Available online at www.sciencedirect.com
ScienceDirect
Computer Speech & Language xxx (2018) xxx-xxx
www.elsevier.com/locate/csl

Language models, surprisal and fantasy in Slavic intercomprehension☆

Q1 Klára Jágrová*, Tania Avgustinova, Irina Stenger, Andrea Fischer
Q2 *Saarland University, Saarbrücken, Germany*

Received 13 February 2017; received in revised form 8 March 2018; accepted 26 April 2018
Available online xxx

Abstract

In monolingual human language processing, the predictability of a word given its surrounding sentential context is crucial. With regard to receptive multilingualism, it is unclear to what extent predictability in context interplays with other linguistic factors in understanding a related but unknown language − a process called intercomprehension. We distinguish two dimensions influencing processing effort during intercomprehension: surprisal in sentential context and linguistic distance. Based on this hypothesis, we formulate expectations regarding the difficulty of designed experimental stimuli and compare them to the results from think-aloud protocols of experiments in which Czech native speakers decode Polish sentences by agreeing on an appropriate translation. On the one hand, orthographic and lexical distances are reliable predictors of linguistic similarity. On the other hand, we obtain the predictability of words in a sentence with the help of trigram language models. We find that linguistic distance (encoding similarity) and in-context surprisal (predictability in context) appear to be complementary, with neither factor outweighing the other, and that our distinguishing of these two measurable dimensions is helpful in understanding certain unexpected effects in human behaviour.

Keywords: Statistical language modelling; Surprisal; Receptive multilingualism; Slavic languages; Sentential context; Think-aloud protocols; Polish; Czech; Reading

1. Introduction

Statistical models are widely used in psycholinguistic modelling of human language (Keller, 2010). Negative log probabilities assigned by statistical models, typically called surprisal scores, correlate well with e.g. human reading times of texts of varying difficulty (Hale, 2001; Levy, 2008) and may thus serve as reasonable indices of the cognitive effort involved in human natural language comprehension. Psycholinguistic and neurolinguistic experiments on cognitive load are usually confined to a monolingual setting − one in which the subjects have native competence in

☆ This paper has been recommended for acceptance by Prof. R. K. Moore.

* Corresponding author.
E-mail address: kjagrova@coli.uni-saarland.de (K. Jágrová), ira.stenger@mx.uni-saarland.de (I. Stenger), afischer@lsv.uni-saarland.de (A. Fischer).

https://doi.org/10.1016/j.csl.2018.04.005

the tested language. Prototypically, the experiments aim to evaluate the relative difference in processing complexity of various formulations that convey effectively the same information. We study the mutual intelligibility of Slavic languages and in contrast to the regular psycholinguistic setting, it is not clear to what extent and in what form such psycholinguistic results translate in case of receptive multilingualism.

In this contribution, we present a qualitative empirical study into the role of sentential context during reading intercomprehension between selected Slavic languages. We hypothesize that both linguistic distance and surprisal based on sentential context influence the processing effort in reading intercomprehension. To investigate the relationship between these two predictors − linguistic distance and surprisal − we discuss three different experiments. In the first experiment, a Croatian (HR) sentence which poses morphosyntactic challenges to Russian native speakers was presented to respondents with Slavic native languages other than HR. They were asked to translate the given sentence into their native language. The results of this experiment indicate that words which are apparently orthographically transparent may influence translations more than within-context surprisal does. In a second experiment, we presented native readers of Czech (CS) with Polish (PL) sentences and elicited translations for these sentences. The CS−PL data was gathered in a series of two-person think-aloud experiments conducted at Charles University in Prague in December 2016. We analyse the stimulus sentences in terms of their orthographic and lexical distance and compare the translations produced in terms of their information density as modelled by trigram Kneser−Ney language models (LMs) (Kneser and Ney, 1995). We find that again, linguistic distance is a critical factor in intercomprehension. However, linguistic distance and in-context surprisal appear to be complementary, with neither factor outweighing the other − our think-aloud protocols reveal that in cases where a word is highly surprising, but also identical to a cognate in their L1 (native language), our test subjects appear to have felt misled by the apparently "weird" context, and instead chose less surprising translations. In addition to the results from the think-aloud translation experiments, we present results from web-based cloze tests with the same stimuli sentences where the translation gaps were placed on the words that turned out to be problematic in the think-aloud experiments. The cloze experiments were conducted over the website freely accessible at http://intercomprehension.coli.uni-saarland.de/en/.

The main purpose of this study is to present a method for estimating the processing difficulty of sentences in reading intercomprehension, using statistical LMs. The qualitative analysis does *not* aim to evaluate a statistically significant number of stimuli in an experiment, but rather to investigate why respondents chose certain translations in certain cases. Results from web-based cloze experiments for the same stimuli are added for a quantitative perspective.

2. Receptive multilingualism and language modelling

Receptive multilingualism, a term often used synonymously for *intercomprehension*, is defined as the ability to understand an unknown but related foreign language without being able to use it actively for speaking or writing (Doyé, 2005). Receptive multilingualism is facilitated by the ability of the human language processing mechanism to quite robustly handle imperfect linguistic signal. As an example, knowing German and English, one can experience practical reading intercomprehension for instance when trying to decipher a Dutch text (e.g. Vanhove, 2014).

Successful intercomprehension is possible and has been well documented and studied for a number of languages. Notable examples are e.g. Danish and Swedish (cf. e.g. Schüppert et al., 2016) or CS and Slovak (e.g. Nábělková, 2007; Golubović, 2016), among others. The mutual intelligibility of certain language combinations, i.e. to what degree and under which circumstances intercomprehension between these languages works, appears to be influenced by a number of linguistic and non-linguistic factors (cf. Gooskens, 2013 for a comprehensive overview of the factors).

2.1. Linguistic distance as a measure for similarity

In research on receptive multilingualism, the *linguistic distance* between two related languages has been tested for being a relatively reliable predictor for their mutual intelligibility (e.g. Golubović and Gooskens, 2015). CS and Slovak, for instance, are very close languages and therefore, mutual intelligibility is possible without any major problems (Nábělková, 2007). Linguistic distance is usually measured on different descriptive levels of languages. Lexical, orthographic, and morphological distances are typically obtained on parallel sets of words or texts (e.g. Golubović and Gooskens, 2015; Golubović, 2016). However, distances of individual words do not inform about the

Please cite this article as: K. Jágrová et al., Language models, surprisal and fantasy in Slavic intercomprehension, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.04.005