Comparing Neural Language Models' island sensitivity to that of human participants in Basque

Michelle Suijkerbuijk
michelle.suijkerbuijk@ru.nl

A much-debated question in linguistics is whether learning language, specifically grammar, requires a language-specific learning capacity or can be achieved from input alone. Neural language models (NLMs) can greatly influence this debate as they learn solely from input and their inductive biases, without any built-in linguistic representations [1]. Recent research has already investigated the NLM's behavior on various grammatical phenomena, including syntactic island constraints, which block filler-gap dependency formation in structures such as relative clauses like (1) [2].

(1) *What$_i$ did you meet [$_{RC}$ the scientist who invented _$_i$]?

Although island violations rarely occur in language input and NLMs cannot fall back on built-in linguistic knowledge, research shows they can successfully model these constraints [3]. These studies, however, almost exclusively focus on English, while NLMs should work equally well across languages to model our universal language learning system [4]. In my previous work, I have already showed that the NLMs are not successful in Dutch, a language different to English in word order [5]. The study I am working on now goes even further beyond English by investigating NLMs' behavior on the relative clause island constraint in Turkish, Basque and Czech; morphologically complex languages with different word orders from different language families. In this talk, I will focus mostly on the Basque study.

While I have already collected human acceptability judgments in Turkish and Czech, I will test around 50 Basque participants during my stay here and 5 NLMs per language after my stay. In all languages, models and humans are tested on the same sentences, manipulated for the presence of island, gap and filler. We predict humans to show that having both a filler and a gap inside an island (e.g., *Pastela, Ikerrek sukaldean egin duen sukaldaria entzun du*) is degraded compared to inside a non-island (e.g., *Pastela, Ikerrek entzun du sukaldariak sukaldean egin duela*), and will investigate whether NLMs can model this human behavior. During this presentation, I will present the methodology of the Basque study in more detail, and I will already show some preliminary results from 10 Basque speakers tested online. Moreover, I will explain what different potential results (1) can tell us about the facilitation or hindrance of morphological complexity and non-English word order on NLMs' grammatical learning, and (2) mean for the debate about how humans learn language.

[1] Contreras Kallens, Pablo, Ross Deans Kristensen-McLachlan & Morten H. Christiansen. 2023. Large Language Models demonstrate the potential of statistical learning in language. *Cognitive Science* 47(3). e13256. https://doi.org/10.1111/cogs.13256.

[2] Sprouse, Jon, Matt Wagers & Colin Phillips. 2012. A test of the relation between working-memory capacity and syntactic island effects. *Language* 88(1). 82-123. https://doi.org/10.1353/lan.2012.0004.

[3] Wilcox, Ethan Gotlieb, Richard Futrell & Roger Levy. 2024. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry* 55(4). 805-848. https://doi.org/10.1162/ling_a_00491.

[4] Bender, Emily M. 2011. On Achieving and Evaluating Language-Independence in NLP. *Linguistic Issues in Language Technology* 6(3). 1-26. https://doi.org/10.33011/lilt.v6i.1239.

[5] Suijkerbuijk, Michelle, Naomi Tachikawa Shapiro, Peter de Swart, Stefan L. Frank. In preparation. The Success of Neural Language Models on syntactic island effects is not universal: strong *wh*-island sensitivity in English but not in Dutch.