



**HAL**  
open science

## Disfluencies in Simultaneous Interpreting, a Corpus-Based Study With Special Reference to Sex

Camille Collard, Bart Defrancq

► **To cite this version:**

Camille Collard, Bart Defrancq. Disfluencies in Simultaneous Interpreting, a Corpus-Based Study With Special Reference to Sex. *New Empirical Perspectives on Translation and Interpreting*, 1, Routledge, pp.36, 2019, 9780429030376. 10.4324/9780429030376 . hal-03967881

**HAL Id: hal-03967881**

**<https://hal.science/hal-03967881>**

Submitted on 1 Feb 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## 12 Disfluencies in Simultaneous Interpreting, a Corpus-Based Study With Special Reference to Sex

*Collard Camille and Bart Defrancq*

### 1. Introduction

The aim of this paper is to analyse the influence of the interpreter's sex,<sup>1</sup> as well as other predictors, in the production of disfluencies (filled and silent pauses, restarted sentences, words with elongated pronunciations, etc.) in simultaneous interpreting (SI). Few researchers have analysed the influence of sex in Interpreting Studies. Yet, there are reasons to believe that sex is a significant variable for interpreting. While the existence of sex differences in cognitive skills in general is uncertain, sex differences have been found for specific tasks related to the production, memory and analysis efforts in interpreting (Hyde & Linn 1988; Maitland et al. 2004; Aerts 2013; Hirnstein et al. 2014 inter alia). The few studies focusing on sex differences in interpreting have concluded that differences exist for disfluencies, omission, hedges and the EVS (Cecot 2001; Mason 2008; Baes 2012; Magnifico & Defrancq 2016 inter alia). Studies on cognitive sex differences are useful for several reasons. First, they can supplement and nuance studies on gender differences either by giving some explanations for the differences found or by suggesting that gender differences are societal, cultural or educational. Second, they can give researchers and trainers a new perspective on their work by encouraging them to consider sex as a significant variable. Halpern et al. (2007) suggest that we can use the knowledge acquired through studies on sex differences to teach female and male students ways to solve problems that correspond to their most efficient cognitive process, which has a positive impact on their performance. Third, it is a well-known fact that the interpreting profession is dominated by women (Lim 2005; Ryan 2015; Hickey 2018). The reasons behind this phenomenon are still relatively unclear but Miller & Halpern (2014) suggest that individuals' relative cognitive strengths are important to career and educational choices. Studies on cognitive sex differences could therefore help explain this phenomenon, while potentially eliminating stereotypes. Studies on sex differences in verbal tasks showing a female advantage might have an impact on men's decisions, as men might not choose interpreting based on their impression of a female superiority

in interpreting. This perception could also have an impact on their performance, given that people's performances appear weaker when they are told that the other sex performs better at the task (Spencer et al. 1999).

Disfluencies are believed to be triggered by cognitive load, which, in turn is believed to be determined both by the cognitive demands imposed by "the individual concurrent tasks" (Seeber 2011: 187) involved in interpreting and by the available resources to carry out those tasks. Whereas most of the literature focuses on the tasks and their effect on cognitive load and properties of interpreting (Gerver 1969; Barik 1975; Gile 1995; Seeber 2011; 2013 *inter alia*), this study will also focus on the available cognitive resources. Females are generally, but not systematically, found to perform better than males in carrying out tasks related to SI (see Section 1.3), which begs the question whether female interpreters are also more resourceful than male interpreters to cope with cognitive load. A quantitative study on disfluencies in female and male interpreting could shed some light on the matter, as disfluencies are believed to be the result of cognitive overload, i.e. a situation in which the available resources cannot meet the demands. The main hypothesis of this study is therefore that female interpreters produce fewer disfluencies than male interpreters. However, females and males perform in particular circumstances that inevitably also have a bearing on their performance. We will therefore also include a series of background variables related to the circumstances in our analysis.

This research project is based on a parallel acoustic aligned and time-tagged sub-corpus of the European Parliament Interpreting Corpus Ghent (EPICG). The data used for this study comprise 30 source speeches and their interpretations for six language pairs (English-French, French-English, English-Dutch, Dutch-English, Dutch-French and French-Dutch) for a total of 14 hours of interpreted speech and 108,245 interpreted words. The following types of disfluencies were selected: filled and silent pauses, false starts and lengthenings. The corpus offers a wide variety of speakers, speeches and languages. It was therefore decided to also look at the variables that potentially influence the production of disfluencies. The following variables are included in the data: interpreter's sex, source and target languages, Ear-Voice Span, source speaker/interpreter speaking time ratio, source speaker's and interpreter's delivery rate, source speaker's number of filled pauses, lengthenings, false starts, silent pauses and the total duration of silent pause.

In the first section, this article explains the concepts of cognitive load associated with disfluencies and with simultaneous interpreting. The literature on sex differences for several cognitive skills relevant to simultaneous interpreting is also covered. In the second section, the corpus is presented, as well as the methodology used to identify and measure the disfluencies and predictors. Descriptive statistics and the results of the statistical analyses for sex differences and of the multiple regressions

Replace with  
180

are presented in the third section. Finally, the results are discussed in the last section.

## 2. Related Research

### 2.1. *Disfluencies and Cognitive Load*

Disfluencies are defined as “phenomena that interrupt the flow of speech and do not add propositional content to an utterance” (Fox Tree 1995: 709). There is considerable variety in the literature on which items are to be considered disfluencies as well as on their classification (Shriberg 1994). In most studies, a distinction is made between repair and non-repair categories. A repair is an occurrence of interrupted speech combined with an attempt at producing an alternative for a previously articulated segment (Levelt 1983). Shriberg’s (1994) own classification in single-token disfluencies and structured disfluencies, for instance, partly rests on that distinction. Except for the lexical false starts, single-token disfluencies do not involve a repair: filled pauses (*uh*, *uhm*), intra-word pauses and hesitation-related lengthening of phones. On the other hand, structured disfluencies include typical cases of repair: repetitions (of one or more words), deletions (cases where a segment is discarded and the structure is started anew), substitutions (cases where segments are replaced with others in the same structure) and insertions (when a segment is repeated with an additional item). As explained in the methodology, we will select a subset of the most easily identifiable disfluencies for this study, i.e. false starts, filled pauses, silent pauses (intra-word but also between words) and lengthenings. The latter three pertain to the non-repair type, while the first is a repair. Silent pauses are a complex type of disfluencies as they are not always a disfluency; they can also be strategically used as a rhetorical device or to maintain the prosodic structure of an utterance (Macgregor et al. 2010).

An alternative classification based on the causes of disfluencies is put forward by Gósy (2007), who differentiates between two major groups of speech disfluencies: (1) disfluencies rooted in uncertainty (UDs) such as hesitations, fillers, repetition, restarts, lengthening and pauses within the word and (2) error-type disfluencies (ETDs) such as Freudian slips, grammatical errors, contamination, false word activation, tip of the tongue, ordering problems and slips.<sup>2</sup>

Disfluencies are frequent in speech: averaging across a number of studies, and excluding silent hesitations, it has been estimated that disfluency in spontaneous speech affects about 6 per 100 words (Bortfeld et al. 2001). Demographic variables seem to influence the frequency of disfluencies: Bortfeld et al. (2001) and Shriberg (1996) both report that men produce more disfluencies than women. Engelhardt et al. (2010), however, do not find significant differences between the sexes. The frequency

of disfluencies also appears to increase with age (Shewan & Henderson 1988). Demographic variables are interesting because they consider the capacity dimension of cognitive load. Cognitive load rests on the idea that the human working memory has only limited capacity, which prevents it from performing several tasks simultaneously at the same speed and the same level of efficiency as when the tasks are performed separately (Welford 1952; Broadbent 1958). Working memory also has limited capacity for storing the information that is necessary to perform the tasks. If capacity does not meet demands, disfluencies tend to appear or increase. Differences on the sex or age dimension could be due to capacity differentials. This study will explore the former dimension.

Disfluencies can be triggered by a variety of situations that impose increased demands: new information (Clark & Fox Tree 2002; Arnold et al. 2003), heavy constituents (Swerts 1998; Arnold et al. 2000; Watanabe et al. 2008) or long sentences (Oviatt 1995; Shriberg 1996). Boomer (1965) and Shriberg (1996) found more disfluencies near the beginnings of turns or sentences, where planning effort is presumably higher. The topic or domain of a conversation is another characteristic that may cause the planning load of utterances to vary. In one study by Schachter et al. (1991), it was found that social science lectures contained more fillers than hard science lectures, while humanities lectures contained the most.

While disfluencies are seen as the consequence of cognitive load, they can also have a positive influence on the listener. The small number of studies that have investigated the effect of disfluencies on comprehension show that under specific circumstances, disfluencies can help the listener by helping the identification of upcoming words (Howell & Young 1991; Fox Tree & Schrock 1999; Brennan & Schober 2001; Fox Tree 2001; Arnold et al. 2003, Arnold et al. 2004; Ferreira et al. 2004). In specific contexts, disfluencies are used as communicative cues as they signal troubles in delivery to the listener and aid in comprehending the intended grammatical and semantic structure (Betz et al. 2015). If a speaker hesitates, the listener expects the speaker to say a difficult, non-predictable word because the listener assumes a higher cognitive effort and will therefore focus more (Kutas & Hillyard 1984; Corley et al. 2007; Collard et al. 2008; Macgregor et al. 2009). Corley & Hartsuiker (2011) find that filled pauses (*uh*, *um*, and the like) serve to signal upcoming delays in a way that informs listeners' reactions and facilitates word recognition. Fox Tree (2001) found that both English and Dutch listeners are faster to identify a target word in a carrier sentence when it follows an *uh* in comparison to a control condition without the *uh*. It is also possible that the benefits for perception emerge from the fact that disfluencies like *uh* and *um*, and any silent pauses preceding or following the filler, considerably delay target word onset themselves. Anecdotal evidence suggests that speakers who are difficult to follow will be more easily understood when they speak more slowly (and therefore pause more).

## 2.2. *Cognitive Load in Simultaneous Interpreting*

Simultaneous interpreting is considered a cognitively demanding task (Gerver 1969; Barik 1975; Gile 1995; Seeber 2011; 2013 *inter alia*). It includes following cognitive processes, such as speech comprehension and production, memory, attention/resource allocation and the monitoring of simultaneous operations (Klaudy 2004). Interpreters are required to deliver their speech in a fluent and efficient way, making speech production and fluency key skills for simultaneous interpreters. The speech production system mobilises the mental lexicon, knowledge of the outside world and a syllabary (Levelt 1999), while being kept in check by a self-monitoring mechanism (Postma 2000) that inspects one's own speech and takes appropriate action when errors are made (Hartsuiker & Kolk 2001). Disfluencies can also be seen as a window into the cognitive processes of speech planning (Goldman-Eisler 1958; Nooteboom 1969; MacKay 1973; Garrett 1975; Shattuck-Hufnagel 1979; Levelt 1989; Bock & Levelt 1994). Disfluencies are connected to the problem of lexical access in the form of false word activations and the disharmony between lexical access and articulatory planning in the form of prolongations and restarts (Tóth 2011).

SI can also be seen as a special case of speech production in noise (Tóth 2011). Indeed the condition of a task performed in noisy environments and interpreting is similar: the need to divide attention. Interpreters must divide their attention between listening to the input utterance and rendering the translation and therefore work under high cognitive pressure. Interpreting is a good example of a process where demands compete for cognitive capacities and is believed to be linked to the ability to manage competing demands on limited cognitive resources (Liu et al. 2004). Gile (1997) and Seeber (2011) have designed capacity-demand models that represent the interpreting process as a “cognitive management problem” where interpreters need to find a balance between available resources and demands imposed by the different sub-tasks of interpreting. When the processing load exceeds the processing capacity, the interpreter will most likely produce errors and omissions, as well as disfluencies. Gile (1995) confirms that restarts and lengthenings, two types of disfluencies occurring in the output of simultaneous interpreters, might be partly explained by the mental energy required for the task.

Disfluencies are therefore likely to be particularly frequent in interpreting. Several studies show that speech produced under high cognitive load typically presents more disfluencies (Jameson et al. 2009; Yap 2012; Schuller & Batliner 2013). Christodoulides & Lenglet (2014) found more filled pauses, false starts, repetitions and deletions as well as longer pauses for interpreting compared to reading (in total, 9.8% of the tokens were disfluent in SI, compared to 0.4% in Reading). However, their study is based on only two experimental subjects. Other studies, carried

out on larger populations show that interpreters produce longer silent pauses compared to the source speech (Tissi 2000; Cecot 2001; Ahrens 2005; Christodoulides 2013), more false starts (Pöchhacker 1995; Tissi 2000), more numerous vowel and consonant lengthening (Tissi 2000) and more filled pauses (Plevoets & Defrancq 2016, 2018). Mead (2000) has also shown that students produced more filled pauses when interpreting into their B language than into their A language because the former is more cognitively demanding than the latter. Bakti (2009) found that restarts and grammatical errors are the most frequently occurring disfluency in a corpus of simultaneous interpreters working from English into Hungarian.

Moreover, several factors are known to increase the cognitive load and are therefore likely to have an influence on the frequency of disfluencies. Seeber (2011) lists two input features: delivery rate and language pair. While delivery rate is reported not to increase cognitive load significantly in the comprehension of spontaneous speech (Voor & Miller 1965), it has been proven to influence interpreters' performances considerably (Gerger 1969; Pio 2003). Seeber (2011) suggests that simultaneous interpreting of Subject-Object-Verb (e.g. German and Dutch) into Subject-Verb-Object (e.g. French and English) structures generates more cognitive load than interpreting SVO into SVO structures. This is mostly due to the additional cognitive processing triggered by the strategies (i.e. waiting, stalling and chunking) applied by interpreters when dealing with syntactic asymmetry between source and target languages. Gile (2008) mentions source text sentence length as one of the factors increasing cognitive load in interpreters. However, he adds that length as such is probably not a factor, but rather the syntactic complexity that often comes with longer sentences. He also points to the lexical density as one of the prime determinants in cognitive load in interpreting. Chmiel & Mazur (2013) report that in an experiment on sight translation performed by trainee interpreters long sentences receive longer fixation times, indicating an increased cognitive load in the interpreter.

Research into the factors of cognitive load in interpreting has been predominantly source-oriented, but it is reasonable to assume that the cognitive load of an interpreter is also influenced by his or her attempts to produce a target text under high cognitive load. In two studies on cognitive load and filled pauses, Plevoets & Defrancq (2016, 2018) found that both high source text delivery rates and high source and target text lexical density triggered more filled pauses in interpreters. High formulaicity has a significant negative effect on the occurrence of filled pauses. Given that lexical density measures the number of content items while formulaicity measures the number of multiple word units, this suggests that the triggers of disfluencies are chiefly lexical. Since disfluencies are assumed to be influenced by the way interpreters divide their attention between different efforts (Gile 1995), we can assume that the length of

the Ear-Voice Span (i.e. the lag time between the source speaker and the interpreter) will also have an impact on disfluencies (Tóth 2011). Gile (1995) considers Ear-Voice Span (EVS) a possible indicator of how the various efforts relate to each other. Indeed, a long EVS might mean that the interpreter prioritises the listening effort over the production effort, while a short EVS potentially means that the interpreter is saving memory capacity. The type of disfluency can also be influenced by the type of effort, as Setton (1999) suggests that long silent pauses mean that the attention is dedicated to the listening task while long filled pauses mean that the attention is almost entirely dedicated to the production task.

It is vital, in a study exploring sex as a predictor for disfluencies in simultaneous interpreting to consider a wide variety of contextual variables, such as the ones discussed in this section. As repeatedly stated in the literature (Diriker 2004; Duflou 2016), simultaneous interpreting is a situated linguistic performance that cannot be dissociated from the context in which it takes place. A study on the effect of sex could very well yield completely misleading results if contextual variables were not taken into account.

### 2.3. *Sex Differences in Cognitive Skills*

Sex differences in cognitive skills are a sensitive and controversial topic that needs to be handled with care. Meta-analyses, which aim at determining whether sex differences exist as a whole or in global categories (mathematical, verbal and spatial among others), tend to indicate that males and females are much more similar than they are different and that sex differences are often exaggerated (Hyde & Linn 1988; Hyde 2005; Miller & Halpern 2014). For example, Hyde & Linn's (1988) meta-analysis on verbal abilities includes various tasks (spelling, reading, writing and vocabulary) and concludes that there is no scientific proof that allows claiming that women and men have different verbal abilities. This being said, several individual studies found sex differences for specific tasks and Hyde & Linn (1988) recognise that females do score significantly higher at one particular task: speech production. Accordingly, the aim of this research project is neither to prove that cognitive sex differences exist in general and outside of interpreting nor to generalise the findings to the rest of the population. This research project aims at exploring the influence of a rarely analysed predictor (sex), alongside other known predictors, on disfluencies during one specific task (simultaneous interpreting).

As mentioned before, studies found sex differences in specific tasks that appear to be relevant for simultaneous interpreting. Besides having better speech production abilities (Hyde & Linn 1988), women have been found to have greater verbal fluency, i.e. the ability to retrieve specific information within restricted search and time parameters, for example the ability to generate words beginning with a single letter in one minute



(Herlitz et al. 1997; Loonstra et al. 2001; Maitland et al. 2004; Hirnstein et al. 2014). A female advantage in generating synonyms has also been found (Hines 1990). Aerts et al. (2013) found that women display a larger sensitivity to the phonemic contrasts during auditory phoneme discrimination and showed more differentiation in real word-pseudoword dissociation. Studies have also found faster processing speed (the speed at which a person can understand and react to the information he or she receives) in women (Keith et al. 2008) and a female advantage in both pre-lexical and lexical processing was found (Majeres 1999), as well as for perceptual speed (the ability to compare or recognise items) (Born et al. 1987; Hedges & Nowell 1995). In other words, women could have an advantage for the listening and analysis effort in SI. For the memory effort, evidence for a female advantage in episodic and some aspects of semantic memory has been found (Kramer et al. 1997; Herlitz et al. 1999; Maitland et al. 2004). Women also tend to perform better than men on immediate and delayed free recall and on recognition tasks with verbal and visual components (Trahan & Quintana 1990; Kimura & Seal 2003). However, Harness et al. (2008) report higher scores for males in a study on recall combined with a distraction task carried out on students, which could be relevant to our purposes if simultaneous interpreting is considered a language production task with the incoming speech as a distractor. Taylor & Francis report that females outperform males in the Rey Auditory Verbal Learning Test (free recall of two lists of nouns, which aims at evaluating short-term auditory-verbal memory, retroactive, and proactive interference, retention of information among others) and the Verbal Paired Associates test (immediate and delayed recall of word pairs, aimed at evaluating explicit episodic memory performance) (Bolla-Wilson & Blecker 1986; Gale et al. 2007).

When it comes to the production of disfluencies, females have been found to suffer less frequently from clinical disfluency (or stuttering) (Guyette & Baumgartner 1988; Yairi & Ambrose 1992). Shriberg (1996) finds that men produced more fillers than women. Similarly, Bortfeld et al. (2001) show that men present higher rates of disfluencies overall (6.80 to 5.12 per 100 words), which is mainly due to higher rates of fillers and repeats. Men produced slightly but not significantly higher rates of restarts than women. Engelhardt et al. (2010), however, do not find significant differences between the sexes.

As far as sex effects on simultaneous interpreting are concerned, only Cecot's (2001) experimental study seems to have taken them into account. She finds that females use more filled pauses, whereas males use more unfilled pauses, and that men's unfilled pauses last longer than women's. There is some evidence that female and male interpreters cope differently with pragmatic challenges in interpreting: Mason (2008) finds more omissions for males in consecutive interpreting in the courtroom and suggests that omission patterns are determined by gendered social behaviour.

Magnifico & Defrancq (2016, 2017) show that female interpreters use more hedges and downtone fewer unmitigated face-threatening acts than male interpreters at the European Parliament and, interestingly, that female interpreter self-repair more often than male interpreters (Magnifico & Defrancq submitted).

### 3. Research Question and Hypotheses

The main research question of this study is what determines differences in the production of disfluencies between individual interpretations. Overall, the cognitively challenging nature of interpreting is expected to lead to more disfluencies in interpreting than in normal speech. Sex as a predictor for disfluencies will be explored, although the meta-literature on differences in linguistic performance and fluency between the sexes is generally inconclusive. Accounts of a female advantage in some linguistic tasks, production tasks in particular, do, however, exist and outnumber accounts of a male advantage. Therefore, our main hypothesis will be that female interpreting will present a lower frequency of disfluencies than male interpreting. To do justice to the complex nature of the interpreting task and to the situated nature of the interpreting activity, a list of potential contextual predictors was also included in the study. Some of them have been identified in previous research as increasing cognitive load and are therefore expected to have an effect on the frequency of disfluencies in interpreting: speaker's delivery rate, EVS, language pair. The influence of the frequency of other types of disfluencies on each disfluency will also be studied.

### 4. Methodology

#### 4.1. A Corpus-Based Approach

The present study is corpus-based, i.e. based on naturalistic data produced in a real-life environment by professionals and have therefore the potential to reinforce the empirical foundations of interpreting research (Shlesinger 1998). Thanks to new technologies that simplify the compilation, transcription and analysis of data, as well as to the availability of interpreting data online, corpus-based Interpreting Studies (CIS) are gaining in popularity. Back in the early 2000s, the University of Bologna compiled the first publicly accessible simultaneous interpreting corpus from data from the European Parliament, the Parliament Interpreting Corpus (EPIC) (Bendazzoli & Sandrelli 2005). Several universities have followed suit, building their own interpreting corpora: CoSi (consecutive and simultaneous interpreting) and DiK (dialog interpreting in public service settings) at the University of Hamburg, EPICG (European Parliament Interpreting Corpus Ghent) at Ghent University, and others at the universities of Rome, Trieste, Poznan, Louvain-la-Neuve and Saarbrücken.

The data used for this study are drawn from EPICG, one of the corpora based on audiovisual recordings of speeches and interpretations at plenary sessions of the European Parliament. Speeches at the European Parliament are generally very short (one to six minutes), and the working conditions only reflect the institutionalised context, which makes the corpus admittedly less representative for the interpreting activity at large. Therefore, the conclusions drawn by data coming from this context do not necessarily apply to all types of interpreting. On the other hand, interpreters working for the plenary sessions of the European Parliament are generally experienced and have undergone an accreditation test, which ensures a baseline interpreting quality in the data. Corpus-based studies are sometimes criticised because they consist of samples taken out of the context in which they occurred (Diriker 2004; Duflou 2016). It is therefore essential to provide metadata to give more information about the context (Burnard 2002). Metadata on the speaker (political group and function, age, sex) and the speech (topic, time of the day, delivery type) were added to the corpus based on information provided through the European Parliament's website.

The data used for this study are collected from a sub-corpus of EPIC Ghent, which comprises 30 source speeches and their interpretations for six language pairs (English-French, French-English, English-Dutch, Dutch-English, Dutch-French and French-Dutch). These data offer a wide diversity of topics, speakers and interpreters. Source and target texts were transcribed according to the Valibel instructions (Bachy et al. 2007) and acoustically aligned on the basis of pauses with the transcription tool EXMARALDA Partitur-Editor. More information on the compilation process, including transcription conventions and annotations can be found in Bernardini et al. (2018). One important aspect of EPICG is the availability of oscillograms for source speech and interpretation. These obviously facilitate the identification of disfluencies such as pauses. In total, the corpus comprises more than 14 hours of interpreted speech and a total word count of 108,245 interpreted words.

For each combination of source and target language, a balanced set of 15 male and 15 female interpretations was aimed at.<sup>3</sup> Given that the interpreters in the corpus are anonymous, the authors were faced with two methodological challenges: (1) sex had to be determined on the basis of the recorded voices only and (2) there is a possibility that some interpreters occur more than once in the database. In order to solve the first challenge, the sex of each interpreter was determined separately by both authors and an independent reviewer, which is a reliable method according to Lass & Puffenberg (1971). Indeed human listeners are able to identify speaker sex with an accuracy of over 95%. The process for our corpus yielded an inter-rater agreement of 99.4%, with the three assessors diverging on only one interpretation. It was concluded that the disagreement came from a human encoding mistake and the three assessors finally agreed on all interpretations. In order to complement the human

identification process, a speaker-diarization software LIUM\_SpkDiarization (Rouvier et al. 2013) was also used. After several necessary adaptations to the software and the corpus's audio data, the human-machine agreement reached 95.6%. However two human assessors disagreed with the sex identified differently by the software and agreed with the human identification. The latter was therefore chosen as the reference. The software's lack of accuracy can be explained by the fact that it is optimised for radio and TV shows and it cannot guarantee the same level of performance for other types of recordings. Moreover, the diversity of languages and speakers and the complexity of the audio data (several speakers take the floor simultaneously and the quality is not always optimal) make the task more challenging for the software.

Additional steps were undertaken in order to tackle the second challenge and ensure a representative and diverse set of interpreters. First, the study analyses languages that are sufficiently common to guarantee that they are covered by a large number of interpreters (English, French and Dutch). Second, speeches were randomly sampled from the European Parliament's website over a six-year period in order to reduce the risk of having the same teams in the interpreting booths. In the final dataset, 93 interpretations are drawn from 21 different dates in 2008, 47 interpretations from 17 different dates in 2009, 14 from 11 different dates in 2010, 9 from 9 different dates in 2011, and 16 from 4 different dates in 2013 and 1 in 2014. As a rule of thumb, a maximum of three interpretations from the same interpreter for each target language and sex is considered as acceptable by the authors.

While human listeners seem to be more successful than the software at identifying sex, the identification of speakers for such a diversified corpus appears to be quite unreliable when performed by human listeners. Therefore, LIUM\_SpkDiarization was used for the identification of identical interpreters, and the results were rated as reliable by two human assessors. However, given the difficulty for these raters to identify the interpreters, no precise human-machine agreement could be determined. Table 12.1 therefore the results of the LIUM\_SpkDiarization's analysis only.

Table 12.1 Identification of identical interpreters by LIUM\_SpkDiarization

Sex	Language	Number of interpreters identified		Total number of unique interpreters
		Twice	Three times	
Females	French	3	0	27
	English	3	1	25
	Dutch	4	0	26
Males	French	3	0	27
	English	2	2	24
	Dutch	2	0	28

While in theory the presence of identical interpreters in the data set is a violation of the assumption of independence of observations within each group (males and females), the effects are likely to be very limited given that for each sub-corpus the ratio interpreters/interpretations is at least 0.80 and no interpreter occurs more than three times in the database.

#### 4.2. Identification and Measurement of Disfluencies

For our research, the following types of disfluencies were selected and manually identified: lengthenings, filled pauses, false starts and silent pauses. These disfluencies are relatively easy to identify and were frequently observed in the corpus. The task of identifying silent pauses and lengthenings was simplified by the presence of an oscillogram where these two types of disfluencies produce perceptible patterns. Where one disfluency occurred right after another or where there were several disfluencies in a row, these were coded as separate occurrences. No distinction was made between the different types of filled pauses (*uhm*, *uh*, *hum* and *euh*). False starts include truncated words (*the pre/president*) no matter whether the same word is repeated or if another word is uttered (*the pre/chairman*). Silent pauses were included if they lasted at least 0.2 seconds in conformity with other studies on the same topics (Goldman-Eisler 1972; Duez 1982; Tissi 2000). Two types of measurements were taken into account: their frequency (i.e. the total number of silent pauses in one speech) and their total duration (i.e. the total duration silent pauses account for in a whole speech, contrary to the duration of one single silent pause). It is important to mention that all pauses were taken into account, not only intra-word pauses. Therefore, when analysing the results, we will make a distinction between the articulated disfluencies (i.e. lengthenings, filled pauses and false starts) and silent pauses, as the latter do not necessarily constitute disfluencies.

While disfluencies were first manually annotated in the transcript, they were counted by a tailor-made script. Their frequency was determined in two ways: on the one hand, we performed a normalization per minute on the text level—the number of occurrences of a particular disfluency is divided by the total duration of the interpretation they occur in, and is then multiplied by 60 in order to obtain a number of disfluencies per minute. This measurement provides an average frequency of disfluencies in a whole speech. Secondly, since disfluencies (and other predictors) can vary highly throughout a speech, source speeches and interpretations were also divided into segments. As a rule of thumb, we determined segments of ca. 10 seconds, but as we wanted to avoid segment boundaries splitting up articulated portions of the acoustic signal, segment length varies. To avoid distortions, only segments of 8 to 12 seconds were included in the study (97.9% of the total number of segments). For each segment, the frequency of disfluencies is determined manually. For

a clearer representation, these data were also normalised (each data was divided by the actual segment duration and multiplied by 10 seconds).

The distributions of all disfluencies being rightly skewed, the one-tailed non-parametric Mann-Whitney U-test was performed with IBM SPSS to check for potential sex differences in each type of disfluency per interpretation. As a reminder, if the  $p$ -value associated with the Mann-Whitney U test is below the significance level (0.05), the difference between men and women is considered as significant. The assumptions for the test are fulfilled: the dependent variables (disfluencies) are continuous, the observations are independent (an interpreter cannot be represented both in the male and the female group) and the distributions of both groups (males and females) for each type of disfluency have a similar shape. The effect size (or  $r$ ) for the Mann-Whitney's U test is calculated by dividing the  $z$ -value by the square root of  $n$  (the size of the sample).

### 4.3. Predictors

Based on the overview of the literature, pilot studies and metadata available in the corpus, several predictors besides the interpreter's sex have been identified as potentially influencing cognitive load and disfluencies and have therefore been chosen as predictors for the present study. Two predictors are categorical and were manually added to each transcription: source language and target language (either English, French or Dutch). The other predictors are continuous and were measured thanks to a tailor-made script after having been manually identified in the transcription. The Ear-Voice Span was measured by manually applying time tags to equivalent words uttered by the speaker and the interpreter and by calculating the duration between the two time tags. Similarly to other studies (Christoffels & De Groot 2004; Timarová et al. 2014), the scale used for EVS in this study is centiseconds, as the transcription tool used for the corpus (EXMARaLDA Partitur-Editor) does not allow for precise identification at the scale of milliseconds. Items chosen as time tags are of any grammatical category (substantives, nouns, verbs, etc.). The source speaker's and interpreter's delivery rate were measured as the number of words uttered per normalised segment, excluding all types of disfluencies. The source speaker/interpreter speaking time ratio is the total speaking time of the source speaker (silent pauses excluded) divided by the interpreter's total speaking time (silent pauses excluded) and gives an impression of how much the interpreter actually speaks compared to the source speaker. Finally, all types of disfluencies were also measured for the source speaker in the same way as for the interpreter and his or her influence on the interpreter's disfluencies were measured. Moreover, the influence of the other disfluencies produced by the interpreter was analysed for each type of disfluency (e.g. the influence of the number of false starts produced by the interpreter on the number of filled pauses uttered

by the interpreter). In total for each type of disfluency the effect of 16 predictors was analysed: interpreter's sex, source and target languages, Ear-Voice Span, source speaker/interpreter speaking time ratio, source speaker's and interpreter's delivery rates, number of filled pauses, lengthenings, false starts, silent pauses and the total duration of silent pause.

Mann-Whitney U tests are performed on each type of disfluencies to identify potential sex differences. In order to assess the influence of the 16 predictors, generalised linear mixed models were conducted with R for each type of disfluency. Mixed models allow for each interpretation to be included as a random variable in order to control for idiosyncratic effects. Contrary to disfluencies, which are measured both as averages per minute and per normalised segment, predictors are analysed in one condition only: frequency per normalised segment. Per-segment measurements are more accurate than averages measured for a whole speech. Given that most predictors are measured per 10 seconds, the generalised linear mixed model was performed on the 10-second segments for disfluencies. Moreover, the authors want to assess the influence of predictors on the actual number of disfluencies instead of an average number of disfluencies. Since there are numerous zero values for each measurement of articulated disfluencies per 10 seconds, the count data Poisson regression was used in order to assess the influence of predictors on each type of disfluency. This method is indeed more robust with data containing numerous zero values. However, for the total duration of silent pauses, the data are continuous and a linear regression was used instead of a Poisson regression.

The results are presented through the  $\beta$ -coefficient, which indicates the individual contribution of each predictor to the regression model. The  $\beta$ -coefficient also indicates to what degree each predictor affects the outcome if the effects of all other predictors are held constant. They are interpreted differently between categorical and continuous variables. For continuous variables, each time the predictor increases by one unit, the disfluency increases or decreases by the number of units indicated by the beta value. For categorical predictors, one of the variables is taken as the reference variable and the others are compared to it. In order to estimate the goodness of fit of the model, the marginal  $R^2$  and the conditional  $R^2$  are also mentioned. The marginal  $R^2$  is the variance explained by the fixed effects, while the conditional  $R^2$  is the variance explained by both the fixed and the random effects.

## 5. Results

### 5.1. Descriptive Statistics

As mentioned in Section 2, disfluencies were measured in two conditions: (1) normalised frequency per minute for each text and (2) normalised frequency per standardised segment of 10 seconds. Table 12.2 shows

the number of disfluencies and the delivery rate in two conditions (normalised per minute and normalised per 10 seconds) for source speakers and interpreters.

The most frequent type of disfluency in both interpreting and spontaneous speech is the silent pause followed by the filled pause. Interpreters clearly produce more disfluencies (filled pauses, lengthening and false starts) than source speakers. Counting only these three types of disfluencies, about 7.8% of uttered tokens are disfluent in interpreting, compared to about 2.6% for original speakers. The frequency of disfluencies in our corpus therefore appears to be slightly lower than that recorded in the experimental study by Christodoulides & Lenglet (2014). The average total duration of silent pauses is very similar between original speakers and interpreters, but the number of pauses is higher in spontaneous speech, which means that interpreters produce fewer but longer pauses than original speakers.

### 5.2. *Sex Differences in Interpreters' Disfluencies*

Table 12.3 shows the descriptive statistics for all types of disfluencies per interpretation and per sex and the results of the ten one-tailed Mann-Whitney U tests performed on the five types of disfluency. The null hypothesis is that the distribution of the disfluency in both conditions will be identical. The significance level for p-values is 0.05.

Significant p-values were found for lengthenings ( $p = .009$ ) and the total duration of silent pauses ( $p = .001$ ), with higher values for men but small effect sizes (respectively 0.18 and 0.23). No differences were found for the remaining disfluencies: number of filled pauses ( $p = .076$ ), false starts ( $p = .109$ ) and silent pauses ( $p = .052$ ).

### 5.3. *Influence of Predictors*

Generalised linear mixed models were conducted on each type of disfluencies with all predictors and the random effect “interpretation”. On top of the full model, the influence of predictors was also measured in single models between each predictor and the EVS.

In order to assess the actual influence of the predictors on the production of disfluencies at a given moment, the data used for the regressions are the 10-second measurements. Given the requirements for the Poisson regression, the data used as dependent variables (i.e. the disfluencies) are the count data (the raw frequencies of disfluencies per normalised segment). The sample size is therefore different from previous analysis, as some short and long segments were removed from the dataset. The predictors are the normalised values, as no differences were found when using the count data for the predictors. The predictors included in the regression are the following: interpreter's sex, source and target languages, Ear-Voice



Table 12.2 Disfluencies and delivery rate for source speakers and interpreters<sup>4</sup>

Variables	Source speakers			Interpreters		
	Mean	% of uttered tokens	SD	Mean	% of uttered tokens	SD
Number of words	158.2	97.5%	19.4	142.1	91.2%	17.4
per 10s	26.08	97.6%	5.14	23.63	92.1%	5.98
Number of filled pauses	2.61	1.7%	1.18	7.52	5.3%	4.63
per 10s	0.38	1.5%	0.00	1.06	4.5%	1.23
Number of lengthening	0.46	0.3%	0.00	3.36	2.4%	2.70
per 10s	0.07	0.3%	0.00	0.54	2.3%	0.85
Number of false starts	0.91	0.6%	0.85	1.50	1.1%	1.19
per 10s	0.16	0.6%	0.00	0.25	1.1%	0.54
Number of silent pauses	22.91		22.82	18.64		4.20
per 10s	3.80		3.90	3.10		1.31
Average total duration of silent pauses (in sec)	10.28		9.82	10.58		3.60
per 10s	1.75		1.66	1.78		1.11

Taylor & Francis  
Not for distribution

Table 12.3 Descriptive statistics and sex differences in disfluencies<sup>5</sup>

<i>Disfluency</i>	<i>Sex</i>	<i>Mean</i>	<i>Median</i>	<i>SD</i>	<i>Mann-Whitney U</i>
Number of filled pause	F	6.95	6.25	4.20	U 3548
	M	8.10	7.33	4.10	p .076 Effect size (r) 0.11
Number of lengthening	F	3.01	2.27	2.79	U 3225.5
	M	3.71	3.39	2.56	p .009 Effect size (r) 0.18
Number of false starts	F	1.60	1.37	1.29	U 3619.5
	M	1.41	1.14	1.08	p .109 Effect size (r) 0.09
Number of silent pauses	F	18.98	18.90	3.94	U 3482
	M	18.29	17.31	4.44	p .052 Effect size (r) 0.12
Average total length of silent pause (in sec)	F	9.68	9.47	2.73	U 2982.
	M	11.49	10.89	4.11	p .001 Effect size (r) 0.23

Span, source speaker/interpreter speaking time ratio, source speaker's and interpreter's delivery rate, number of filled pauses, lengthenings, false starts, silent pauses and the total duration of silent pauses.

### 5.3.1. *Interpreter's Filled Pauses*

The sixteen predictors were included in the generalised linear mixed model. Results are presented in Table 12.4.

The marginal  $R^2$  is 0.159 and the conditional  $R^2$  is 0.304. One predictor did yield significant p-value in the single model, but do not have significant p-value in the full model: duration of silent pauses with  $p = 0.997$  ( $p = .001$  for the single model). Five predictors do not have significant p-values in both models: interpreter's language: Dutch ( $p = .104$ ), source speaker's language ( $p = .390$  and  $p = .258$ ), source speaker's false starts ( $p = 0.653$ ), lengthenings ( $p = .682$ ), filled pauses ( $p = .051$ ).

Table 12.4 Parameter Estimates for interpreter's filled pauses per 10 seconds

Predictors	$\beta$	Std. Error	<i>z</i> value	Sig.
<b>Significant predictors</b>				
Interpreter's sex: male	0.097	0.046	2.091	.037
Interpreter's sex: female	0			
Interpreter's language: French	0.223	0.057	3.919	<.001
Interpreter's language: English	0			
EVS	0.001	0.000	7.417	<.001
Interpreter's delivery rate	-0.057	0.004	-15.657	<.001
Interpreter's false starts	0.108	0.026	4.154	<.001
Interpreter's lengthenings	0.165	0.017	9.740	<.001
Interpreter's number of silent pauses	0.059	0.014	4.364	<.001
Interpreter's duration of silent pauses	-0.229	0.038	-6.047	<.001
Source speaker/interpreter speaking time ratio	-0.405	0.194	-2.083	.037
Source speaker's delivery rate	0.022	0.004	5.608	<.001
Source speaker's number of silent pauses	-0.041	0.015	-2.730	.006
<b>Not significant predictors</b>				
Interpreter's language: Dutch	0.091	0.056	1.626	.104
Source speaker's language: Dutch	0.076	0.067	1.130	.258
Source speaker's language: French	-0.062	0.072	-0.860	.390
Source speaker's false starts	-0.016	0.037	-0.450	.653
Source speaker's duration of silent pauses	-0.229	0.038	-6.047	.997
Source speaker's filled pauses	0.039	0.020	1.953	.051
Source speaker's lengthenings	0.021	0.051	0.409	.682

Three predictors did not yield significant *p*-values when included individually in the model but are significant in the full model: interpreter's sex with *p* = .037 (*p* = .265 for the single model), interpreter's number of silent pauses with *p* < .001 (*p* = .803 for the single model) and source speaker's delivery rate with *p* < .001 (*p* = .073 for the single model). Seven predictors (and partially the interpreter's language) have significant *p*-values in both models: interpreter's language French (*p* < .001), EVS (*p* < .001), interpreter's delivery rate *p* < .001), false starts (*p* < .001), lengthenings (*p* < .001), duration of silent pauses (*p* < .001), source speaker/interpreter speaking time ratio (*p* = .037) and source speaker's number of silent pauses (*p* = .006).

Female interpreters produce fewer filled pauses than male interpreters. French interpreters produce more filled pauses than English interpreters, while Dutch as a target language is not significant. Five predictors are associated with increased production of filled pauses by the interpreter:

the EVS, the interpreter's number of false starts, lengthenings and silent pauses and the source speaker's delivery rate. Four predictors are negatively associated with the occurrence of filled pauses: the interpreter's delivery rate and duration of silent pauses, the source speaker's number of silent pauses and the source speaker/interpreter speaking time ratio.

### 5.3.2. *Interpreter's Lengthenings*

Sixteen predictors were included in the generalised linear mixed model. Results are presented in Table 12.5.

The marginal  $R^2$  is 0.186, and the conditional  $R^2$  is 0.250. Three predictors did yield significant p-values when included individually in the model but are not significant in the full model: interpreter's false starts with  $p = .173$  ( $p = .001$  for the single model), source speaker/interpreter speaking time ratio with  $p = .318$  ( $p = .007$  for the single model) and source speaker's number of silent pauses with  $p = .314$  ( $p = .037$  for the single

Table 12.5 Parameter Estimates for interpreter's lengthenings per 10 seconds

	$\beta$	Std. Error	z-value	Sig.
<b>Significant predictors</b>				
Interpreter's sex: male	0.257	0.058	4.415	<.001
Interpreter's sex: female	0			
Interpreter's language: French	0.281	0.073	3.870	<.001
Interpreter's language: English	0			
Source speaker's language: French	-0.505	0.091	-5.535	<.001
Source speaker's language: English	0			
EVS	0.001	0.000	4.167	<.001
Interpreter's delivery rate	-0.087	0.005	-17.095	<.001
Interpreter's filled pauses	0.172	0.017	10.353	<.001
Interpreter's number of silent pauses	-0.047	0.019	-2.505	.012
Interpreter's duration of silent pauses	-0.230	0.046	-5.007	<.001
Source speaker's delivery rate	0.016	0.006	2.974	.003
Source speaker's filled pauses	0.097	0.026	3.794	<.001
<b>Not significant predictors</b>				
Interpreter's language: Dutch	-0.115	0.079	-1.460	.144
Source speaker's language: Dutch	-0.103	0.078	-1.317	.188
Interpreter's false starts	0.052	0.038	1.362	.173
Source speaker/interpreter speaking time ratio	-0.216	0.217	-0.998	.318
Source speaker's false starts	-0.015	0.051	-0.290	.772
Source speaker's lengthenings	0.078	0.066	1.181	.238
Source speaker's number of silent pauses	-0.020	0.020	-1.007	.314
Source speaker's duration of silent pauses	0.009	0.009	1.027	.305

model). Five predictors do not have significant p-values in both models: source speaker's duration of silent pauses ( $p = .305$ ), interpreter's language Dutch ( $p = .144$ ), source speaker's language Dutch ( $p = .188$ ), source speaker's number of false starts ( $p = .772$ ) and lengthenings ( $p = .238$ ).

Eight predictors (as well as the interpreter's and source speaker's languages to some extent) have significant p-values in both models: interpreter's sex ( $p < .001$ ), interpreter's language French ( $p < .001$ ), original speaker's language French ( $p < .001$ ), EVS ( $p < .001$ ), interpreter's delivery rate ( $p < .001$ ), number of filled pauses ( $p < .001$ ), silent pauses ( $p = .012$ ) and average duration of silent pauses ( $p < .001$ ), original speaker's delivery rate ( $p = .003$ ) and number of filled pauses ( $p < .001$ ).

Female interpreters produce fewer lengthenings than male interpreters. French interpreters produce more lengthenings than English interpreters, but Dutch as a target language is not significant. English as a source language triggers more lengthenings than French; Dutch is not significant. Four predictors are associated with increased production of lengthenings by the interpreter: EVS, interpreter's number of filled pauses, source speaker's number of filled pauses and delivery rate. The remaining three predictors are negatively associated with the production of lengthenings: interpreter's delivery rate, number and duration of silent pauses.

### 5.3.3. *Interpreter's False Starts*

Sixteen predictors were included in the generalised linear mixed model. Results are presented in Table 12.6.

The marginal  $R^2$  is 0.055 and the conditional  $R^2$  is 0.086. Two predictors did yield significant p-values when included individually in the model but are not significant in the full model: interpreter's lengthenings with  $p = .227$  ( $p < .001$  for the single model) and source speaker/interpreter speaking time ratio with  $p = .734$  ( $p = .001$  for the single model). Eight predictors do not have significant p-values in both models: interpreter's sex ( $p = .075$ ), source speaker's language French ( $p = .100$ ), interpreter's number of silent pauses ( $p = .540$ ), source speaker's delivery rate ( $p = .463$ ), source speaker's filled pauses ( $p = .056$ ), false starts ( $p = .978$ ), silent pauses ( $p = .127$ ) and lengthenings ( $p = .928$ ).

One predictor is not significant in the single model but is significant in the full model: source speaker's language Dutch with  $p = .005$  ( $p = .081$  for the single model). Five Predictors have significant p-values in both models: interpreter's language French ( $p < .001$ ) and Dutch ( $p < .001$ ), EVS ( $p = .006$ ), interpreter's delivery rate ( $p < .001$ ), filled pauses ( $p < .001$ ), and duration of silent pauses ( $p < .001$ ) and source speaker's duration of silent pauses ( $p = .019$ ).

English interpreters seem to produce more false starts than Dutch and French interpreters, while English as a source language triggers more false starts than Dutch. Several predictors are positively associated with the

Table 12.6 Parameter Estimates for interpreter's false starts per 10 seconds

<i>Predictors</i>	$\beta$	<i>Std. Error</i>	<i>z value</i>	<i>Sig.</i>
<b>Significant predictors</b>				
Interpreter's language: French	-0.754	0.105	-7.186	<.001
Interpreter's language: Dutch	-0.871	0.099	-8.781	<.001
Interpreter's language: English	0			
Source speaker's language: Dutch	-0.318	0.112	-2.838	.005
Source speaker's language: English	0			
EVS	0.001	0.000	2.738	.006
Interpreter's delivery rate	-0.032	0.008	-4.353	<.001
interpreter's filled pauses	0.116	0.026	4.411	<.001
Interpreter's duration of silent pauses	-0.236	0.070	-3.354	<.001
Source speaker's duration of silent pauses	-0.030	0.013	-2.337	.019
<b>Not significant predictors</b>				
Interpreter's lengthenings	0.047	0.039	1.207	.227
Source speaker's lengthening	0.010	0.111	0.090	.928
Interpreter's sex	-0.137	0.077	-1.784	.075
Source speaker's language: French	-0.195	0.118	-1.653	.100
Interpreter's number of silent pauses	0.017	0.028	0.615	.540
Source speaker's delivery rate	-0.006	0.008	-0.734	.463
Source speaker's filled pauses	-0.084	0.044	-1.910	.056
Source speaker's false starts	-0.002	0.075	-0.028	.978
Source speaker's number of silent pauses	0.045	0.029	1.527	.127
Source speaker/interpreter speaking time ratio	-0.111	0.324	-0.342	.734

production of false starts: EVS and interpreter's number of filled pauses. Other predictors have a negative influence on the production of false starts: the interpreter's delivery rate and duration of silent pauses, as well as the source speaker's duration of silent pauses.

#### 5.3.4. *Interpreter's Silent Pauses*

As a reminder, two types of measurements were analysed for silent pauses: the frequency of silent pauses per normalised segment and the average total duration of silent pauses per normalised segment (i.e. not the average duration of a single silent pause, but the average total duration of all silent pauses per normalised segment). The same sixteen predictors were analysed for both types of measurement, but as the total duration of silent pauses is a continuous variable (and not a count data), a linear regression was used instead of a Poisson regression. The presentation of the results is therefore different.

The results for the frequency of silent pauses are presented in Table 12.7.

Table 12.7 Parameter Estimates for interpreter's number of silent pauses

Predictor	$\beta$	Std. Error	<i>z</i> value	Sig.
<b>Significant predictors</b>				
Interpreter's sex: male	-0.077	0.020	-3.787	<.001
Interpreter's sex: female	0			
Interpreter's language: French	-0.100	0.028	-3.565	<.001
Interpreter's language: English	0			
Interpreter's number of filled pauses	0.031	0.008	3.793	<.001
Interpreter's duration of silent pauses	0.274	0.020	13.761	<.001
Source speaker's number of filled pauses	-0.040	0.012	-3.410	<.001
Source speaker's number of silent pauses	0.044	0.008	5.667	<.001
Source speaker's duration of silent pauses	-0.026	0.004	-7.310	<.001
Source speaker/interpreter speaking time ratio	-0.742	0.098	-7.601	<.001
<b>Not significant predictors</b>				
Source speaker's number of lengthening	-0.026	0.029	-0.878	.380
Interpreter's language: Dutch	-0.045	0.026	-1.731	.083
Source speaker's language: French	0.053	0.029	1.805	.071
Source speaker's language: Dutch	0.024	0.029	0.826	.409
Interpreter's delivery rate	0.001	0.002	-0.010	.992
Interpreter's lengthenings	-0.018	0.012	-1.508	.132
Interpreter's false starts	0.007	0.016	0.446	.655
Source speaker's delivery rate	0.001	0.002	0.493	.622
Source speaker's false starts	0.010	0.021	0.475	.635
EVS	-0.000	0.000	-1.736	.083

The marginal  $R^2$  is 0.116 and the conditional  $R^2$  is 0.135. Three predictors did yield significant p-values when included individually in the model but are not significant in the full model: source speaker's language French with  $p = .071$  ( $p < .001$  for the single model), interpreter's delivery rate with  $p = .992$  ( $p < .001$  for the single model) and lengthenings with  $p = .132$  ( $p = .005$  for the single model). Six predictors do not have significant p-values in both models: interpreter's language Dutch ( $p = .083$ ), source speaker's language Dutch ( $p = .409$ ), interpreter's false starts ( $p = .655$ ), source speaker's delivery rate ( $p = .622$ ), source speaker's false starts ( $p = .635$ ), EVS ( $p = .083$ ) and source speaker's lengthenings ( $p = .380$ ).

Two predictors did not yield significant p-values when included individually in the model but are significant in the full model: interpreter's number of filled pauses with  $p < .001$  ( $p = .511$  for the single model) and source speaker's filled pauses with  $p < .001$  ( $p = .135$  for the single model). Five

predictors (and the interpreter's language to some extent) have significant p-values in both models: interpreter's sex ( $p < .001$ ), interpreter's language French ( $p < .001$ ), interpreter's duration of silent pauses ( $p < .001$ ), source speaker/interpreter speaking time ratio ( $p < .001$ ), source speaker's number of silent pauses ( $p < .001$ ) and average duration ( $p < .001$ ).

Female interpreters seem to produce more silent pauses than male interpreters. French interpreters produce fewer silent pauses than English interpreters. Three predictors are associated with increased production of silent pauses by the interpreter: interpreter's number of filled pauses and duration of silent pauses and source speaker's number of silent pauses. The remaining three predictors are negatively associated with the production of silent pauses: source speaker's number of filled pauses and duration of silent pauses and the source speaker/interpreter speaking time ratio.

The same predictors were analysed for the average total duration of silent pauses and results are found in Table 12.8.

Table 12.8 Effect of predictors on interpreter's duration of silent pauses

Predictor	$\beta$	Std. Error	df	t value	Sig.
Interpreter's sex: male	0.193	0.024	1182.528	8.1322	<.001
Interpreter's sex: female	0				
Interpreter's language: French	-0.079	0.031	2265.815	-2.558	.011
Interpreter's language: Dutch	-0.070	0.028	3455.451	-2.512	.012
Interpreter's language: English	0				
Source speaker's language: French	-0.111	0.038	519.856	-2.936	.004
Source speaker's language: Dutch	-0.101	0.035	864.554	-2.909	.004
Source speaker's language: English	0				
Interpreter's delivery rate	-0.058	0.002	4351.998	-31.041	<.001
Interpreter's filled pauses	-0.093	0.008	4332.238	-11.659	<.001
Interpreter's lengthenings	-0.094	0.011	4358.546	-8.260	<.001
Interpreter's false starts	-0.067	0.016	4333.803	-4.214	<.001
Interpreter's number of silent pauses	0.153	0.007	4357.972	21.684	<.001
Source speaker's delivery rate	0.014	0.002	4149.168	6.305	<.001
Source speaker's false starts	0.053	0.021	4325.042	2.576	.01
Source speaker's filled pauses	0.075	0.012	4049.241	6.369	<.001
Source speaker's lengthenings	0.123	0.029	4360.578	4.306	<.001
Source speaker's number of silent pauses	-0.064	0.008	4084.211	-8.046	<.001
Source speaker's duration of silent pauses	0.083	0.003	4120.363	29.162	<.001
Source speaker/interpreter speaking time ratio	2.249	0.037	4311.460	60.899	<.001
EVS	0.000	0.000	4271.201	2.735	.006



Three predictors did not yield significant p-values when included individually in the model but are significant in the full model: interpreter's language Dutch with  $p = .012$  ( $p = .735$  for the single model), source speaker's language Dutch with  $p = .004$  ( $p = .466$  for the single model) and source speaker's number of silent pauses with  $p < .001$  ( $p = .529$  for the single model). Thirteen predictors (and the interpreter's and source speaker's language to some extent) have significant p-values in the full model.

Male interpreters produce longer silent pauses than female interpreters. English interpreters produce longer silent pauses, while English as a source language triggers longer silent pauses. Eight predictors are associated with increased production of lengthening by the interpreter: the EVS, the interpreter's number of silent pauses, the source speaker's duration of silent pauses, delivery rate, frequency of false starts, filled pauses and lengthenings as well as the source speaker/interpreter speaking time ratio. The remaining five predictors are negatively associated with the production of lengthening: interpreter's delivery rate, frequency of filled pauses, lengthenings, false starts and the source speaker's number of silent pauses.

## 6. Discussion

### 6.1. Sex Differences

This study confirms results found in the literature: interpreters produce more filled pauses (Plevoets & Defrancq 2016), lengthenings (Tissi 2000), false starts (Pöchhacker 1995; Tissi 2000) and longer silent pauses (Tissi 2000; Cecot 2001; Ahrens 2005; Christodoulides 2013) than source speakers. Counting ~~only these three types of disfluencies~~, about 7.8% of uttered tokens are disfluent in interpreting, compared to about 2.6% for source speeches. The average total duration of silent pauses per minute is very similar for source speakers and interpreters, but source speakers produce a higher number of silent pauses than interpreters, which means that their silent pauses are shorter than interpreters' silent pauses.

The literature showed that men produce more fillers and repeats than women (Shriberg 1996; Bortfeld et al. 2001). However Cecot (2001) found that female interpreters use more filled pauses and men use more and longer unfilled pauses. The results of the Mann-Whitney U test performed on our data give mitigated results. Male interpreters produce more lengthenings and have a higher total duration of silent pauses but no differences are found for the other types of disfluencies.

The mixed models confirm the results of the Mann-Whitney U test as they show that male interpreters produce more lengthenings and longer duration of silent pauses. However, the regressions also show that male interpreters produce more filled pauses and fewer silent pauses. As a reminder, the mixed models were performed on the per 10-second measurements while the Mann-Whitney U test was conducted per interpretation.

Replace  
« : »  
with  
« . »

Replace  
strikethrough  
text with  
« uttered  
disfluencies  
only »

The difference observed might therefore be due to the difference in measurement and to the fact that the influence of other predictors is included in the regression, while it is not in the Mann-Whitney U tests.

## 6.2. Influence of Predictors

A summary of the influence of the 16 predictors on all disfluencies can be found in Table 12.9.

Table 12.9 Summary of the influence of predictors on disfluencies (beta values)

	<i>Filled pause</i>	<i>Lengthening</i>	<i>False start</i>	<i>Silent pause</i>	<i>Duration of silent pause</i>
Interpreter's sex: male	0.097	0.257		-0.077	0.193
Interpreter's sex: female	0	0		0	0
Interpreter's language: Dutch			-0.871		-0.070
Interpreter's language: French	0.223	0.281	-0.754	-0.100	-0.079
Interpreter's language: English	0	0	0	0	0
Source speaker's language: Dutch			-0.318		-0.101
Source speaker's language: French		-0.505			-0.111
Source speaker's language: English		0	0		0
EVS	0.001	0.001	0.001		0.001
Interpreter's delivery rate	-0.057	-0.087	-0.032		-0.058
Interpreter's filled pauses		0.172	0.116	0.031	-0.093
Interpreter's lengthenings	0.165				-0.094
Interpreter's false starts	0.108				-0.067
Interpreter's number of silent pauses	0.059	-0.047			0.153
Interpreter's duration of silent pauses	-0.229	-0.230	-0.236	0.274	
Source speaker/interpreter speaking time ratio	-0.405			-0.742	2.249
Source speaker's delivery rate	0.022	0.016			0.014
Source speaker's filled pauses		0.097		-0.040	0.075
Source speaker's lengthenings					0.123
Source speaker's false starts					0.053
Source speaker's number of silent pauses	-0.041			0.044	-0.064
Source speaker's duration of silent pauses			-0.030	-0.026	0.083

Female interpreters produce fewer filled pauses and lengthenings and have a shorter average duration of silent pauses but produce more silent pauses than male interpreters. There is no sex difference for false starts. French and English as target languages are statistically significant for all types of disfluencies.

While French interpreters produce more filled pauses and lengthenings than English interpreters, they produce fewer false starts and silent pauses and have a shorter duration of silent pauses. Dutch interpreters produce fewer false starts and have shorter duration of silent pauses than English interpreters, but no significant differences are found for the other types of disfluencies. When it comes to the influence of source languages, French and Dutch trigger shorter duration of silent pauses than English, Dutch triggers fewer false starts and French triggers fewer lengthenings than English. The fact that Dutch as a source language is not significant for most disfluencies and triggers fewer false starts and shorter duration of silent pauses somehow contradicts the literature according to which language pairs with an SOV constituent order in the source language (such as Dutch) increase cognitive load (Seeber 2011) and therefore might trigger more disfluencies. French as a source language seems therefore not to trigger many disfluencies, while English does.

The Ear-Voice Span is mentioned in the literature as a predictor of disfluencies (Tóth 2011) and its significance is confirmed for four types of disfluencies: filled pauses, lengthenings, false starts and the duration of silent pauses. The positive effect of EVS can have two potential explanations: (1) the more the interpreter hesitates, the longer it takes to produce a sentence, the longer the EVS become and (2) a longer EVS is often associated with high cognitive load; similarly disfluencies are also associated with high cognitive load. Their positive correlation could therefore be interpreted in terms of both variables' individual association with cognitive load. The interpreter's delivery rate is negatively associated with all disfluencies except for the frequency of silent pauses. This seems logical, as disfluencies will tend to decrease the fluency, and therefore the delivery rate, of the interpreter (as a reminder, articulated disfluencies are not included in the word count for delivery rate). The source speaker/interpreter speaking time ratio has a negative influence on two types of disfluencies (number of filled and silent pauses) but a positive influence on the duration of silent pauses. This means that when the source speaker speaks more than the interpreter, the interpreter produces fewer filled pauses and silent pauses, but has a longer duration of silent pauses. Less speaking time for the interpreter logically means less opportunity for disfluencies. Similarly, if the duration of silent pauses increases, it means that the speaking time decreases, since silent pauses are not included in the speaking time.

The production of filled pauses is increased by the production of lengthenings, false starts and silent pauses by the interpreter, and also triggers an increase in the production of the other types of disfluencies (except

for the duration of silent pauses). However the production of lengthenings and false starts do not mutually trigger an increase. The production of lengthenings is decreased by silent pauses while lengthenings have no influence on false starts. The interpreter's duration of silent pauses is negatively influenced by all types of uttered disfluencies, and vice versa. This might simply be explained by the fact that longer duration of silent pauses decreases the time dedicated to speech and therefore also reduces the number of articulated disfluencies. Similarly, the production of silent pauses logically increases the duration of silent pauses.

The source speaker's delivery rate was mentioned in the literature (Plevoyets & Defranco 2016; Seeber 2011) as increasing the cognitive load and the number of filled pauses. Our data show that this predictor indeed has a positive influence on three types of disfluencies: filled pauses, lengthenings and the duration of silent pauses. The source speaker's uttered disfluencies have a positive influence on the duration of silent pauses by the interpreter. The source speaker's number of filled pauses has a positive influence on the interpreter's lengthening, but a negative influence on the production of silent pauses. The source speaker's number of lengthening and false starts only has a positive influence on the duration of silent pauses. The source speaker's number of silent pauses decreases the production of filled pauses and the duration of silent pauses, but increases the frequency of silent pauses.

## 7. Conclusion

The aim of this study was to identify the variables determining the frequency of four types of disfluencies in interpretations: lengthenings, filled pauses, false starts and silent pauses. We particularly focused on the interpreter's sex as a predictor because the literature indicates that women and men perform differently on linguistic tasks. In many of the studies, women are found to perform better than men. Meta-studies, on the other hand, generally fail to confirm these observations. Several other predictors known to increase cognitive load in interpreters and, therefore, likely to trigger disfluencies in interpreters, were also examined: these are variables relating to language pairs (source language and target language), properties of the source and the target texts (delivery rate) and their relation in time (EVS) and, finally, to the occurrence of other types of disfluencies both in speakers and in interpreters).

If female interpreters produce fewer disfluencies because of their higher verbal skills, it might mean that they are less subjected to cognitive load. The results show that interpreters indeed produce more filled pauses, lengthenings, false starts and longer silent pauses than source speakers. This confirms the general idea that disfluencies are caused by cognitive load: interpreting is assumed to be cognitively more demanding than spontaneous speech and is therefore expected to be more prone to

disfluencies. When it comes to sex differences and women's greater fluency, the results also tend to confirm our tentative hypothesis.

Both the Mann-Whitney U test and the mixed models show that the hypothesis is confirmed for lengthening and the duration of silent pauses but that no difference exists for false starts. Results are more mitigated for filled pauses and silent pauses but given the increased accuracy of the mixed models, both in terms of measurement and inclusion of other predictors, we can consider that its results are more reliable and that men also produce more filled pauses, but fewer silent pauses.

It is important to remember that the higher frequency of disfluencies in male interpreters is not necessarily a bad sign. As mentioned in the introduction, disfluencies are not viewed as errors but as solutions to errors in speech planning (Betz et al. 2015), and they can play a positive role by helping the identification of upcoming words for example (Howell & Young 1991; Fox Tree & Schrock 1999; Brennan & Schober 2001; Fox Tree 2001; Arnold et al. 2003; Arnold et al. 2004; Ferreira et al. 2004). It could, however, be useful for trainers to pay particular attention to the types of disfluencies produced by students and determine what the causes are and if a particular attention is needed to fix them. As for the textual and contextual variables explored in our study, as potential triggers of cognitive load and therefore of disfluencies, our results show that they do not always have an influence on the production of disfluencies.

The following predictors tend to increase the production of most disfluencies: the Ear-Voice Span, source speaker's and interpreter's number of filled pauses and source speaker's delivery rate. Other predictors are negatively associated with most disfluencies: interpreter's delivery rate and duration of silent pauses. The other predictors yielded mixed results. The source speaker/interpreter speaking time ratio decreases the production of filled and silent pauses but increases the duration of silent pauses. Interpreter's lengthenings and false starts increase the production of filled pauses but decrease the duration of silent pauses. The interpreter's number of silent pauses increases filled pauses and the duration of silent pauses, but decreases lengthenings. Source speaker's lengthenings and false starts only increase the duration of silent pauses, while the source speaker's number of silent pauses decreases the number of filled pauses and the duration of silent pauses.

Dutch as a target or source language is only significant in some cases, and in non-predictable ways: different constituent orders, as between Dutch and one of the other languages, do not seem to trigger disfluencies. English seems to trigger the most disfluencies and French the least. French interpreters produce more filled pauses and false starts but fewer lengthenings and silent pauses.

It is important to mention that some triggers of disfluencies have not been taken into account in this research project because they are difficult to operationalise and require a depth of analysis that could not be

attained in the framework of this research: new information (Clark & Fox Tree 2002; Arnold et al. 2003), heavy constituents (Swerts 1998; Arnold et al. 2000; Watanabe et al. 2008) or long sentences (Oviatt 1995; Shriberg 1996; Gile 2008; Chmiel & Mazur 2013), the position of disfluencies in sentences (Boomer 1965; Shriberg 1996), target lexical density (Plevoets & Defrancq 2016) and the topic or domain (Schachter et al. 1991). The results might therefore be different when these predictors are included. Moreover, the aim of this study was to compare several types of disfluencies between male and female interpreters and not to give a comprehensive overview of the types of disfluencies in interpreting. As a consequence, not all types of disfluencies were taken into account, such as repetitions, grammatical errors, articulation rate and alterations in voice quality (Jameson et al. 2009; Yap 2012; Schuller & Batliner 2013). Obviously, the inclusion of more variables would greatly enhance our understanding of disfluencies and the associated cognitive load in interpreters. We would also wholeheartedly welcome complementary experimental research on disfluencies in interpreting to confirm or contradict our corpus-based findings in highly controlled circumstances.

## Notes

1. Throughout this paper, the authors refer to the interpreter's sex and not the interpreter's gender. While the expression "gender" is commonly used, notably in reference to gender differences, it generally refers to an individual's self-conception and role within society. In fact gender differences studies tend to focus on communicative and linguistic differences (Coates 1993; Chambers & Trudgill 1998) while the present study analyzes the cognitive aspects and therefore only takes the subjects' biological sex into account and makes no assumptions on their gender.
2. A *lengthening* is defined as "a marked prolongation of one or more phones (often limited to one syllable), resulting in above-average syllable and word duration" (Betz et al. 2015; Betz & Wagner 2016). A false start is an utterance that is aborted and then restarted. Lexical identical repetitions happen when the speaker produces the same lexical form multiple times in a row.
3. According to the estimations of the Organisation Intersex International Europe, there is a probability that 1% of the interpreters in the corpus are intersex. Unfortunately there is no possibility of knowing whether interpreters in the corpus are intersex. While the authors are conscious of this possibility, they consider that the size of the corpus means that this low probability does not obviate much of the discussion.
4. Sample size for per-min values is 180 and 4467 for per 10-sec values.
5. N was 180 (90 for females and 90 for males).

Add a space  
before 180

## References

- Aerts, Annelies, van Mierlo, Pieter, Hartsuiker, Robert, Hallez, Hans, Santens, Patrick & De Letter, Miet. 2013. Neurophysiological investigation of phonological input: Aging effects and development of normative data. *Brain and Language* 125(3). 253–263.

- Ahrens, Barbara. 2005. Prosodic phenomena in simultaneous interpreting: A corpus-based analysis. *Interpreting* 7(1). 51–76.
- Arnold, Jennifer E., Fagnano, Maria & Tanenhaus, Michael K. 2003. Disfluencies signal thee, um, new information. *Journal of Psycholinguistic Research* 3. 25–36.
- Arnold, Jennifer E., Tanenhaus, Michael. K., Altmann, Rebecca J. & Fagnano, Maria. 2004. The old and thee, uh, new: Disfluency and reference resolution. *Psychological Science* 15. 578–582.
- Arnold, Jennifer E., Wasow, Thomas, Losongco, Anthony & Ginstrom, Ryan. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language* 76(1). 28–55.
- Bachy, Sylviane, Dister, Anne, Francard, Michel, Geron, Geneviève, Giroul, Vincent, Hambye, Philippe, Simon, Anne Catherine & Wilmet, Régine. 2007. Conventions de transcription régissant les corpus de la banque de données VALIBEL. ([www.uclouvain.be/cps/ucl/doc/valibel/documents/conventions\\_valibel\\_2004.pdf](http://www.uclouvain.be/cps/ucl/doc/valibel/documents/conventions_valibel_2004.pdf)). (Accessed 10.10.2015).
- Baes, Eva. 2012. *Analyse des différences entre interprètes masculins et féminins au Parlement européen en termes de décalage* [Analysis of the differences in décalage between male and female interpreters at the European Parliament]. Master thesis. Hogeschool Gent, Ghent.
- Bakti, Maria. 2009. Speech disfluencies in simultaneous interpreting. In De Crom, Dries (ed.), *(Trans)formation of Identities: Selected Papers of the CETRA Research Seminar in Translation Studies 2008*. 1–18. Leuven: CETRA.
- Barik, Henri C. 1975. Simultaneous interpretation: Qualitative and linguistic data. *Language and Speech* 16(3). 237–270.
- Bendazzoli, Claudio & Sandrelli, Annalisa. 2005. An approach to corpus-based interpreting studies: Developing EPIC (European Parliament Interpreting Corpus). In Gerzymisch-Arbogast, Heidrun & Nauert, Sandra (eds.), *Proceedings of the Marie Curie Euroconferences MuTra: Challenges of Multidimensional Translation*, Saarbrücken. 2–6 May.
- Bernardini, Silvia, Ferraresi, Adriana, Russo, M. Mariachiara, Collard, Camille & Defrancq, Bart. 2018. Building interpreting and intermodal corpora: A how to for a formidable task. In *Making Way in Corpus-Based Interpreting Studies*. 21–42. Singapore: Springer.
- Betz, Simon & Wagner, Petra. 2016. Disfluent lengthening in spontaneous speech. In Jokisch, Oliver (ed.), *Elektronische Sprachsignalverarbeitung (ESSV), Studentexte zur Sprachkommunikation*. 135–144. Dresden: TUD Press.
- Betz, Simon, Wagner, Petra & Schlangen, David. 2015. Micro-structure of disfluencies: Basics for conversational speech synthesis. In *Interspeech. 2222–2226. Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech) 2015*. <https://pub.uni-bielefeld.de/download/2758941/2764759>
- Bock, Kathryn & Levelt, Willem. 1994. Language production: Grammatical encoding. In Gernsbacher, Morton Ann (ed.), *Handbook of Psycholinguistics*. 945–984. San Diego: Academic Press.
- Bolla-Wilson, Karen & Bleecker, Margit. 1986. Influence of verbal intelligence, sex, age, and education on the Rey Auditory Verbal Learning Test. *Developmental Neuropsychology* 2. 203–211.
- Boomer, Donald. 1965. Hesitation and grammatical encoding. *Language and Speech* 8(3). 148–158.

- Born, Marise P., Bleichrodt, Nico & van der Flier, Henk. 1987. Cross-cultural comparison of sex related differences on intelligence tests. *Journal of Cross-Cultural Psychology* 18. 283–314.
- Bortfeld, Heather, Leon, Silvia D., Bloom, Jonathan E., Schober, Michael F. & Brennan, Susan E. 2001. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech* 44(2). 123–147.
- Brennan, Susan E. & Schober, Michael F. 2001. How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language* 44. 274–296.
- Broadbent, Donald E. 1958. *Perception and Communication*. London: Pergamon Press.
- Burnard, Lou. 2002. A retrospective look at the British National Corpus. In Kettemann, Bernhard & Marko, Georg (eds.), *Language and Computers: Studies in Practical Linguistics*. 51–70. Paris: Editions de Boccard.
- Cecot, Michela. 2001. Pauses in simultaneous interpretation: A contrastive analysis of professional interpreters' performances. *The Interpreters' Newsletter* 11. 63–85.
- Chambers, Jack K. & Trudgill, Peter. 1998. *Dialectology*, 2nd Ed. Cambridge: Cambridge University Press.
- Chmiel, Agnieszka & Mazur, Iwona. 2013. Eye tracking sight translation performed by trainee interpreters. In Way, Catherine, Vandepitte, Sonia, Meylaerts, Reine & Bartłomiejczyk, Magdalena. (eds.), *Tracks and Treks in Translation Studies: Selected Papers from the EST Congress*. 189–205. Leuven: Belgium.
- Christodoulides, George. 2013. Prosodic features of simultaneous interpreting. In Mertens, Piet & Simon, Anne Catherine (eds.), *Proceedings of the Prosody: Discourse Interface Conference*. 33–37. Leuven: Belgium.
- Christodoulides, George & Lenglet, Cédric. 2014. Prosodic correlates of perceived quality and fluency in simultaneous interpreting. In *Speech Prosody* 7. 1002–1006. Ireland: Trinity College Dublin.
- Christoffels, Ingrid K. & De Groot, Annette M. 2004. Components of simultaneous interpreting: Comparing interpreting with shadowing and paraphrasing. *Bilingualism: Language and Cognition* 7(3). 227–240.
- Clark, Herbert H. & Fox Tree, Jean E. 2002. Using uh and um in spontaneous speaking. *Cognition* 84. 73–111.
- Coates, Jennifer. 1993. *Women, Men and Language*, 2nd Ed. London: Longman.
- Collard, Philip, Corley, Martin, MacGregor, Lucy J. & Donaldson, David I. 2008. Attention orienting effects of hesitations in speech: Evidence from ERPs. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34. 696–702.
- Corley, Martin & Hartsuiker, Robert. 2011. Why um helps auditory word recognition: The temporal delay hypothesis. *PLoS One* 6(5). 197–192.
- Corley Martin, MacGregor, Lucy J. & Donaldson, David I. 2007. It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition* 105. 658–668.
- Defrancq, Bart. 2013. Women and men interpreting. *Paper presented at the first "Talking to the World" Conference, Newcastle, 12 September*.
- Diriker, Ebru. 2004. *De-/Re-Contextualizing Conference Interpreting: Interpreters in the Ivory Tower?* Amsterdam: John Benjamins.
- Duez, Danielle. 1982. Silent and Non-Silent Pauses in Three Speech Styles. *Language and Speech* 25(1). 11–28.



- Duflou, Veerle. 2016. *Be(com)ing a Conference Interpreter: An Ethnography of EU Interpreters as a Professional Community*. Amsterdam: John Benjamins.
- Engelhardt, Paul E., Corley, Martin, Nigg, Joel T. & Ferreira, Fernanda. 2010. The role of inhibition in the production of disfluencies. *Memory and Cognition* 38(5). 617–628.
- Ferreira, Fernanda, Lau, Ellen F. & Bailey, Karl G. D. 2004. Disfluencies, language comprehension, and tree adjoining grammars. *Cognitive Science* 28. 721–749.
- Fox Tree, Jean E. 1995. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language* 34. 709–738.
- Fox Tree, Jean E. 2001. Listeners' uses of um and uh in speech comprehension. *Memory and Cognition* 29. 320–326.
- Fox Tree, Jean E. & Schrock, Josef C. 1999. Discourse markers in spontaneous speech: Oh what a difference an oh makes. *Journal of Memory and Language* 40. 280–295.
- Gale, Shawn D., Baxter, Leslie, Connor, Donald J., Herring, Anne & Comer, James. 2007. Sex differences on the Rey Auditory Verbal Learning Test and the Brief Visuospatial Memory Test-Revised in the elderly: Normative data in 172 participants. *Journal of Clinical and Experimental Neuropsychology* 29(5). 561–567.
- Garrett, Merrill F. 1975. Syntactic process in sentence production. In Bower, Gordon H. (ed.), *Psychology of Learning and Motivation: Advances in Research and Theory*, Vol. 9. 133–177. New York: Academic Press.
- Gerver, David. 1969. The effect of source language presentation rate on the performance of simultaneous conference interpreters. In Pöchhacker, Franz & Shlesinger, Miriam (eds.), *The Interpreting Studies Readers*. 52–66. London: Routledge.
- Gile, Daniel. 1995. *Basic Concepts and Models for Interpreter and Translator Training*. Amsterdam: John Benjamins. 274p.
- Gile, Daniel. 1997. Conference interpreting as a cognitive management problem. In Danks, Joseph H., Shreve, Gregory M., Fountain, Stephen B. & McBeath, Michael (eds.), *Cognitive Processes in Translation and Interpreting*. 196–214. Thousand Oaks, CA: Sage Publications.
- Gile, Daniel. 2008. Local cognitive load in simultaneous interpreting and its implications for empirical research. *Forum* 6. 59–77.
- Goldman-Eisler, Frieda. 1958. Speech production and the predictability of words in context. *Quarterly Journal of Experimental Psychology* 10. 96–106.
- Goldman-Eisler, Frieda. 1972. Segmentation of input in simultaneous interpretation. *Journal of Psycholinguistic Research* 1. 127–140.
- Gósy, Mária. 2007. Disfluencies and Self-monitoring. *Govor* 26. 91–110.
- Guyette, Thomas & Baumgartner, John. 1988. Stuttering in the adult. In Lass, Norman J., McReynolds, L., Northern, Jerry L. & Yoder, D. (eds.), *Handbook of Speech-Language Pathology and Audiology*. 640–654. Philadelphia, PA: B. C. Decker Inc.
- Halpern, Diane F., Benbow, Camilla P., Geary, David C., Gur, Ruben C., Hyde, Janet S. & Gernsbacher, Morton. A. 2007. The science of sex differences in science and mathematics. *Psychological Science in the Public Interest* 8(1). 1–51.
- Harness, Ashley, Jacot, Lorri, Scherf, Shauna, White, Adam & Warnick, Jason E. 2008. Sex differences in working memory. *Psychological Reports* 103(1). 214–218.

- Hartsuiker, Robert J. & Kolk, Herman H. J. 2001. Error monitoring in speech production: A computational test of the perceptual loop theory (Statistical data included). *Cognitive Psychology* 42(2). 113~~(45)~~.
- Hedges, Larry V. & Nowell, Amy. 1995. Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science* 269(5220). 41–45.
- Herlitz, Agneta, Airaksinen, Eija & Nordström, Eva. 1999. Sex differences in episodic memory: The impact of verbal and visuospatial ability. *Neuropsychology* 13. 590–597.
- Herlitz, Agneta, Nilsson, Lars-Goran & Bäckman, Lars. 1997. Gender differences in episodic memory. *Memory and Cognition* 25. 801–811.
- Hickey, Sarah. 2018. *The Female Perspective on the Feminisation of Conference Interpreting*. Master thesis, NUI Galway, Galway.
- Hines, Melissa. 1990. Gonadal hormones and human cognitive development. In Balthazart, J. (ed.), *Brain and Behaviour in Vertebrates 1: Sexual Differentiation, Neuroanatomical Aspects, Neurotransmitters, and Neuropeptides*. 51–63. Basel, Switzerland: Karger.
- Hirnstein, Marco, Andrews, Lisa C. & Hausmann, Markus. 2014. Gender-stereo typing and cognitive sex differences in mixed- and same-sex groups. *Archives of Sexual Behavior* 43. 1663–1673.
- Howell, Peter & Young, Keith. 1991. The use of prosody in highlighting alteration in repairs from unrestricted speech. *Quarterly Journal of Experimental Psychology* 43. 733–758.
- Hyde, Janet S. 2005. The gender similarities hypothesis. *American Psychologist* 60. 581–592.
- Hyde, Janet S. & Linn, Marcia C. 1988. Gender differences in verbal-ability: A meta-analysis. *Psychological Bulletin* 104. 53–69.
- Jameson, Anthony, Kiefer, Juergen, Müller, Christian, Grossmann-Hutter, Barbara, Wittig, Frank & Rummer, Ralf. 2009. Assessment of a user's time pressure and cognitive load on the basis of features of speech. In Crocker, Matthew & Siekmann, Jörg (eds.), *Resource-Adaptive Cognitive Processes*. 171–204. Berlin: Springer.
- Keith, Timothy Z., Reynolds, Matthew R., Patel, Puja G. & Ridley, Kristen P. 2008. Sex differences in latent cognitive abilities ages 6 to 59: Evidence from the Woodcock-Johnson III Tests of cognitive abilities. *Intelligence* 36(6). 502–525.
- Kimura, Doreen & Seal, Brooke. 2003. Sex differences in recall of real or non-sense words. *Psychological Reports* 93. 263–264.
- Klaudy, Kinga. 2004. *Bevezetés a fordítás elméletébe* [Introduction to the theory of translation]. Budapest: Scholastica.
- Kramer, Joel, Kaplan, Edith, Delis, Dean, O'Donnell, Louise & Prifitera, Aurelio. 1997. Developmental sex differences in verbal learning. *Neuropsychology* 11. 577–584.
- Kutas, Marta & Hillyard, Steven A. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature* 307(5947). 161–163.
- Lass, Norman & Puffenberg, Marcia. 1971. A comparative study of rate evaluations of experienced and inexperienced listeners. *Quarterly Journal of Speech* 57(1). 89–93.
- Levelt, Willem. 1983. Monitoring and self-repair in speech. *Cognition* 14. 41–104.
- Levelt, Willem. 1989. *Speaking: From Intention to Articulation*. Cambridge: The MIT Press.

Replace  
 strikethrough  
 number with  
 «-157 »

- Levelt, Willem. 1999. Models of word production. *Trends in Cognitive Sciences* 3(6). 223–232.
- Lim, Hyang-Ok. 2005. Meeting students' expectations. *Forum* 1(3). 175–204.
- Liu, Minhua, Schallert, Diane L. & Carroll, Patrick J. 2004. Working memory and expertise in simultaneous interpreting. *Interpreting* 6(1). 19–42.
- Loonstra, Ann, Tarlow, Alison & Sellers, Alfred. 2001. COWAT metanorms across age, education and gender. *Applied Neuropsychology* 8(3). 161–166.
- Macgregor, Lucy, Corley, Martin & Donaldson, David I. 2009. Not all disfluencies are equal: The effects of disfluent repetitions on language comprehension. *Brain and Language* 111(1). 36–45.
- Macgregor, Lucy, Corley, Martin & Donaldson, David I. 2010. Listening to the sound of silence: Investigating the consequences of disfluent silent pauses in speech for listeners. *Neuropsychologia* 48(14). 3982–3992.
- MacKay, Donald G. 1973. Aspects of the theory of comprehension, memory and attention. *The Quarterly Journal of Experimental Psychology* 25(1). 22–40.
- Magnifico, Cédric & Defrancq Bart. 2016. Impoliteness in interpreting: A question of gender? *Translation and Interpreting* 8(2). 26–45.
- Magnifico, Cédric & Defrancq Bart. 2017. Hedges in conference interpreting: The role of gender. *Interpreting* 19(1). 21–46.
- Magnifico, Cédric & Defrancq, Bart. 2019. Self-repair as a norm-related strategy in simultaneous interpreting and its implications for gendered approaches to interpreting. *Target* 31(3). 352–377.
- Maitland, Scott, Herlitz, Agneta, Nyberg, Lars Bäckman Lars & Nilsson, Lars-Göran. 2004. Selective sex differences in declarative memory. *Memory and Cognition* 32(7). 1160–1169.
- Majeres, Raymond L. 1999. Sex differences in phonological processes: Speeded matching and word reading. *Memory and Cognition* 27. 246.
- Mason, Marianne. 2008. *Courtroom Interpreting*. Maryland: University Press of America.
- Mead, Peter. 2000. Control of pauses by trainee interpreters in their A and B languages. *The Interpreters' Newsletter* 10. 89–102.
- Miller, David I. & Halpern, Diane F. 2014. The new science of cognitive sex differences. *Trends in Cognitive Sciences* 18(1). 37–45.
- Nooteboom, Sieb G. 1969. The tongue slips into patterns. In Sciarone, A. G., van Essen, A. J. & van Raad, A. A. (eds.), *Nomen (Society): Leyden Studies in Linguistics and Phonetics*. 114–132. The Hague: Mouton Publishers.
- Oviatt, Sharon. 1995. Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language* 9(1). 19–35.
- Pio, Sonia. 2003. The relation between ST delivery rate and quality in simultaneous interpretation. *The Interpreters' Newsletter* 12. 69–100.
- Plevoets, Koen & 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.
- Plevoets, Koen & Bart Defrancq. 2018. The cognitive load of interpreters in the European Parliament: A corpus-based study of predictors for the disfluency *ub(m)*. *Interpreting* 20(1). 1–29.
- Pöchhacker, Franz. 1995. Slips and shifts in simultaneous interpreting. In Tommola, J. (ed.), *Topics in Interpreting Research*. 73–90. Turku: The University of Turku, Centre for Translation and Interpreting.

- Postma, Albert. 2000. Detection of errors during speech production: A review of speech monitoring models. *Cognition* 77(2). 97–132.
- Rouvier, Mickael, Dupuy, Grégor, Gay, Paul, Khoury, Elie, Merlin, Teva & Meignier, Sylvain. 2013, August. An open-source state-of-the-art toolbox for broadcast news diarization. *Interspeech*, Lyon.
- Ryan, Rachael. 2015. Why so few men?: Gender imbalance in conference interpreting. (<http://aiic.net/p/7347>). (Accessed 07.06.2018).
- Schachter, Stanley, Christenfeld, Nicholas, Ravina, Bernard & Bilous, Frances. 1991. Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology* 60. 362–367.
- Schuller, Björn W. & Batliner, Anton M. 2013. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Hoboken: Wiley. 344p.
- Seeber, Kilian. 2011. Cognitive load in simultaneous interpreting: Existing theories: New models. *Interpreting* 13(2). 176–204.
- Seeber, Kilian. 2013. Cognitive load in simultaneous interpreting: Measures and Methods. *Target: International Journal of Translation Studies* 25(1). 18–32.
- Setton, Robin. 1999. *Simultaneous Interpretation: A Cognitive and Pragmatic Analysis*. Amsterdam and Philadelphia: John Benjamins. 399p.
- Shattuck-Hufnagel, Stefanie. 1979. Speech errors as evidence for a serial-ordering mechanism in sentence production. In Cooper, W. E. & Walker, E. C. T. (eds.), *Sentence Processing*. 295–342. Hillsdale, NJ: Lawrence Erlbaum.
- Shewan, Cynthia, M. & Henderson, Vicki L. 1988. Analysis of spontaneous language in the older normal population. *Journal of Communication Disorders* 21(2). 139–154.
- Shlesinger, Miriam. 1998. Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta* 43(4). 486–493.
- Shriberg, Elizabeth. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Doctoral dissertation. University of California, Berkeley.
- Shriberg, Elizabeth. 1996. Disfluencies in Switchboard. In *Proceedings of the International Conference on Spoken Language Processing*. 11–14. Philadelphia: Addendum.
- Spencer, Steven J., Steele, Claude M. & Quinn, Diane M. 1999. Stereotype threat and women's math performance. *Journal of Experimental Social Psychology* 35. 4–28.
- Swerts, Marc. 1998. Filled pauses as markers of discourse structure. *Journal of Pragmatics* 30. 485–496.
- Timarová, Sarka, Čeňková, Ivana, Meylaerts, Reine, Hertog, Erik, Szmalec, Arnau & Duyck, Wouter. 2014. Simultaneous interpreting and working memory executive control. *Interpreting* 16(2). 139–168.
- Tissi, Benedetta. 2000. Silent pauses and disfluencies in simultaneous interpretation: A descriptive analyzes. *The Interpreters' Newsletter* 10. 103–127.
- Tóth, Andrea. 2011. Speech disfluencies in simultaneous interpreting: A mirror on cognitive processes. *SKASE Journal of Translation and Interpretation* 5(2). 23–31.
- Trahan, Donald E. & Quintana, Joseph W. 1990. Analysis of gender effects upon verbal and visual memory performance in adults. *Archives of Clinical Neuropsychology* 5(4). 325–334.

- Voor, John B. & Miller, Joseph M. 1965. The effect of practice on the comprehension of worded speech. *Speech Monographs* 32. 452–455.
- Watanabe, Michiko, Hirose, Keikichi, Den, Yasuharu & Minematsu, Nobuaki. 2008. Filled pauses as cues to the complexity of up-coming phrases for native and non-native listeners. *Speech Communication* 50(2). 81–94.
- Welford, Alan T. 1952. The “psychological refractory period” and the timing of high speed performance: A review and a theory. *British Journal of Psychology* 43. 2–19.
- Yairi, Ehud & Ambrose, Noline A. 1992. Longitudinal study of stuttering in children: A preliminary report. *Journal of Speech and Hearing Research* 35. 755–760.
- Yap, Tet Fei. 2012. *Speech Production under Cognitive Load: Effects and Classification*. Doctoral dissertation. University of New South Wales, Cardiff.

Taylor & Francis  
Not for distribution

Proof

**Taylor & Francis**  
Not for distribution

Proof