COMPREHENSION OF CLOSELY RELATED LANGUAGES: A VISUAL WORLD EYE TRACKING STUDY

Jacek Kudera[1], *Philip Georgis*[1], *Hasan Md Tusfiqur Alam*[1], *Bernd Möbius*[1], *Tania Avgustinova*[1], *Dietrich Klakow*[1,2]

[1]*Department of Language Science and Technology, Saarland University*
[2]*Spoken Language Systems (LSV), Saarland University*
kudera@lst.uni-saarland.de

Abstract: We present results of an eye tracking experiment which aimed at testing sentence comprehension in closely related Slavic languages. Since none of the participants were trained in translation studies or Slavic linguistics, the study illustrates effects of intercomprehension. The participants were exposed to auditory stimuli in Bulgarian, Czech, Polish, and Russian accompanied by a visual scene. The analysis of anticipatory eye movements has shown that native speakers of one Slavic language listening to sentences in another Slavic language, turn their attention to and begin fixating on the referent objects as soon as they identify a predicate. This experiment provides evidence for surprisal-based effects in intercomprehension.

1 Introduction

Intercomprehension among speakers of closely related languages is grounded in the degree of similarity of the languages used in communication. Such a phenomenon depends on structural properties shared by a speaker's L1 and by the perceived L2, as well as on their shared phonological inventories. Therefore, comprehension of spoken stimuli from a closely related language depends on the phonetic distance and the degree of correspondence between the perceived phrase and its equivalent in the native language of the listener. Furthermore, the information-theoretic notion of surprisal [1, 2], which corresponds to processing effort [3], can quantify the degree of difficulty in comprehension among speakers of closely related languages or vernaculars, when they each speak their own L1. The Visual World Paradigm (VWP) format allows for the measurement of language comprehension in real time through anticipatory eye movements [4, 5, 6]. In this study, a VWP eye tracking experiment [7, 8] was employed to investigate the influence of phonetic distance and surprisal on sentence comprehension in a multilingual setting including four Slavic languages, namely, Bulgarian, Czech, Polish, and Russian.

The underlying assumption for speech processing and visual scene inspection was given by Cooper [9], who observed that when participants are simultaneously exposed to spoken and visual stimuli, their eye movements are synchronized with different linguistic events conveyed in the auditory modality. This observation laid the foundation for VWP and allows us to examine whether participants use predicative constraints to shift their attention to the visual referent [10]. Our study proposes an extension of such a paradigm to a multilingual setting. Therefore, with the application of the VWP method, it is possible to examine if participants turn their attention to a visual field representing the direct object of a sentence from a closely related language, perceived in the auditory modality. It is assumed that such an effect can be enhanced by similar surface forms of verbs in both languages, measured by their phonetic resemblance to one another, as well as by a degree of stimulus (un)expectedness.

1.1 Aims and hypotheses

The primary goal of this study was to investigate the comprehension of sentences from a closely related language in the auditory modality with a visually enhanced environment. It is assumed that information extracted at the predicate, coming from a closely related language, can be used to guide eye movements to whichever object in the visual context satisfies the restriction of a perceived predicate. Furthermore, it was hypothesized that comprehension of a sentence, measured by a gaze anticipation effect in the VWP setup, is driven by the information-theoretic notion of surprisal as well as phonetic distance between corresponding predicates in subjects' L1 and the stimulus language. Since previous studies have shown a unilateral pattern of spoken intercomprehension among speakers of the Slavic languages, a secondary aim of this work was to investigate the directionality of sentence comprehension across the language groups.

1.2 Related work

Eye tracking in the VWP has emerged as an important method for understanding real-time language comprehension. Previous studies using this methodology have shown a close alignment between fixations and estimates of lexical activation [11]. Importantly to our study, phonetic similarity has been shown to trigger an attention shift to a referent object [12]. Past VWP studies on non-native spoken word recognition have shown a greater lexical competition for non-native input, but confirmed the unidirectional effect of comprehension, even when investigating less closely languages than the ones tested here, such as Dutch and English [13]. In accordance with previous experiments, which have shown that surprisal is not determined solely by linguistic context but can be also enhanced by a visual environment [6], our study introduced a situated version of surprisal by merging the cross-lingual predictability of a predicate with a corresponding visual scene. The studies mentioned above demonstrate that VWP is a suitable method for investigating verb-mediated referential processing [10] and therefore could be extended to a multilingual comprehension scenario.

2 Method

A VWP study was designed involving four Slavic languages. A scalable webcam eye-tracker, *Webgazer* [14], was used for gaze estimation and the detection of visual field preferences. Subjects were exposed to auditory stimuli presented simultaneously with a visual scene. The participants were instructed to listen to the sentences and look at the pictures. Then, a pseudo-task involving answering a question in their L1 regarding their understanding of the foreign sentence was given. Subjects were informed that pictures can provide clues toward sentence comprehension, therefore they should pay attention to the objects presented on the screen. The start of the next trial was triggered after the user had recorded an answer. Head movements were unrestricted during the recording session. Since syntactic constraints can influence sentence comprehension in a VWP setup [10], in this experiment stimuli were composed of fixed phrases, familiar to the subjects already after the trial session. The experiment setup precluded the use of a chin rest, but the head pose was tracked in the background. To ensure data quality, drift check intervals were presented after each visual scene [8]. In case of a lost mesh, participants were asked to adjust their head position and to recalibrate the eye tracker prior to starting the following trial. The experiment lasted around 25 minutes, although the length varied depending on subjects' ability to maintain a consistent head pose.

2.1 Phonetic distance

Phonetic distances between transcribed verb pairs were calculated using a weighted average of three component measures expressing phonetic dissimilarity on the word level: dissimilarity of consonantal segments (0.5), dissimilarity of vocalic segments (0.3), and dissimilarity of syllable structure (0.2) [15]. This weighting scheme was set according to the hypothesis that consonantal segments are the most salient phonetic factor in an intercomprehension scenario, and may impact the comprehension of related word forms to a relatively greater extent than dissimilarities in the vowel space or syllable structure. Phonetic dissimilarities between individual consonant or vowel pairs were calculated from feature vectors representing each segment's phonological distinctive features. The final component quantifies the dissimilarity of the syllable structure of the two words by calculating the length-normalized Levenshtein distance [16] of the IPA strings encoded as sequences of 'C' (consonant) and 'V' (vowel) characters.

2.2 Surprisal

Surprisal, or self-information, is an information theoretic measure that quantifies the (un)expectedness of a particular outcome, measured in bits, inversely proportional to its probability [1]. More specifically, surprisal is calculated as the negative logarithm (base 2) of the probability of the outcome. Thus, outcomes with higher probabilities produce lower surprisal values, and, conversely, less likely outcomes yield higher surprisal. A perfectly predictable outcome with probability of 1.0 results in a surprisal value of zero. Surprisal of verb pairs was measured according to Word Adaptation Surprisal (WAS) [17], given in Equation 1, applied to aligned phonetic transcriptions in IPA:

$$WAS = -\sum_{i=1}^{n} \log_2(p(wordL1_i|wordL2_i)) \tag{1}$$

where $p(wordL1_i|wordL2_i)$ represents the probability of the i^{th} phonetic segment of a word in a listener's native language given its aligned equivalent $wordL2_i$ in a non-native language.

This assesses the total surprisal of the sounds of one word given aligned equivalents in another word. Probabilities of phonetic correspondences between each pair of languages were extracted automatically from pairwise phonetic alignments of 434 cognate sets in the four languages. WAS has been shown to correlate with intelligibility among Slavic languages [17] and can be interpreted here as a quantification of the processing effort required by a native speaker of one language to comprehend lexical items from a related language. To account for the differing lengths of words, the WAS values were normalized by the length of the phonetic alignment.

2.3 Regions of interest

Four regions of interest were predefined for each visual scene. The four pictures were presented in the corners of a screen at 150 x 150 pixels, equally distanced from the center of the screen, which depicted the agent of each carrier phrase, a girl. The girl in the picture had closed eyes to avoid the suggestion that she was pointing to any one visual field. No extraneous visual elements were included in the pictures to ensure the visual salience of the depicted object. Clip art images depicted easily recognizable nouns of similar complexity. The visual field of each object in the scene was defined by its outermost contour. The images were controlled for visual complexity, quality, and size. The trials were randomized for each recording session, and images were randomly assigned to each of the screen corners. One corresponding object was randomly placed at each visual scene. For instance, the referent object in a Polish sentence *Ona*

chce pić teraz kawę, glossed as follows, was the only drinkable object (a cup of coffee) among the images presented, which were otherwise not objects that can be drunk or eaten.

Ona	*chce*	*pić*	*teraz*	*kawę*
3SG	want-PRS:3SG	drink-INF	now	coffee.ACC:SG

'She wants to drink a coffee now'

2.4 Audio stimuli

Auditory stimuli consisted of fixed SVO-type sentences in four Slavic languages. The intervals from the verb offset to the onset of a filler, and from the filler offset to the object onset, were equalized across the utterances to 200 ms. The verbs and objects were separated by a filler word *now* in the relevant language. All the verbs were transitive and controlled for high collocation strength with their respective direct object. Due to the relatively flexible word order in the Slavic languages, the structure of the stimuli with a filler preceding the object sounded natural and did not violate any syntactic constraints. The audio samples were synthesized with a TTS system for each tested language. The intonation contour was manually optimized to achieve a natural sound in sentences that included two short pauses. Such a segmentation scheme also allowed for equalizing the width of the time window of the analyzed fixations. Only female voices were used in the synthesis and each sample has been verified for intelligibility and naturalness by native speakers of the tested languages.

2.5 Participants

In total, 100 participants (25 native speakers per each tested language) aged 19–57 (mean 28) completed the experiment. All of the participants reported using their declared L1 in everyday communication. Only data from subjects who reported no vision or hearing related difficulties were included in the analysis. Therefore, no glasses were allowed. Participants were recruited via a crowd-sourcing platform and paid for their participation. Importantly, subjects had not received previous training in Slavic studies, linguistics, or translation. Recording sessions in which participants did not complete the entire session, either due to a sudden background lighting change or due to issues with holding a pose, were not analyzed. The data obtained from these uncompleted sessions were discarded.

2.6 Procedure

Prior to the experiment, the eye-tracker needed to be calibrated. In a first step, participants were asked to adjust the background lighting and to sit comfortably to define a center pose. Calibration was conducted through multiple repetitions of projecting a screen with fixation points (15 to 20) while maintaining a center pose. After the calibration was completed and a face mesh successfully cached, participants began the comprehension test.

The participants' task was to answer the question that immediately followed the stimulus and to press a key indicating the appropriate response. The question was given in the native language of participants. For example Czech native speakers listened to sentences in Bulgarian, Polish, or Russian followed by comprehension checks in Czech. A trial session was conducted prior to starting the experiment. During the trial, the participants were also exposed to stimuli in their own native language to ensure a good understanding of the experimental setup. The fixation dots were presented after each trial, along with a short recalibration procedure, if nec-

essary. Instructions for each phase of the experimental setup were given in the native language of participants. The procedure involving exposure to L1 sentences was used as a control.

The time to first fixation [7] on the visual field of the direct object has been taken into account in the analysis. The fixations were analyzed in temporal reference to the linguistic events from the auditory stimuli. The x/y coordinates of fixations were compared with the time to the first fixation on a target visual field. Data concerning fixations on the visual target before the verbal component of the audio was played were discarded from the analysis. Such fixations were treated as random and not triggered by information carried by the sentence predicate, which had not yet been perceived.

2.7 Data analysis

The descriptive statistics for continuous variables were calculated with the *psych* package in R [18]. Then, a Kruskal-Wallis test was conducted to examine the influence of language on stimulus comprehension with a post-hoc Dunn's test and Holm-Bonferroni correction for pairwise multiple comparisons of the ranked data [19]. In the next step, the effects of phonetic distance, surprisal and language of stimuli on the amount of time until the first fixation on the target visual field were explored by applying a generalized linear model with the *stats* package.

3 Results

A Shapiro-Wilks test showed a deviation from the normal distribution. Therefore, non-parametric tests were conducted instead. Basic descriptive statistics for the continuous variables are presented in Table 1. In order to conduct more fine-grained analyses and gain insight into pairwise language similarities, an additional grouping variable was added (coded in *xx-yy* format, where *xx* refers to the subjects' native language and *yy* to the language of the stimuli). The basic descriptive statistics for the grouped data are presented in the supplementary material (section 6: Data availability).

Table 1 – Basic descriptive statistics for continuous variables

Variable	n	M	SD	Mdn	Min	Max	Skew	Kurt	W	*p*
Time	2915	0.44	0.41	0.32	0.00	1.63	0.87	−0.30	0.88	< 0.001
Phon. dist.	3851	0.37	0.12	0.38	0.08	0.64	−0.31	−0.51	0.98	< 0.001
Surprisal	3851	6.56	2.22	6.80	0.62	11.20	−0.33	−0.45	0.98	< 0.001

Time - time until the first fixation on the target object, *phon. dist.* - phonetic distance *n* - sample size, *M* - mean, *SD* - standard deviation, *mdn* - median, *min* - minimum value, *max* - maximum value, *skew* - skewness, *kurt* - kurtosis, *W* - Shapiro-Wilk test, *p* - p-value.

The differences in sub-sample sizes vary between time until the first gaze and the independent variables due to analyzing the first gaze on the visual field of the direct object. If an anticipation effect was not discovered, the measurement was excluded from the analysis. In the next step, a contingency table from cross-classifying factors was created (see supplementary material: Secion 6). In addition, a Chi-squared test was conducted. The results showed a significant relation between the subjects' native languages and the languages of stimuli ($\chi^2(9, 3852) = 1286; p < 0.001$) with large effect size (Cramer's V = 0.33). Therefore, an analysis of simultaneous effects of multiple variables was conducted. The significance of the effects of phonetic distance and surprisal and pairs of languages was investigated by a generalized linear model.

The nominal variable *language pair* consisted of a set of 12 possible values (four native languages times three languages of exposure) and was changed into a set of dichotomous variables (e.g., bg-pl, bg-cs, bg-ru). The regression analysis was conducted using the step algorithm with the bidirectional search mode [20]. Then the model with the lowest value of the Akaike information criterion was determined [21]. The summary statistics are given in the supplementary material (section 6: Data availability).

In this analysis, phonetic distance did not appear to be a significant modifier of attention shift and therefore further analyses regard only the information-theoretic notion of surprisal. The effect measured on trials in which subjects listened to Polish sentences did not reach the threshold of statistical significance. The greatest effect was observed in a group of Czech native speakers exposed to Russian stimuli. A strong effect was also measured for Russian native speakers listening to Bulgarian sentences. In contrast, the least significant effect was observed for Bulgarian and Russian native speakers exposed to Czech sentences.

Additionally, a post-hoc Kruskal-Wallis test was conducted to examine differences between surprisal values across the language pairs. The highest values of surprisal were measured for the Bulgarian-Russian cluster, whereas the lowest were found for Czech and Polish, which are both West Slavic languages. Overall, significant differences were found among twelve language pairs ($H = 2654, p < 0.0001, df = 11$) and the effect size was large ($\eta^2 = 0.69$). Therefore, a post-hoc Dunn's test with Holm-Bonferroni correction for pairwise multiple comparisons was conducted. The results showed that 58 out of total 66 pairwise comparisons were significant. For a detailed pairwise analysis see the supplementary material (section 6: Data availability).

4 Discussion

In line with studies concerning exposure to one's native language only, subjects have shown to be able to identify a predicate from a non-native, but closely related, language. The significance of the effect of surprisal suggests that comprehension of closely related languages can be driven by the notion of stimulus (un)expectedness rather than by a resemblance between phonetic surface forms. The results suggest that subjects were able to identify direct objects better when the Word Adaptation Surprisal between corresponding predicates was low, as opposed to when the phonetic distance between these corresponding word forms was low. This outcome is intriguing, as one might otherwise assume that more similar-sounding words would facilitate intercomprehension to a greater extent than the complexity of sound mappings between languages, which lay listeners may not be aware of. On the other hand, surprisal calculated from regular sound correspondences observed in shared cognates may be able to better distinguish related forms from chance phonetic resemblances, given that unrelated words frequently exhibit chance similarities and that diachronic processes of sound change do not always produce similar-sounding reflexes of inherited phonemes across related languages. Because both cognate and non-cognate verb pairs were included as stimuli, participants needed to first recognize stimuli as cognate to a word in their native language. Therefore, the strong effect of surprisal follows from the fact that it is likely a better predictor of cognacy than a measure of phonetic distance alone.

Interestingly, the asymmetries across the tested languages point to differences in comprehension depending on subjects' L1 and the language of exposure. The intercomprehension pattern reflects the typological division of the Slavic languages only to a limited extent. Strong intelligibility effects were discovered for language pairs that do not belong to the same subgroup of Slavic languages. It was therefore concluded that information received at the verb level was used to direct attention towards the direct object even when exposed to a non-native language.

5 Conclusions

In this study, real-time comprehension of spoken stimuli from a non-native yet closely related language was tested in a visual environment. The collected data support the hypothesis that sentence processing in a closely related, non-native language, as measured by anticipatory eye movements, is driven by the information-theoretic notion of surprisal measured on corresponding predicates. Furthermore, the data obtained in this study exhibit the asymmetrical character of intercomprehension across the four groups of native Slavic speakers.

The data gathered in this study have shown that Slavic native speakers can immediately establish a dependency between the predicate of a sentence and its direct object even if the sentence is perceived in a related, non-native language. Furthermore, the relationship between corresponding predicates can be quantified by means of surprisal, which corresponds to the (un)expectedness of stimuli rather than to the degree of resemblance in their surface forms. The study has demonstrated that information extracted at the predicate successfully guides eye movements to an object in a visual setting which satisfies the verb constraints.

This study has shown that comprehension of a predicate causes the attention to shift towards the object before the onset of the referential noun. However, the effect of the language of stimuli moderates this relation depending on the specific stimuli and the subject's L1. Such a finding is in line with previous investigations on the unidirectional character of intercomprehension. Regardless of the discovered asymmetries, this study supports an argument for a surprisal-driven intelligibility effect among speakers of closely related languages.

6 Data availability

The experimental data are publicly available in the following Open Science Framework repository: https://osf.io/2wsek/

7 Acknowledgements

Research funded by the Deutsche Forschungsgemeinschaft (German Research Foundation), Project ID 232722074 – SFB 1102.

References

[1] SHANNON, C. E.: *A mathematical theory of communication. The Bell system technical journal*, 27(3), pp. 379–423, 1948.

[2] HALE, J.: *A probabilistic earley parser as a psycholinguistic model.* In *Second meeting of the north American chapter of the association for computational linguistics*. 2001.

[3] ANKENER, C. S., M. SEKICKI, and M. STAUDTE: *The influence of visual uncertainty on word surprisal and processing effort. Frontiers in psychology*, 9, pp. 1–17, 2018.

[4] KAMIDE, Y., G. T. ALTMANN, and S. L. HAYWOOD: *The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. Journal of Memory and language*, 49(1), pp. 133–156, 2003.

[5] TANENHAUS, M. K., M. J. SPIVEY-KNOWLTON, K. M. EBERHARD, and J. C. SEDIVY: *Integration of visual and linguistic information in spoken language comprehension. Science*, 268(5217), pp. 1632–1634, 1995.

[6] SEKICKI, M. and M. STAUDTE: *Eye'll help you out! how the gaze cue reduces the cognitive load required for reference processing*. *Cognitive science*, 42(8), pp. 2418–2458, 2018.

[7] DUCHOWSKI, A. T.: *Eye tracking methodology: Theory and practice*. Springer, 2017.

[8] CARTER, B. T. and S. G. LUKE: *Best practices in eye tracking research*. *International Journal of Psychophysiology*, 155, pp. 49–62, 2020.

[9] COOPER, R. M.: *The control of eye fixation by the meaning of spoken language: a new methodology for the real-time investigation of speech perception, memory, and language processing. Cognitive psychology*, 1974.

[10] ALTMANN, G. T. and Y. KAMIDE: *Incremental interpretation at verbs: Restricting the domain of subsequent reference*. *Cognition*, 73(3), pp. 247–264, 1999.

[11] FARRIS-TRIMBLE, A. and B. MCMURRAY: *Test–retest reliability of eye tracking in the visual world paradigm for the study of real-time spoken word recognition*. *Journal of Speech, Language, and Hearing Research*, 56, 2013.

[12] ALLOPENNA, P. D., J. S. MAGNUSON, and M. K. TANENHAUS: *Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models*. *Journal of memory and language*, 38(4), pp. 419–439, 1998.

[13] WEBER, A. and A. CUTLER: *Lexical competition in non-native spoken-word recognition*. *Journal of memory and language*, 50(1), pp. 1–25, 2004.

[14] PAPOUTSAKI, A., P. SANGKLOY, J. LASKEY, N. DASKALOVA, J. HUANG, and J. HAYS: *Webgazer: Scalable webcam eye tracking using user interactions*. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3839–3845. AAAI, 2016.

[15] KUDERA, J., P. GEORGIS, B. MÖBIUS, T. AVGUSTINOVA, and D. KLAKOW: *Phonetic distance and surprisal in multilingual priming: Evidence from slavic*. In *Interspeech 2021*, pp. 3944–3948. 2021. doi:10.21437/Interspeech.2021-1003.

[16] LEVENSHTEIN, V.: *Binary Codes Capable of Correcting Deletions, Insertions and Reversals*. *Soviet Physics Doklady*, 10, p. 707, 1966.

[17] STENGER, I., T. AVGUSTINOVA, and R. MARTI: *Levenshtein distance and word adaptation surprisal as methods of measuring mutual intelligibility in reading comprehension of slavic languages*. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2017"*. Moscow, 2017.

[18] R CORE TEAM: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

[19] DUNN, O. J.: *Multiple comparisons using rank sums*. *Technometrics*, 6(3), pp. 241–252, 1964.

[20] AKAIKE, H.: *Prediction and entropy*. In *Selected Papers of Hirotugu Akaike*, pp. 387–410. Springer, 1998.

[21] AKAIKE, H.: *Information theory and an extension of the maximum likelihood principle*. In *Selected Papers of Hirotugu Akaike*, pp. 199–213. Springer, 1998.