7.11.24, 12.00-13.30 Uhr, IfL SR links – Lunch & Linguistics-Spezial

Cinthia Sayuri Misaka - *Doctoral student in Linguistics at Universidade Estadual de Campinas. Currently, ZUKOnnect Fellow at University of Konstanz. Master in Neurolinguistics at Universidade Estadual de Campinas. Bachelor in Letters at Universidade de São Paulo and specialization in Japanese language and studies by the Tokyo University of Foreign Studies in Japan (MEXT scholarship recipient). My interests include L1 and L2 acquisition, language development in typical and atypical children (ASD), bilingualism/multilingualism, and psycholinguistics.*

Using Large Language Models (LLMs) in child's corpus analysis: a Brazilian-Portuguese case

This study explores the application of Large Language Models (LLMs) in corpus analysis of children's spoken data, specifically focusing on the automated extraction and classification of morphological and syntactic elements. By comparing the performance of three LLMs—GPT 3.5 Turbo (OpenAI), Gemini (Google), and Claude 3 - Haiku (Anthropic)— the study discusses on their capacity to analyze and categorize possessive pronouns along with related linguistic elements such as articles and nouns. The child's corpus, drawn from CEDAE's (Center of Cultural Documentation Alexandre Eulalio) naturalistic speech database, features interactions between a child (21-28 months old) and their caregiver. The results indicate that GPT 3.5 Turbo (90.4%) and Claude-Haiku (89.6%) outperformed Gemini (77.1%) in accurately extracting and classifying possessive pronouns and related linguistic elements. While manual review of the annotations remains essential, the use of LLMs can significantly reduce the time required for annotation, particularly in large language corpora. Also, we highlight the role played by the prompt in order to provide clear, structured guidance for the models on how to process linguistic data. In this context, the prompt serves as an instruction set that directs the LLMs to focus on specific language structures, such as possessive pronouns and related elements like articles and nouns. This ensures that the models carry out consistent and accurate analyses in alignment with linguistic research goals.

Fernando Sabatin: *Master's degree student in Linguistics at Universidade Estadual de Campinas. Bachelor in Letters (Portuguese) at Universidade Estadual de Campinas with an Academic Exchange at Umeå University (Sweden, Fall of 2019). Member of LAPROS — Acquisition, Processing, and Syntax Laboratory (Unicamp — tinyurl.com/ymbphbhy); Co-Founder & Coordinator of InCognitus — Study Group on Language and Consciousness. Research interests: lexical access; embedded words; mental lexicon; language processing; neurophysiological methods.*

Lexical access: Stimulus Onset Asynchrony effects on the activation of final embedded words

Embedded words are words embedded within larger words and share no semantic relationship with them (e.g. *pain* in *champagne* — final embedded word). Two interesting questions about this phenomenon are (i) whether an embedded word reaches lexical activation when the carrier word is heard or read, and (ii) whether it influences word processing. This phenomenon offers an

opportunity to better understand the dynamics of word segmentation in spoken language use and, moreover, allows us to test models of lexical access. For instance, the cohort model (Marslen-Wilson 1987) predicts that final embedded words are not activated, while TRACE (McClelland & Elman 1986) does. Studies on this topic report inconsistent findings: some report facilitatory priming (Isel & Bacri 1999; Luce & Cluff 1998; Vroomen & De Gelder 1997; Shillcock 1990), whereas others report no priming effects (Norris et al. 2006; Gow & Gordon 1995) or even inhibitory priming (Shatzman 2006; Marslen-Wilson et al. 1994). I argue that these inconsistencies may result, at least partially, from distinctions in the experimental design, such as the variation in Stimulus Onset Asynchrony (SOA): if the interval between prime and target is too long (> 300 ms), the embedded word may already be inhibited by the time the target is presented, hence no priming effect is observed. To address this issue, I designed a cross-modal priming experiment, where a portuguese carrier word containing a final embedded word, e.g., *fé* ('faith') in *café* ('coffee') is inserted into a spoken sentence as a prime for a visual target (written word). I compare conditions with and without priming effects, contrasting SOA = 0 ms and SOA = 500 ms (2×2 Latin square — priming x SOA). The goal is to assess whether embedded words are indeed activated, and if so, if they are rapidly inhibited. Moreover, sentences contain a predictability effect (evaluated through a CLOZE test) to favor the carrier words. In this presentation, I present the literature discussion, implications for lexical access models, and preliminary results of the ongoing experiment. (Funding: FAPESP 2021/00377−3)