# Informational load as a trigger for disfluencies in interpreting. A corpus-based regression analysis.

**Koen Plevoets**
Ghent University
koen.plevoets@ugent.be

**Bart Defrancq**
Ghent University
bart.defrancq@ugent.be

## Abstract

One of the major aspects of an interpreting task is the high cognitive load for the interpreter. Gile (1995) pinpoints the interpreter's lack over the conceptual content and his reduced background knowledge (in comparison to the speaker) as potential sources of problems during interpreting, and for simultaneous interpreting he lists the additional obstacles of the lack of control over the original speech rate and the mutually detrimental influence of the speaking and listening task. Psycholinguistic research (Clark et al. 2002; Corley et al. 2008; Watanabe et al. 2008) has revealed that information overload is prone to give rise to disfluencies in the utterance, e.g. *uh* or *uhm*. In light thereof, it is no surprise that disfluencies figure prominently in interpreting (Bakti 2009, Tissi 2000, Tóth 2011). However, previous research is inconclusive as to whether disfluencies occur to the same extent in interpreting as in spontaneous speech, as no systematic quantitative comparison has yet been undertaken.

This paper will analyse the relation between interpreting, informational load and disfluencies in a corpus of interpreted language as compared to a corpus of spontaneous speech. The corpus of interpreted language was compiled at Ghent University between 2010 and 2013. It consists of French, Spanish and Dutch interpreted speeches in the European Parliament from 2006 until 2008. The audio fragments were transcribed according to the guidelines of the VALIBEL corpus (Bachy et al. 2007). For our purposes, a sub-corpus of French source speeches and their Dutch interpretations was selected, amounting to a total corpus size of 140 000 words. The sub-corpus has additionally been annotated for lemmas, parts-of-speech and chunks (Van de Kauter et al. 2013). The corpus which serves as the reference for spontaneous speech is the sub-corpus of political debates of the Spoken Dutch Corpus (Oostdijk 2000). This sub-corpus contains 220 000 words of Netherlandic Dutch and 140 000 words of Belgian Dutch, which were collected between 1998 and 2003, and it is annotated for lemmas and parts-of speech.

In both corpora, each sentence (or 'discourse unit') was subsequently coded for informational measures such as lexical density and syntactic depth, in order to capture the informational load experienced by the speakers or interpreters. The measurement of lexical density is based on the POS-tags, where all nouns, non-auxiliary verbs, adjectives and adverbs are counted as content words and all pronouns, auxiliary verbs, prepositions, conjunctions and determiners are counted as function words (all remaining interjections and fillers are treated as a rest category). The coding for syntactic depth was done manually: each sentence was screened and annotated for the number of different syntactic subordinations, the maximal degree of syntactic subordination and the average degree of syntactic subordination. The last step in the data retrieval consisted in counting the number of the disfluencies *uh* and *uhm* per sentence, as the aim of the analysis is to predict the frequency of the disfluencies on the basis of the informational load of each sentence. Due to the heavy skewness of the frequency data, it was decided to run the analysis by means of Robust Regression (Maronna et al. 2006).

The results confirm the intuitive assumptions in that the data for interpreted Dutch exhibit a different pattern from both the data of the French source language and of spontaneous Dutch, which in turn are very similar. The observations for interpreted Dutch show a distinctly positive effect of the informational measures on the frequency of the disfluencies: the higher the informational load is during interpreting, the more this results in disfluencies by the interpreter. A striking finding for both the French source data and the spontaneous Dutch data is that the effect in either case is negative. This result may be attributed to the highly prepared nature of the parliamentary speeches, which are sometimes read out verbatim from a written text. The same patterns moreover show up in separate analyses for *uh* and *uhm*. All these findings point to interesting prospects for further research. The immediate next step will be to take account of the position of the disfluency in the utterance, as we conjecture that disfluencies tend to occur before informationally heavy chunks in non-interpreted language, but at the onset of whole utterances in interpreted language.

## References

Bachy, S., Dister, A., Francard, M., Geron, G., Giroul, V., Hambye, P., Simon, A.C. and Wilmet, R. 2007. *Conventions de transcription régissant les corpus de la banque de données VALIBEL*. University of Louvain-la-Neuve. Available online at http://www.uclouvain.be/cps/ucl/doc/valibel/documents/conventions_valibel_2004.PDF.

Bakti, M. 2009. “Speech disfluencies in simultaneous interpreting”. In D. De Crom (ed.) *(Trans)formation of identities: Selected papers of the CETRA research seminar in translation studies 2008*. Leuven: CETRA, 1-18.

Clark, H.H. and Fox Tree, J.E. 2002. “Using *uh* and *um* in spontaneous speaking”. *Cognition* 84: 73-111.

Corley, M. and Stewart, O.W. 2008. “Hesitation disfluencies in spontaneous speech: The meaning of *um*”. *Language and Linguistics Compass* 2: 589-602.

Gile, D. 1995. *Basic concepts and models for interpreter and translator training*. Amsterdam: John Benjamins.

Maronna, R., Martin, D. and Yohai, V. 2006. *Robust statistics: Theory and methods*. Hoboken, New Jersey: John Wiley and Sons.

Oostdijk, N. 2000. “The Spoken Dutch Corpus: Overview and first evaluation”. *Proceedings of the Second International Conference on Language Resources and Evaluation*: 887-894.

Tissi, B. 2000. “Silent pauses and disfluencies in simultaneous interpretation: A descriptive analysis”. *The Interpreters’ Newsletter* 10: 103-127.

Tóth, A. 2011. “Speech disfluencies in simultaneous interpreting: A mirror on cognitive processes”. *SKASE Journal of Translation and Interpretation* 5: 23-31.

Van de Kauter, M., Coorman, G., Lefever, E., Desmet, B., Macken, L. and Hoste, V. 2013. “LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit”. *Computational Linguistics in the Netherlands Journal* 3: 103-120.

Watanabe, M., Hirose, K. Den, Y. and Minematsu, N. 2008. “Filled pauses as cues to the complexity of up-coming phrases for native and non-native listeners”. *Speech Communications* 50: 81-94.