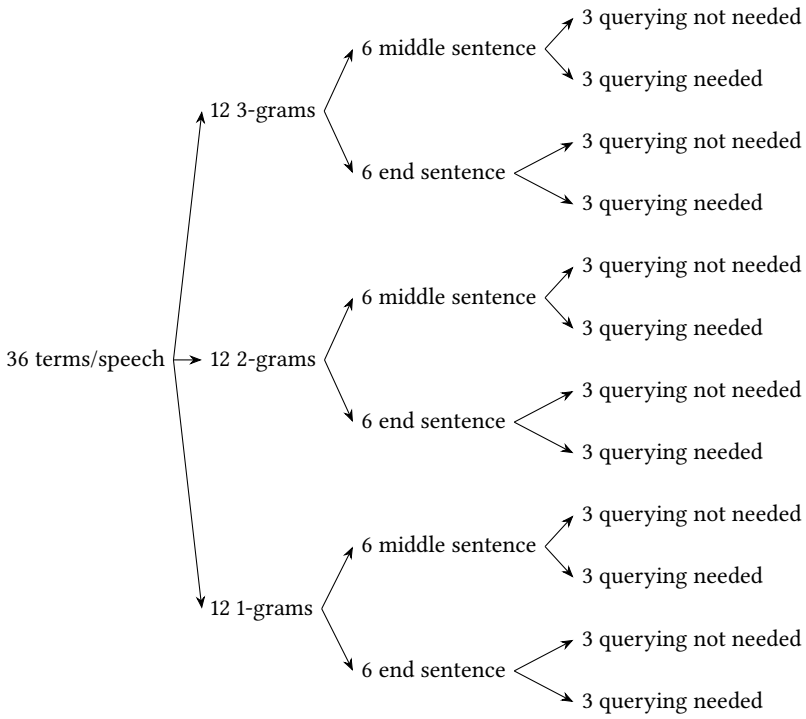

Figure 8: Features and distribution of the stimuli used for data collection

5 Stimuli validation

One of the aims of the exploratory study was to verify whether the stimuli prepared for data collection, to be used also in a future experiment involving a larger sample, elicited the reaction we expected from the test subjects, i.e. a query in the glossary. This is necessary to make sure that enough queries are performed in the glossary to provide sufficient data for a comparison between the three terminology management solutions we focus our analysis on – Word glossaries, Excel glossaries and second-generation CAI tools. While a certain degree of inter-subject variability can be expected, we must verify whether our a-priori classification of the stimuli holds true on a general level. This is the focus of the first part of the transcription analysis, which will be presented in section 4.1.

Another goal of our research is to verify whether the use of a CAI tool leads to better terminological quality in comparison to more traditional terminology management solutions, e.g. Word and Excel glossaries. First observations made

in our sample are briefly discussed in 4.2, where we also provide a framework to analyze errors and omissions in relation to the tools used for glossary query.

§?? presents the results of our observations in relation to the strategies adopted by the test subjects to interpret the stimuli. Given the small size of the sample, with this exploratory study we aim at developing a methodology to be used for further research, rather than at drawing conclusions, which will require a larger set of data.

5.1 Stimuli classification

As previously stated, half of the stimuli were classified as needing a glossary query. In order to verify whether this was true, the sample was checked for the total number of terms searched, the number of terms searched that were classified as needing to be searched in the glossary ("QN") and the number of terms searched that we did not expect to require a query in the glossary ("NO QN").

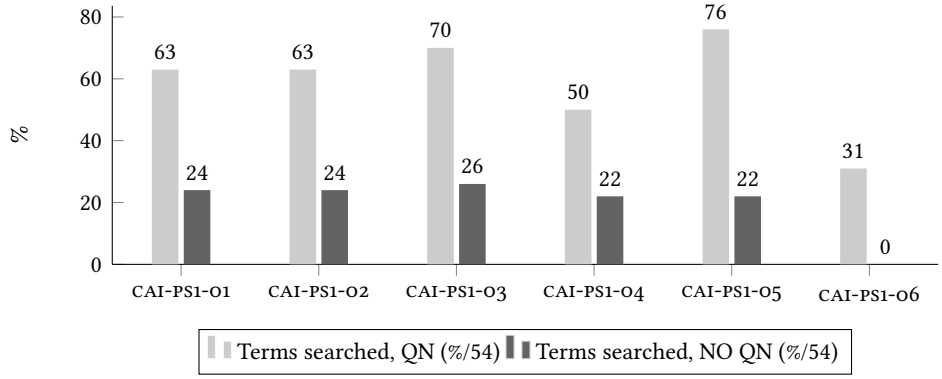

Figure 9: Search behavior per stimuli category

As shown in Figure 9, the percentage of terms classified as needing a query that were actually searched varies among the test subjects, while it is quite similar in the case of terms classified as not needing a query. A notable exception is test subject CAI-PS1-06, who searched a much lower number of terms than the other test subjects. The percentages are very similar for the German natives (participants CAI-PS1-01, CAI-PS1-02 and CAI-PS1-03), although they looked up different terms.

We also verified which terms classified as needing a glossary query had not been looked up by any subject. 5 terms out of 54 were not looked up by anyone and should therefore either be moved to the non-query category or replaced by

more specialized, less frequent terms. Of the terms classified as not needing a query in the glossary, only 1 out of 54 was looked up by all test subjects. It should therefore either be classified differently or replaced.

If we take into consideration the position of the stimuli in the target sentences, something interesting emerges from initial data analysis, which deserves further exploration in a bigger sample, especially in correlation with pupillometry measures. While the difference is, for some test subjects, more evident than for others, the stimuli placed at the end of the sentence seem to elicit more queries than the stimuli placed in the middle of the sentence (see Figure 10). This might be explained with the fact that, when a term is placed at the end of the sentence, anticipation might lead the participants to prepare themselves to adopt a coping mechanism, such as a glossary query. The "preparation" could also result in a sentence structure that favors a glossary query, requiring less restructuring or making it difficult to omit the term completely. This could, however, result in higher cognitive load, because if the query is not successful, more cognitive resources will be needed to adopt a different strategy, possibly affecting the rendition of the following sentences. A stimulus placed in the middle of the sentence could prompt the interpreter to immediately choose a strategy different than consulting the glossary available, such as generalization or the use of a synonym. While this may lead to a less precise rendition of the original, it may also come with lower cognitive load experienced.

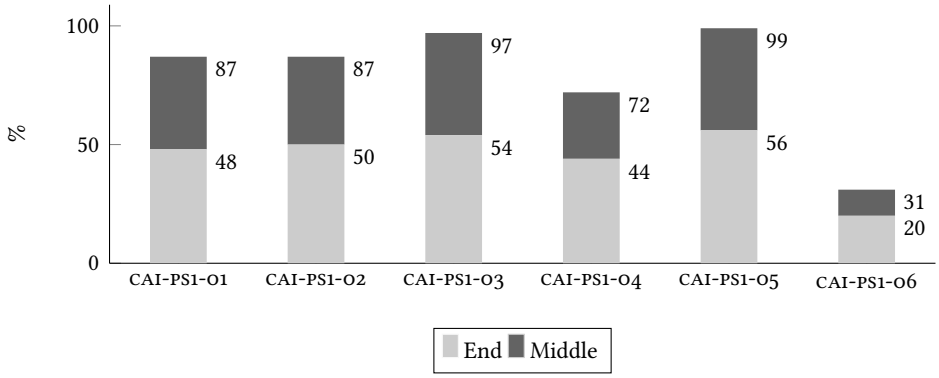

Figure 10: Stimulus position and percentage of terms searched. Percentage expressed on a total of 54 terms per category.

Even though the stimuli classified as requiring a query were equally distributed in terms of position - half of them placed in the middle and half at the end of the sentence - the difference in the search behavior might also be due to the terms

themselves, rather than only to their position. This can be further tested on a larger sample, by switching the position of the stimuli or by using a different set of stimuli.

If we take into consideration the morphological complexity - here defined as the number of elements making up the terms[11] - we notice that unigrams are searched more often than bigrams and trigrams in the case of the stimuli classified as not needing a query (see Figure 11). This might be explained with the fact that, when faced with a bigram or a trigram, participants need to decide which element of the term should be looked up, which requires additional cognitive resources. For this reason, they might choose to directly adopt a different strategy. A unigram does not require them to make this decision, and so the act of querying the glossary is more straightforward. No clear trend can be identified for the stimuli that should require a query.

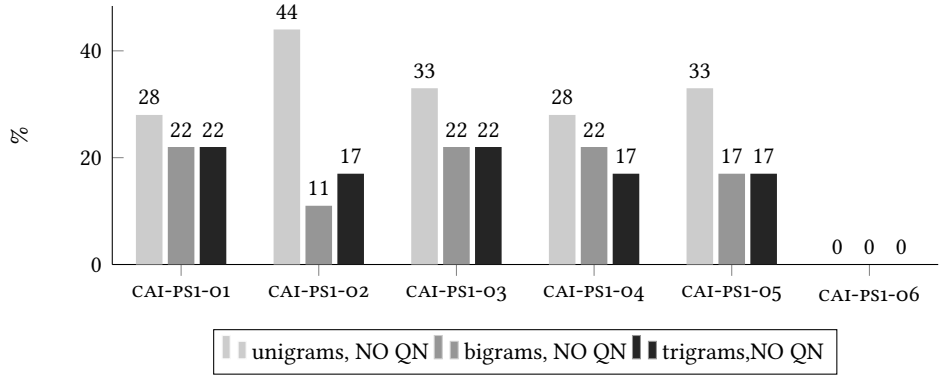

Figure 11: Morphological complexity and percentage of terms searched (terms not needing a query). N=18/category

All in all, our a-priori classification of the stimuli was confirmed by the sample, if we exclude the outlier CAI-PS1-06. Further research will be needed to check our hypotheses on the role played by the position and the morphological complexity of the stimuli.

5.2 Tools used and precision of renditions

With the aim of gaining initial data on how the tool used influences the precision of the test subjects' renditions, we compared the level of precision observed for the Word glossary, the Excel glossary and InterpretBank, when a glossary query

[11]For instance, a trigram is considered morphologically more complex than a bigram.

was chosen as the strategy to interpret the terms. Our classification of the renditions is loosely based on **Wadensjö1998** and is made up of 3 main categories:

- close renditions (precision 2 – P2): no information lost, precise rendition, use of equivalent as per glossary or adequate synonym
- acceptable renditions (precision 1 – P1): some information is lost (e.g. through paraphrasing, the loss of an adjective in trigrams, a drop in register), but the general meaning is maintained
- zero/unacceptable rendition (precision 0 – P0): the rendition completely or largely deviates from the original message (the content is different), or the information is not present (zero rendition).

This classification certainly presents some degree of subjectivity, but it is nonetheless useful as a broad guideline to evaluate the precision of the test subjects' deliveries.

Figures 12, 13 and 14 sum up the degree of terminological precision achieved when performing a glossary query with Word, Excel and InterpretBank glossaries.

Inter-subject variability is too high to draw initial conclusions on this aspect, but Excel seems to lead to the worst performance, since we can notice more occurrences of zero renditions or unacceptable renditions than for Word and InterpretBank. This is probably due to the fact that, when working with Excel, our test subjects did not have the possibility to view all the results of a query, but only to manually skip to the next occurrence, which might make the query too cumbersome to be performed in the very small amount of time available to the interpreter. InterpretBank seems to perform slightly better than Word, but this should be further verified. Usability probably plays a role in this respect, so eye-tracking measures will be key in determining how the user interface facilitates or hinders the identification of the equivalent needed.

As for the morphological complexity, we expected queries performed with InterpretBank to be more effective – leading to a higher level of precision – than queries performed with Word and Excel, especially for more complex terms (trigrams). In the small sample analyzed in the pilot study, queries performed with InterpretBank lead to higher precision for unigrams in 5 cases out of 6. The only exception is participant CAI-PS1-06, for whom we have very few data points when compared to the rest of the sample. For bigrams and trigrams the results are less uniform - queries with InterpretBank are more effective than Word and Excel in

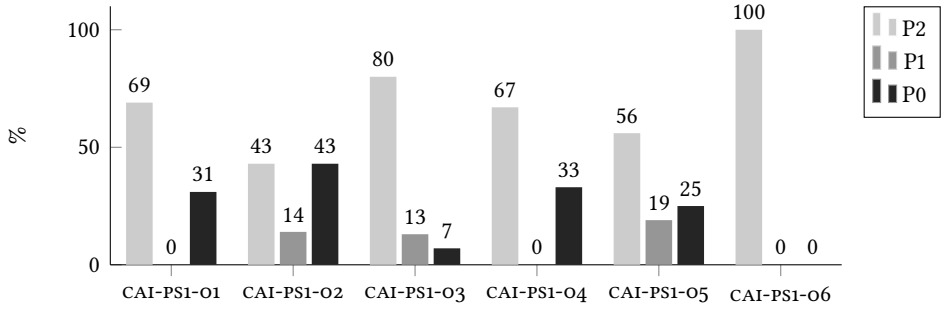

Figure 12: Precision of renditions with Word

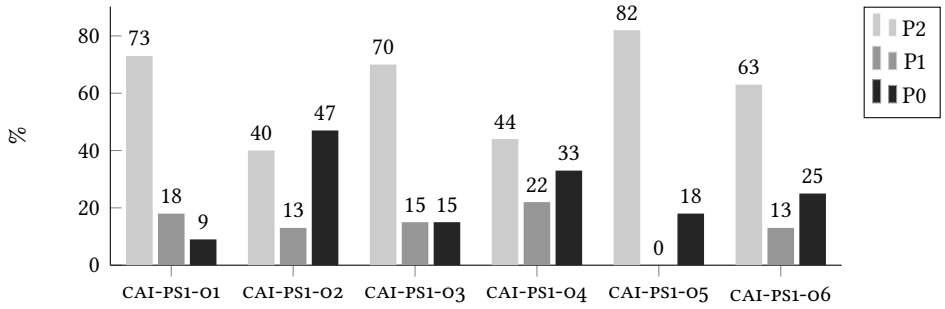

Figure 13: Precision of renditions with Excel

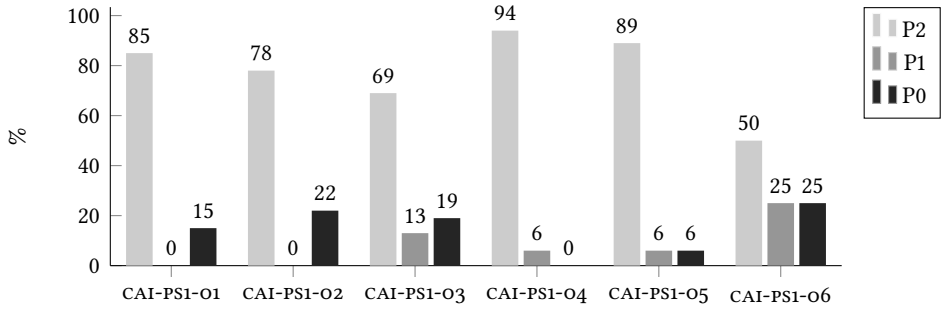

Figure 14: Precision of renditions with InterpretBank

3 participants out of 6. While the sample analyzed is too small to draw conclusions, this aspect can be further analyzed in a larger sample, where differences might be significant.

It should be noted that in order to facilitate the analysis, we first took into consideration only the terms and the content conveyed by them, not by the whole sentences. This analysis hence remains focused at a microscopic level, that of terminology. Since the ultimate goal of CAI tools is that of improving the overall quality of the rendition, we deem necessary to expand our analysis to the sentence level, to verify whether a higher level of precision achieved in the rendition of the terminological unit results in a correct and complete rendition of the sentence it is embedded in – and of the following ones – or whether, on the contrary, the query, despite being successful, leads to errors or omissions. To this aim, the transcriptions of the test subjects' renditions were annotated following **Barik1971**'s (**Barik1971**) classification of omissions, additions and errors in SI. In the data analysis, we decided to focus on three categories, which represent the most serious issues encountered in the rendition, namely E4 (substantial phrasing change), E5 (gross phrasing change) and a third category which corresponds to a complete omission of the sentence, which we labeled as M5.[12] On the other hand, to make our analysis easier, we grouped in one category the renditions that did not present any issue or only presented less serious issues, such as skipping omissions and mild phrasing changes. The classification of errors and omissions provides an element of subjectivity, which might be constrained by taking into consideration only clearly wrong sentences or total omissions.

Given the small sample of the study and the subsequent high level of inter-subject variability in terms of amount of terms searched, we were not able to identify any clear trends from this data alone. The statistical significance of the data will have to be verified on a bigger sample. Nonetheless, the pilot study was useful to define a working method that can be applied to further research and possibly expanded to also take into consideration the features of the stimuli.

5.3 Tools used and interpreting strategies

We conclude our analysis by looking at the strategies adopted, to establish whether a correlation can be found with the tools used. The classification of the interpreting strategies is based on **Bartlomiejczyk2006**. In our analysis, we focused on the "strategies of production" (ibid.), which can be observed by analyzing the product of SI, while we did not take into consideration overall strategies, which would re-

[12] See **Barik1971** for a complete classification of errors, omissions and additions in SI.

quire additional methods to be identified. From the analysis of the transcriptions, 10 main strategies, or coping-tactics (**Gile1995**), emerged:

1. Glossary search (GS)
2. Approximation (A): use of a synonym or a closely related term
3. Compression (C): use of a hyperonym, some precision is missing
4. Omission (o): not strictly considered a strategy, it is mostly unintentional
5. Paraphrase (P)
6. Reproduction (R): no translation, the term is reported as in the source language
7. Transfer (T): ad-hoc translation
8. Syntactic transformation (ST)
9. World knowledge (WK): reference to one's pre-existing knowledge
10. Substitution (S): the term is replaced by another term, not related to it.

Figure 15 reports an overview of the strategies used by the test subjects for all tools and all stimuli. Data clearly show that, with one exception, a glossary query was the strategy most used by the test subjects. This can be easily explained by the fact that the test subjects had not prepared for the assignment. The second most used strategies are approximation, omission and world knowledge.

The third most used strategies are world knowledge, paraphrase and omission.

If we look at the strategies adopted when using different tools to look up terms in the glossary, we can notice that, when using InterpretBank (see Figure 16), a glossary query is the favorite strategy, except for one subject (the same as in the general analysis), who seems to resort mainly to approximation. The second most used strategies are omission and approximation, while the third most used strategies are world knowledge, approximation and omission.

Querying the glossary was the favorite strategy also when Excel (Figure 18) was used, in 4 cases out of 6, while the other two resorted, respectively, mainly to omission and paraphrase, and to world knowledge. There is not a clear preference as to the second most used strategy, while the third most used, in 4 cases out of 6, is world knowledge, followed by approximation and both omission and compression.

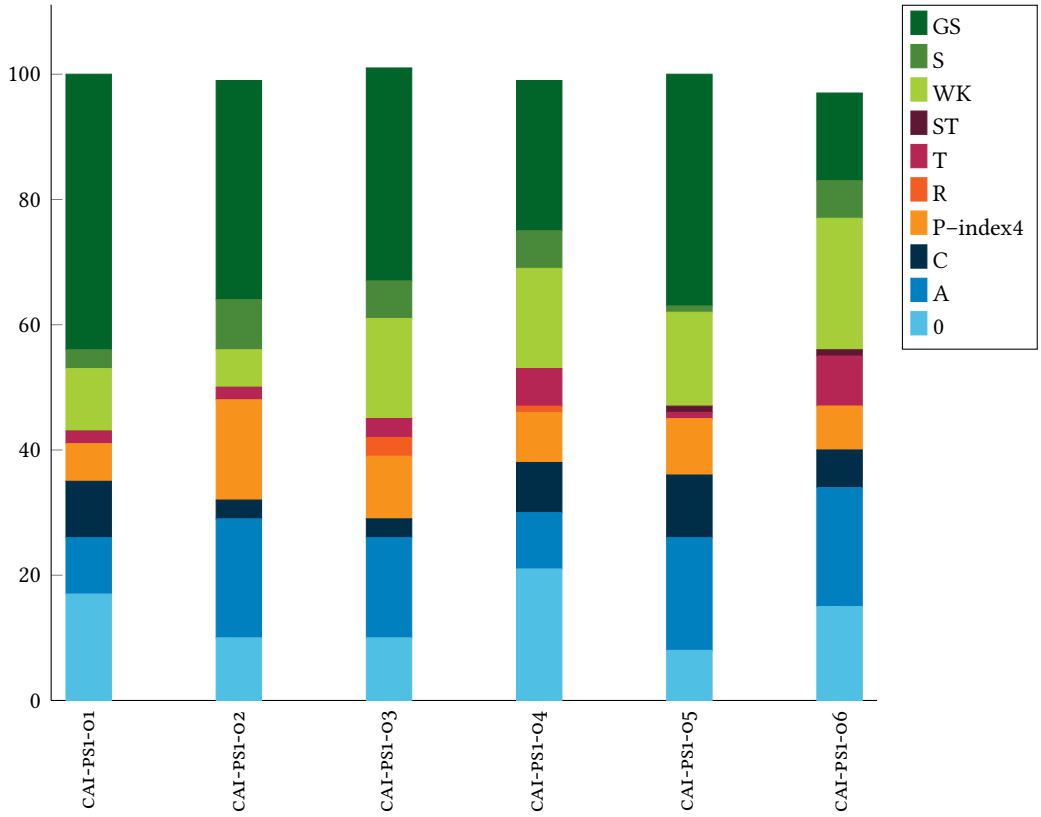

Figure 15: Overview of strategies used (all tools and stimuli)

In the third case, in which the test subjects could look up terms in a Word glossary (see Figure 17), a glossary query also seems to be the favorite strategy, while paraphrasing is the second most used strategy in the sample. The third most used strategy is omission.

Even though a glossary search was the preferred strategy by almost all participants irrespective of the tool used, the percentage of queries performed with InterpretBank seems to be higher across the board, except for one participant. By looking at these initial data, we can thus hypothesize that test subjects find it easier to perform a glossary query when using InterpretBank, probably due to better usability, and that reference to previous knowledge, approximation, paraphrasing or outright omission are the preferred coping tactics when the glossary is not queried. This should be tested on a larger sample by triangulating data from transcriptions with eye-tracking data.

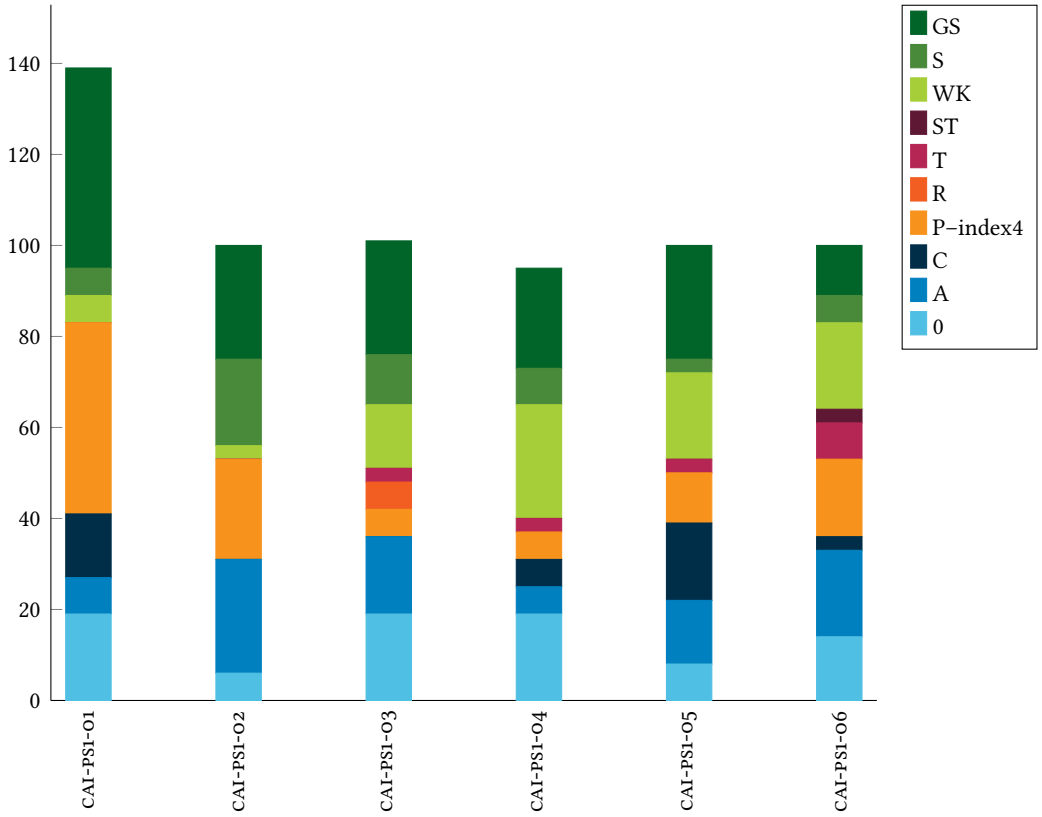

Figure 16: Strategies adopted – InterpretBank glossary

6 Conclusions and further research

The paper presented first results from an exploratory study aimed at developing a research methodology to investigate the use of computer-assisted interpreting tools during simultaneous interpreting. The pilot study is part of a PhD research project that aims at collecting data both on the procedural and the terminological aspect of SI with CAI, combining product- and process-based measures.

After discussing the theoretical framework chosen for the study, we presented our main hypotheses on cognitive load during SI with CAI. In particular, we expect cognitive load to be higher during SI with CAI than during traditional SI, but to be lower for CAI tools such as InterpretBank than for traditional terminology management solutions like Word and Excel glossaries. We also expect the terminological quality to be better when a CAI tool is used. While the hypotheses on

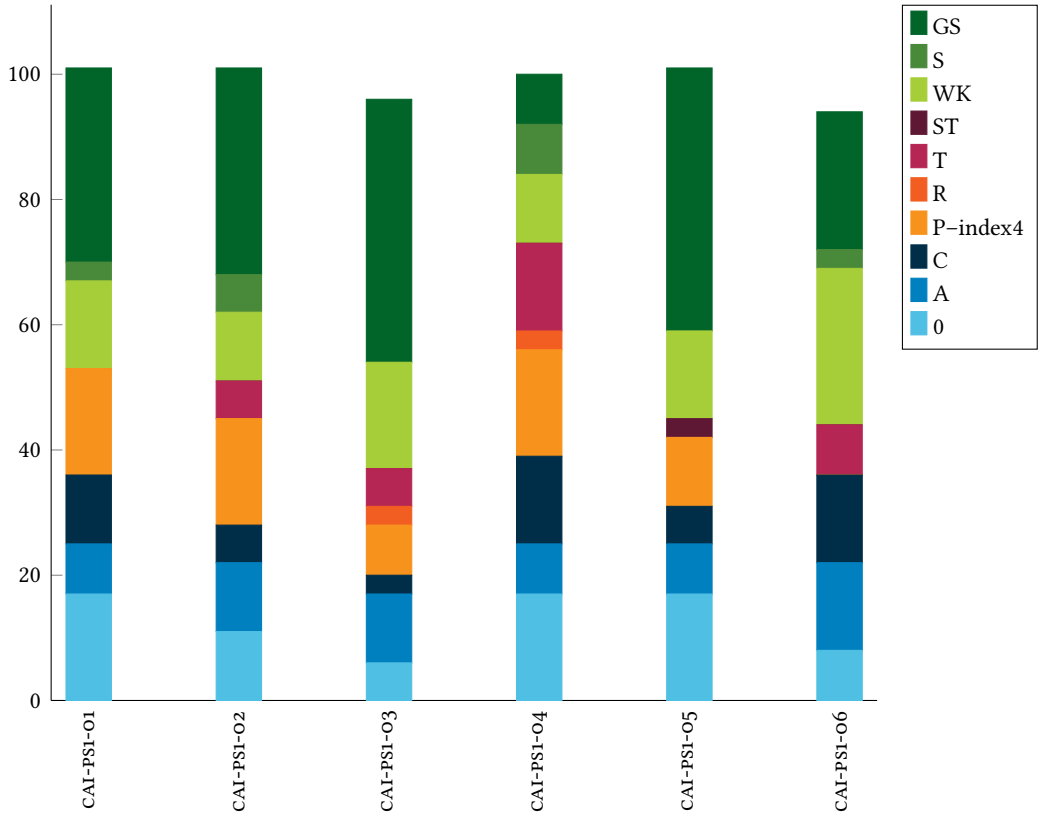

Figure 17: Strategies adopted – Word glossary

cognitive load will require the analysis of pupillometry and eye-tracking measures to be verified, the analysis of the interpretations helped shed some light on the terminological quality of SI performed with the support of CAI tools and of traditional table glossaries.

First data from the transcriptions of the test subjects' deliveries have proved helpful to verify the adequacy of the stimuli created for the experiment, showing that the a-priori classification of the stimuli used is overall confirmed by the test subjects' search behavior, in particular when it comes to the stimuli classification into terms expected to require a glossary query and terms not requiring a query. The position of the stimuli seems to play a role in the search behavior, while their morphological complexity does not seem to have a significant impact on it. InterpretBank seems to provide the highest degree of precision, and the glossary query appears to be the favorite kind of strategy to apply to cope with specialized

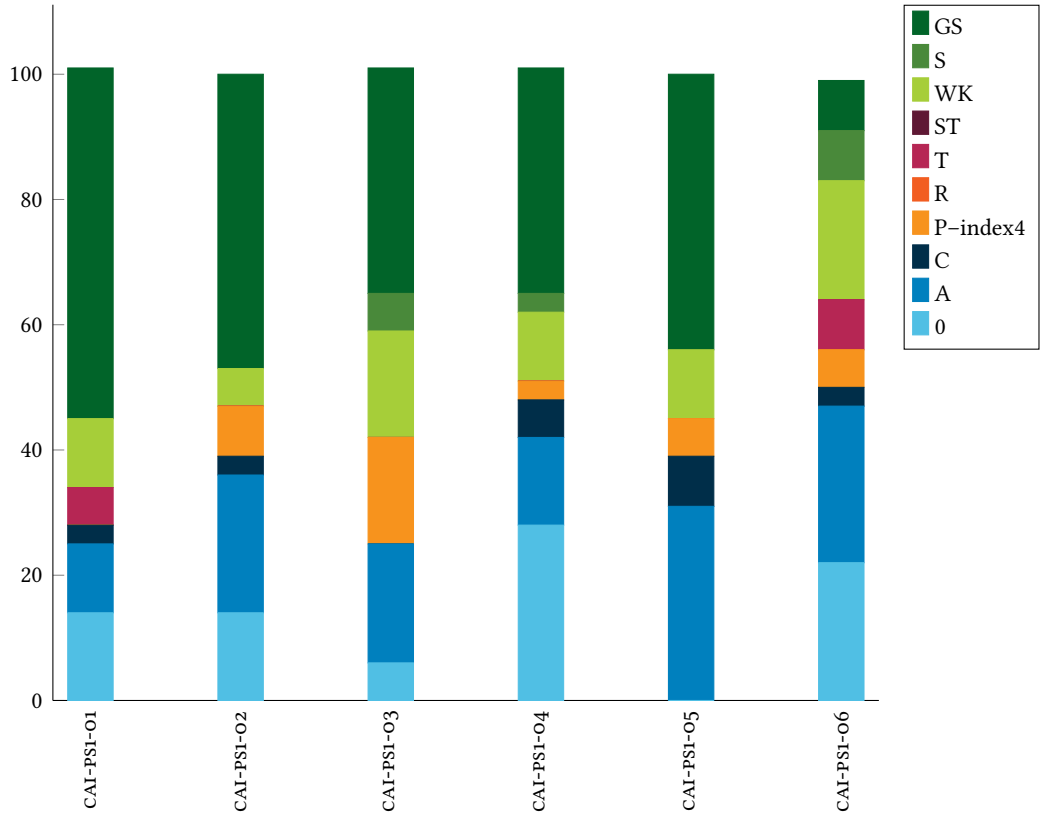

Figure 18: Strategies adopted – Excel glossary

terms when InterpretBank can be used to search for terminology. All of these aspects will need to be further investigated in future studies.

Further analysis of process-related and usability data, in particular of pupillometry in triangulation with eye-tracking measures, will be necessary to gain information that can shed some light on the hypotheses on cognitive load and help formulate further hypotheses.

Finally, future studies should also include the option to query the glossary through automatic speech recognition, which can be expected to be the most "cost-effective" option in terms of cognitive load added and level of precision, coherence and cohesion achieved in the interpretation.

Interpreting and technology

Since the 1970s, the notion of a lexeme, an abstract lexical unit identifying what is common to a set of words belonging to the same inflectional paradigm, has become a cornerstone of theoretical thinking on morphology and a standard tool for description. The present volume collects papers that crucially use, discuss or question the lexeme in the context of contemporary morphology, with particular emphasis on its place in the description of word formation through the concept of a *Lexeme Formation Rule*. It will be of interest to any descriptive linguist, theoretical linguist, or psycholinguist with an interest in morphology and its interface with syntax and lexical semantics.