Surprisal in Intercomprehension

Tania Avgustinova

Saarland University, Department of Language Science and Technology
avgustinova@lst.uni-saarland.de

Abstract. Empirically, mutual intelligibility of closely related languages relies on similarities measurable as linguistic distances, which are characteristically symmetric for a given language pair. This contribution illustrates the potential of an information-theoretic modelling of Slavic intercomprehension, particularly in capturing asymmetries observable in receptive multilingualism.

Keywords: Slavic languages, receptive multilingualism, linguistic distance, surprisal, language modelling, web-based experiments.

1 Background

A large-scale interdisciplinary research collaboration at Saarland University[1] (Crocker et al. 2016) investigates the hypothesis that language use may be driven by the optimal utilization of the communication channel. The information-theoretic concepts of *entropy* (Shannon, 1949) and *surprisal* (Hale 2001; Levy 2008) have gained in popularity due to their potential to predict human linguistic behavior. The underlying assumption is that there is a certain total amount of information contained in a message, which is distributed over the individual units constituting it. Capturing this distribution of information is the goal of *surprisal-based modeling* with the intention of predicting the *processing effort* experienced by humans upon encountering these units. The ease of processing linguistic material is thus correlated with its contextually determined predictability, which may be appropriately indexed by Shannon's notion of information.

Multilingualism pervasiveness suggests that human language competence is used quite robustly, taking on various types of information and employing multi-source compensatory and guessing strategies. While it is not realistic to require from every single person to master several languages, it is certainly beneficial to strive and promote a significantly higher degree of receptive skills facilitating the access to other languages. Taking advantage of linguistic similarity – genetic, typological or areal – is the key to acquiring such abilities as efficiently as possible. Awareness that linguistic structures known of a specific language apply to other varieties in which similar phenomena are detectable is indeed essential.

[1] The research presented here is funded by the German Science Foundation (DFG); Project-ID 232722074 –SFB 1102.

1.1 Project INCOMSLAV[2]

Receptive multilingualism, a term often used synonymously for *intercomprehension*, is defined as the ability to understand an unknown but related foreign language while being unable to use it for speaking or writing (Doyé 2005). Successful intercomprehension is possible and has been well documented and studied for a number of languages. It provides an outstanding evidence about the human language processing mechanism as remarkably robust in handling imperfect linguistic signal.

Our ongoing research[3] correlates linguistically established and diachronically motivated similarities between closely related languages, as manifested in degrees of their mutual intelligibility, with conditional entropy and surprisal scores in experimentally observed intercomprehension of written and spoken stimuli. The graphical representation in Fig.1 summarizes its components.

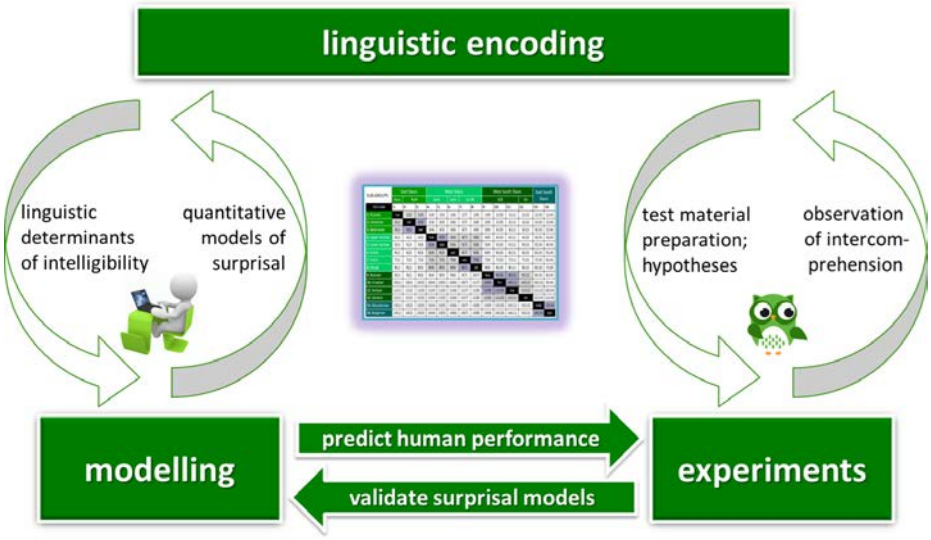

Fig. 1. The INCOMSLAV workflow. For an interactive visualization of currently available results, cf. http://intercomprehension.coli.uni-saarland.de/en/SlavMatrix/Results/

1.2 Information and Surprisal

Traditionally, linguistics has associated the informational content of a sentence or discourse with its semantics, typically expressed using logical, symbolic frameworks: The inherent meaning of words and constituents is combined compositionally to determine

---

[2] The project "Mutual Intelligibility and Surprisal in Slavic Intercomprehension: INCOMSLAV" is part of the Collaborative Research Center 1102 "Information Density and Linguistic Encoding".

[3] Cf. (Fischer et al. 2015, 2016; Stenger et al. 2017; Jágrová et al. 2017; Jágrová et al. 2018; Jágrová & Avgustinova 2019; Stenger et al. 2019; Mosbach et al. 2019; Stenger et al. 2020; Stenger, Jágrová & Avgustinova 2020; Stenger & Avgustinova 2020a,b)

a sentential or discourse message, and support logical inference. Recently, however, there has been considerable interest in a more abstract, probabilistic characterization of the information that is conveyed by an unfolding language signal. One view that has emerged, either tacitly or explicitly, in various areas of linguistics is that speakers manage the information density of the language they produce to make comprehension easier (Hawkins 2004; Levy and Jaeger 2007; Aylett and Turk 2004; Genzel and Charniak 2002). The crucial insight behind many of these proposals is that, while the inherent linguistic complexity and meaning conveyed by a particular expression may be in some sense constant across various usages, the information conveyed by the expression is a function of its predictability: Highly predictable expressions convey less information than more surprising ones, and thus entail lower cognitive effort for comprehenders. Assuming speakers are sensitive to the resource limitations of hearers, decisions regarding the choice of linguistic encoding for a particular message will often be conditioned by the context, with contextually predictable messages or expressions being encoded more densely and surprising material being encoded less densely.

The central role played by predictiveness under this view is given broad support by modern psycholinguistics and computational linguistics. Just as the predictability of a word in a particular context has long been known to influence reading times for people, language models developed for automatic speech recognition, part of speech tagging, and related language processing tasks crucially rely on contextualized probability estimates for linguistic inputs in order to achieve high performance in the face of otherwise pathological ambiguity. Recent research using the visual world paradigm has further demonstrated that listeners actively anticipate what speakers are likely to say next, based on the linguistic and non-linguistic context – see (Staudte et al. 2011) for discussion. Such active prediction offers a natural explanation for the facilitated integration of linguistic input when it is consistent with expectations, and slower processing when it is not.

Information Theory (Shannon 1948) defines the information conveyed by a linguistic unit in terms of its likelihood of occurrence in a particular context – i.e. its predictability. Given the varying constraints that a particular context exerts upon what linguistic unit may follow, predictability is defined in probabilistic terms, as follows:

$$Predicativility(unit_i) = Probability(unit_i|Context) \quad (1)$$

A further result of Information Theory is that the predictability of a unit in context can be used to determine the amount of information that is conveyed by that unit in terms of bits – resulting in a measure commonly known as surprisal – using the following formula:

$$Surprisal(unit_i) = \log \frac{1}{Probability(unit_i|Context)} \quad (2)$$

Two fundamental properties of this characterization are (a) that linguistic events with low probability convey more information than those with high probability, and (b) the information conveyed by an expression (e.g. word) is not determined solely by the expression itself, but the context in which the expression occurs. Stated simply, surprisal captures the intuition that linguistic expressions that are highly predictable, in a given

context, convey less information than those which are surprising. To sum up, surprisal[4] is taken to be the *expectability* of a certain unit in a given context, and is usually interpreted as the amount of information this unit conveys relative to the context. Note that this does not necessarily include capturing the information at hand itself – rather, it suffices to *quantify* its impact relative to the surrounding context:

$$\begin{aligned} Surprisal(unit|Context) &= -\log_2\big(Probability(unit|Context)\big) \\ Total\ surprisal\ of\ a\ message &= \sum Surprisal(unit|Context) \end{aligned} \tag{3}$$

For linguistics, such an approach is promising in shedding light on certain aspects of language variation that are hitherto not sufficiently explained. It naturally extends to all facets of linguistic communication, thus offering a deeper understanding of the relationship between the nature of variation provided by linguistic systems and the way it is exploited in actual language use. For instance, (Hawkins 2014) argues, major patterns of variation across languages are structured by general principles of efficiency in language use and communication.

2 Linguistic Encoding and Cross-lingual Surprisal

Along with the expressiveness needed for communication, human languages provide *a multitude of choices* regarding how information to be transmitted may be encoded.

2.1 Encoding Density

Transmission of information is central to this framework of ideas: It is assumed that *getting across our message* is what motivates much of our communication efforts. A given message M – broadly understood as *contextualized meaning* – can be encoded via information chunks of different nature: words, nodes, features (incl. their distribution and concentration), number of open (i.e. unfilled) dependencies (independently of their length), etc. An encoding E is (informationally) denser to the extent that it uses fewer relevant units and/or less (syntactic) structure to transmit the message M. Operationally, encoding density corresponds to the *amount of relevant units per message*: the higher this amount is, the lower the encoding density. Obviously, with different relevant units, the encoding density of a message $E(M)$ may vary.

2.2 Cross-lingual Information Transmission

The transmission generally involves encoding and decoding mechanisms. An implicit assumption often made in surprisal-based modeling is that the encoding and decoding standards of the producer and receiver match, i.e. that there is no explicit distinction

[4] Surprisal as complexity metrics is the logarithm (with base 2, i.e. counted in bits) of the reciprocal of the probability of an event. For an event x the surprisal is $log_2\left(\frac{1}{P(x)}\right)$, which is referred to as "self-information" of an event (i.e. the information of observing this outcome rather than any of the others that were possible in some predefined universe of events).

between encoding and decoding mechanisms. This naturally leads to a preferential study of the decoding differences between varying encodings of the same or very similar messages in a strictly monolingual setting. The widespread phenomenon of *intercomprehension*, or cross-lingual communication enabled by the mutually intelligibility of the used genetically and typologically related languages or language varieties, substantially differs from this standard scenario. In monolingual human language processing, the predictability of a unit given its surrounding 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 message in a related but unknown language. In an intercomprehension scenario, a speaker of language L_1 encodes an intended message according to language L_1 standards. A speaker of language L_2, who does not know the encoding/decoding standards of L_1, receives the encoded message. Since the receiver does not know the correct L_1 decoding mechanism, he or she applies his or her own language decoding procedure instead. As a result, successful information transmission may be systematically hindered due to a mismatch between L_1 encoding and L_2 decoding mechanisms. For predicting the success of cross-lingual information transmission, our modelling needs to rely not only on surprisal in context but also crucially on linguistic similarities.

When trying to define surprisal for this setting, we can take it to mean either (a) "how unexpected is the (correct) *adaptation* of an unknown-language unit given the current system" or (b) "how informative is a new unknown element given the *accessible, adapted context information so far*". Both interpretations are valid ways of looking at cross-lingual surprisal that differ in their focus: The former version places emphasis on the *translation* aspect of divergent information coding schemes, while the latter variant focuses on the *information conveyed* by a message.

2.3 Transparency and Surprisal

In order to distinguish between these concurrent views on cross-lingual surprisal, we introduce the notion of *transparency*. Intuitively, by transparency we want to capture whether the information contained in a message M encoded according to some standards E is accessible to a decoder D which follows potentially different standards. We assume that all encodings producible within one language L are transparent to all native speakers of this language L. We denote the set of encodings of a given message M producible in language L by $E(M)$.

Apparently transparent encodings (AT). We define an encoding $E(M)$ of message M to be *apparently transparent* to a decoding participant D if D succeeds in decoding $E(M)$ to some M', i.e. $D\big(E(M)\big) = M'$. For *apparent transparency*, the decoded message M' does not have to match the encoded message M – the decoding must have simply been successful.

Fully transparent encodings (FT). We define an encoding $E(M)$ to be *fully transparent, positively transparent,* or simply *transparent* to a decoding participant D if D succeeds in decoding $E(M)$ to M, i.e. $D(E(M)) = M$.

Partially transparent encodings (PT). An encoding $E(M)$ is *partially transparent* to decoder D if $D(E(M)) = M'$ where M' is contained in M, but some aspects of M are missing in M', i.e. $M' \subset M$.

Deceptive encodings (DT). An encoding $E(M)$ is *deceptively transparent, negatively transparent*, or simply *deceptive* to a decoding participant D if D succeeds in decoding $E(M)$ to some M', but M' is not contained in M, i.e. $M' \not\subset M$.

Natural transparency between languages (NT). An encoding E used by language L_1 is *naturally transparent* to speakers of language L_2 if the decoding mechanism of L_{21} correctly handles E without additional knowledge of L_1, i.e. when E is fully transparent to L_1-agnostic speakers of L_2.

The concept of transparency enables us to speak concisely of adaptation vs. information aspects. The above formulations require a rigorous definition of messages, and thus information itself, in order to be used directly. Such definitions are, potentially, not always feasible. Ultimately, we are interested in the *natural transparency* (cf. NT above) between various closely related (here: Slavic) languages. Besides, the degree of natural transparency between two language L_1 and L_2 can serve as a measure of linguistic similarity of their lexicons and grammars. Two further factors are critical for successful intercomprehension, namely, the awareness of interference phenomena ("false friends", cf. DT above) and the accessibility of pre-knowledge enabling inference (expectation; overgeneralization).

If an encoding $E(M)$ produced by transmitter S is fully transparent to a decoder D, then we expect $D(E(M))$ to have the same surprisal for D in D's context as $E(M)$ does for S in S's context:

$$Surprisal\big(D\big(E(M)\big)|Context_D\big) = Surprisal(E(M)|Context_S) \quad (4)$$

3 Similarities and Asymmetric Intercomprehension

Linguistic phenomena may be unique to a language, shared between two languages, or common to many languages from a family. As Townsend and Janda (1996:25) point out, '*[m]ost Slavs speak of understanding each other without much difficulty, but this is usually exaggerated and applies mostly to a simple concrete level*.' Ringbom (2007:11), distinguishes between *objective* (established as symmetrical) and *perceived* (not necessarily symmetrical) cross-linguistic similarities. Various constellations are indeed possible, e.g., speakers of language A may understand language B better than language C, i.e. [A(B)>A(C)], while speakers of language B may understand language C better than language A, i.e. [B(C)>B(A)], etc. Asymmetric intelligibility can be of linguistic nature, e.g., if language A has more complicated rules and/or irregular developments than language B, this results in structural asymmetry (Berruto 2004). As a matter of fact, transparency of vocabulary along with phonetic, morphological and syntactic structures, is typically asymmetric across languages. Asymmetric intelligibility can also be due to extra-linguistic and socio-cognitive factors like attitude, language exposure, age, level of education, or 'unequal' language status when speakers of a 'smaller' (or less prestigious) language usually understand the 'larger' (or more prestigious) one better than vice versa (Vanhove 2014).

All this suggests that in order to model intercomprehension we need methods that capture observed and expected asymmetries, account for the information conveyed by a linguistic unit and accordingly scale the cognitive effort required to process this information.

3.1 Testing Human Performance

In (Stenger, Jágrová & Avgustinova 2020) we report on a web-based resource for conducting intercomprehension experiments with native speakers of Slavic languages and present our methods for measuring linguistic distances and asymmetries in receptive multilingualism. Through a website, which serves as a platform for online testing, a large number of participants with different linguistic backgrounds can be targeted.[5] A statistical language model is used to measure information density and to gauge how language users master various degrees of (un)intelligibilty. The key idea is that intercomprehension should be better when the model adapted for understanding the unknown language exhibits relatively low average distance and surprisal. All obtained intelligibility scores, together with distance and asymmetry measures for the different language pairs and processing directions, are available as an integrated online resource in the form of a Slavic intercomprehension matrix (SlavMatrix)[6], which is further maintained and completed as new data and correlations become available. Intercomprehension scores, obtained from the respondents, actually reveal what is known as *inherent* intelligibility based on structural linguistic similarities (Gooskens 2019).[7]

The challenge on **word-level intelligibility** is designed as a cognate guessing task. The participants are exposed to randomized stimuli in two conditions: **written** (visual perception) or **spoken** (auditory perception), with the task to write within 10 seconds a translation of each word in their native Slavic language. The allocated time is supposed to be sufficient for typing even the longest words, but not long enough for using a dictionary or an online translation aid. The results are automatically categorized as 'correct' or 'wrong' via pattern matching to predefined correct answers and acceptable alternatives, with an integrated tolerance for lower/upper case and diacritical signs. The respondents get an immediate feedback in the shape of an emoticon– a "thumb up" for a successful translation or a "sad face" for a wrong or missing translation. In the final analysis, the responses are checked manually typographical errors.

The challenge on **phrase-level intelligibility** is designed as a translation of noun and adjective sequences, with the adjective occurring pre- or post-nominally. For each stimulus phrase, the participants have 20 seconds for entering a translation into their native

---

5 All experiments are available at http://intercomprehension.coli.uni-saarland.de with an interface in 11 Slavic languages, English and German.

6 As of March 2020, about 2000 native speakers participated in the challenges.

7 The website provides an additional try-again functionality for already completed experiments. Learners of Slavic languages can thus repeat completed tasks and compare their initial results (corresponding to *inherent* intelligibility) with the intercomprehension scores achieved after a focused teaching intervention (revealing the so-called *acquired* intelligibility).

Slavic language. The individual target words, together with the words directly preceding them, are also tested in a base form (if applicable) in the word-intelligibility challenge.

The challenge on **sentence-level intelligibility** is designed as a cloze (fill-in-the-gap) task. The respondents see initially only the first word of the sentence in the unknown language. They are prompted to click on the word so that the next word in the sentence appears. After they have clicked through and consequently read the entire stimulus sentence in that way, a box appears at the position of the last word, which should be translated into their native Slavic language. This method ensures that participants read each sentence word by word. There are two separate time limits: one for clicking and reading through the sentence and one for entering the translation of the target word. The latter is automatically set by the system to 20-30 seconds, depending on the length of the sentence. The time limit for clicking and reading through the whole sentence is set to a maximum of 300 seconds.

3.2 Example: Polish through Czech

Studying the role of a predictive context in intercomprehension, (Jágrová and Avgustinova, 2019) found that surprisal significantly correlates with target words that are non-cognates or false friends, and could show that in solving the task, the respondents relied on context rather than on word similarity. Altogether 149 Polish target words were tested on Czech participants both in highly predictive sentential contexts, with cloze probability ≥90% according to (Block and Baldwin 2010), and without context, in a word-intelligibility challenge. The stimuli contained 65.1% cognates, 11.4% non-cognates, and 23.5% false friends for the selected language pair.

Hypothesis: Successful disambiguation of target words in a closely related foreign language relies on both cross-lingual similarity (measured by linguistic distance, LD) and predictability in sentential context (in terms of surprisal obtained from 3-gram LMs). In a monolingual setup, the more predictable a word is in context, the lower is the cognitive effort to process the information provided by the word – this corresponds to a low surprisal value. On the contrary, words that are unpredictable in context, and thus cause greater cognitive effort, have higher surprisal values. In the current multilingual setup, target words that have a low LD to the respondent's native language and are predictable in the context are expected to be translated correctly more often than words that are less similar and unpredictable. Consequently, the correct answers per target word should correlate better with both LD and surprisal rather than only with LD. Obviously, the amount of a correctly perceived sentential context plays a crucial role here, because if the context is not intelligible enough, then its supportive power in terms of predictability might lose its effect. Moreover, with a context that is helpful enough, it should be possible to understand even non-cognates and maybe even detect false friends in the stimulus sentences. However, the effects of a semantic priming are hardly predictable by the 3-gram LMs applied here. **Research questions:** Are Polish target words more comprehensible for Czech respondents when they are presented in a highly predictive sentential context? If so, do surprisal values obtained from 3-gram LMs correlate with intelligibility scores obtained in intercomprehension experiments?

Online experiments. Czech speaking respondents with no previous knowledge of Polish had to translate the target words (*pierścionek→prstýnek* 'ring' and *siłownia→posilovna* 'gym') in two conditions: **with** predictive context (Fig.1) and freely **without** any context (Fig.2).

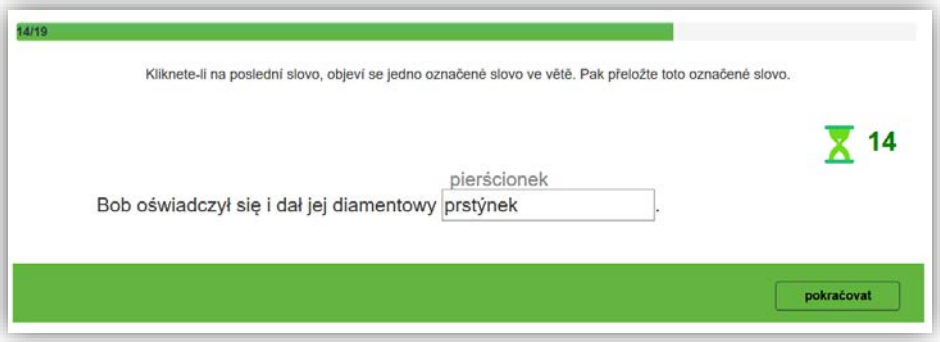

Fig. 2. Experimental screen in cloze translation experiments as seen by Czech respondents. The instruction on top says 'When you click on the last word, a marked word will appear. Then translate this marked word.'

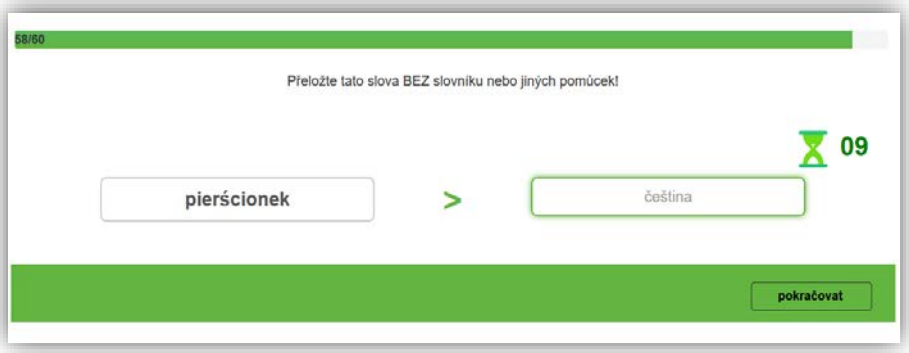

Fig. 3. Experimental screen in free translation experiments as seen by Czech respondents. The instruction on top says 'Translate these words without a dictionary or other aids.'

Modelling. Surprisal is an information-theoretic measure of unpredictability. Statistical LMs inform us about the probability that a certain word follows a certain other word. Thus, surprisal reflects frequency and predictability effects in the corpus on which the LM was trained.[8] Whenever there is a drop in surprisal after a word, the word

[8] The Polish stimuli sentences were scored by an LM trained on the Polish part of InterCorp (Čermák and Rosen, 2012) and the Czech literal translations were scored by an LM trained on the Czech National Corpus (Jagrova et al. 2017).

with the lower surprisal is interpreted as highly predictable after its preceding word. The LD is obtained from the orthographic distance (calculated as the Czechoslovak to Polish pronunciation-based LD, i.e. always towards the closest Czech or Slovak translation equivalent under the assumption that the Czech readers have receptive skills in Slovak) and the lexical distance (determined by the number of non-cognates per sentence in the language pair). Consider the surprisal graphs produced on the bases of corpora by a 3-gram LM for the Polish sentences in (Fig.4) and (Fig.5). Whenever there is a drop in surprisal after a word, the word with the lower surprisal is interpreted as highly predictable after its preceding word.

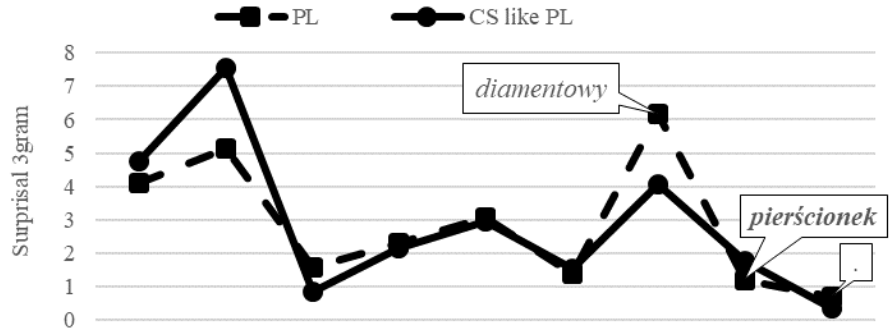

Fig. 4. 'Bob proposed and gave her a diamond ring' (Block and Baldwin 2010)

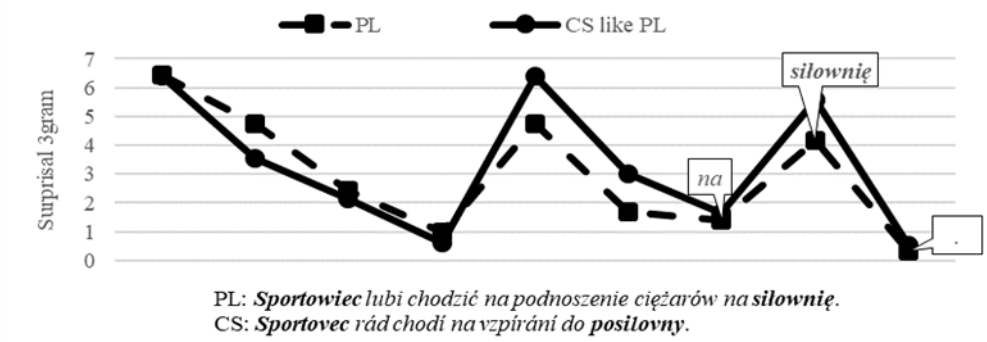

Fig. 5. 'The sportsman likes to do weightlifting at the gym.' The Polish translator was instructed to keep the target word at the last position in the sentences. Therefore, some translations might vary slightly from their original English versions. Original version as of (Block and Baldwin 2010): 'The athlete is enjoying lifting weights at the gym.'

Human performance: For the example in Fig.4, 90% of the respondents translated the Polish target *pierścionek* ‘ring’ correctly, while without context only 45.5% gave the correct Czech cognate *prstýnek*. The dropping surprisal curve after *diamentowy* confirms that the target *pierścionek* ‘ring’ is highly predictable after *diamentowy* ‘diamond [adjective]’. In other words, the LM prediction matches the actual human performance. However, not all successful responses in the study were context-driven. In Fig.5 there is an increase in surprisal at the target *siłownię* ‘gym[acc]’, even though it has a cloze probability of 95% (Block and Baldwin 2010) implying high predictability in context. In other words, the LM prediction deviates from the actual human performance. The experimentally obtained higher rate of correct translations in context (58.1% vs. 30.3% without context) can be explained here by the thematic association of the target word *siłownię* ‘gym[acc]’ with the sentence-initial *sportowiec* ‘athlete / sportsman’, rather than by the predictive power of the two immediately preceding words (*ciężarów na* ‘weights[gen.pl.] at’). Still, a semantic prime can only play a role in intercomprehension if it is correctly recognized as such. In such a case, the respondent may expect for the target position a filler fitting the prime, even though the actual stimulus found there is unfamiliar or unidentifiable.

To sum up, a target word in a stimulus language (here: Polish) can be predicted not only by its collocates in the immediate context, but also due to a semantic prime leading the respondent (a native speaker of Czech) to a correct interpretation and a response that might be associated with the correct translation.

4 Outlook

As different Slavic languages can be mutually more recognizable in pronunciation than in writing, the next major topic on our research agenda is the cross-lingual information transmission in oral intercomprehension scenarios.

References

M. Aylett, A. Turk (2004) The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. Language and speech, 47(1), 31-56.

G. Berruto (2004) Sprachvarietät – Sprache (Gesamtsprache, historische Sprache). In U. Ammon et al. (Eds.), Soziolinguistik. Ein internationales Handbuch zur Wissenschaft von Sprache und Gesellschaft, 1. Teilband. Walter de Gruyter, Berlin, New York, pp. 188–195.

C. K. Block, C. L. Baldwin (2010) Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. Behavior Research Methods 42(3): 665–670.

M. Crocker, V. Demberg, E. Teich (2016) Information Density and Linguistic Encoding (IDeaL). KI – Künstliche Intelligenz 30 (1): 77-81

F. Čermák, A. Rosen (2012) The case of InterCorp, a multilingual parallel corpus. International Journal of Corpus Linguistics 13(3), 411–427

P. Doyé (2005) Intercomprehension. Guide for the development of language education policies in Europe: from linguistic diversity to plurilingual education. Reference study, Strasbourg, DG IV, Council of Europe.

A. Fischer, K. Jágrová, I. Stenger, T. Avgustinova, D. Klakow, R. Marti. (2015). An Orthography Transformation Experiment with Czech-Polish and Bulgarian-Russian Parallel Word Sets. In: B.Sharp, W.Lubaszewski, R.Delmonte (eds.) Natural Language Processing and Cognitive Science 2015 Proceedings. Ca Foscarina Editrice, Venezia.

A. Fischer, K. Jágrová, I. Stenger, T. Avgustinova, D. Klakow, R. Marti. (2016). Orthographic and Morphological Correspondences between Related Slavic Languages as a Base for Modeling of Mutual Intelligibility. In: N. Calzolari, et al.(eds.) Language Resources and Evaluation Conference LREC 2016, pp. 4202-4209, included linguistic resources, Portorož (Slovenia)

D. Genzel, E. Charniak (2002). Entropy rate constancy in text. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 199-206).

C. Gooskens (2019) Receptive multilingualism. Multidisciplinary perspectives on multilingualism: The fundamentals LCB 19: 149–174.

J. Hale (2001). A probabilistic Earley parser as a psycholinguistic model. In Second Meeting of the North American Chapter of the Association for Computational Linguistics.

J. A. Hawkins (2004). Efficiency and complexity in grammars. Oxford University Press on Demand.

J. A. Hawkins (2014) Cross-linguistic variation and efficiency. OUP Oxford.

K. Jágrová, T. Avgustinova (2019) Itelligibility of highly predictable Polish target words in sentences presented to Czech readers. CICLing 2019. Springer's Lecture Notes in Computer Science (preprint, appendix)

K. Jágrová, T. Avgustinova, I. Stenger, A. Fischer. (2018) Language Models, Surprisal and Fantasy in Slavic Intercomprehension. In R. K. Moore, P. Fung, S. Narayanan (eds.), Computer Speech and Language. Elsevier

K. Jágrová, I. Stenger, R. Marti, T. Avgustinova (2017) Lexical and orthographic distances between Bulgarian, Czech, Polish, and Russian: A comparative analysis of the most frequent nouns. In J.Emonds, M.Janebová (eds.), Language Use and Linguistic Structure. Proceedings of the Olomouc Linguistics Colloquium 2016/, 401–416. Olomouc: Palacký University.

R. Levy (2008). Expectation-Based Syntactic Comprehension. Cognition 106(3): 1126–1177.

T. F. Jaeger, R. P. Levy (2007) Speakers optimize information density through syntactic reduction. In Advances in neural information processing systems (pp. 849-856).

M. Mosbach, I. Stenger, T. Avgustinova, D. Klakow. (2019) incom.py - A Toolbox for Calculating Linguistic Distances and Asymmetries between Related Languages. In: G.Angelova, R.Mitkov, I.Nikolova, I.Temnikova (eds.), Proceedings of Recent Advances in Natural Language Processing, RANLP 2019, Varna, Bulgaria, 2-4 September 2019, pages 811-819

H. Ringbom (2007) Cross-linguistic similarity in foreign language learning. Multilingual Matters LTD, Clevedon.

C .E. Shannon, (1948) A mathematical theory of communication. Bell System Technical Journal 27: (379–423), 623–656.

M. Staudte, A. Heloir, M. Crocker, M. Kipp (2011) On the importance of gaze and speech alignment for efficient communication. In Proceedings of the 9th international gesture workshop.

I. Stenger, T. Avgustinova (2020a) How intelligible is spoken Bulgarian for Russian native speakers in an intercomprehension scenario? In: V.Micheva, D.Blagoeva, M.Vitanova, M.Tsibranska-Kostova, S.Kolkovska, T.Aleksandrova (Eds.) Proceedings of the International Annual Conference of the Institute for Bulgarian Language (Sofia 2020), Volume II: 142-151

I. Stenger, T. Avgustinova (2020b) Visual vs. auditory perception of Bulgarian stimuli by Russian native speakers. In: V.P.Selegej et al. (Eds.), Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference 'Dialogue' (2020). Issue 19, 684–695

I. Stenger, T. Avgustinova, K. Belousov, D. Baranov, E.Erofeeva (2019) Interaction of linguistic and socio-cognitive factors in receptive multilingualism [Vzaimodejstvie lingvističeskich i sociokognitivnych parametrov pri receptivnom mul'tilingvisme], 25th International Conference on Computational Linguistics and Intellectual Technologies (Dialogue 2019), Proceedings, Moscow, Russia: http://www.dialog-21.ru/digest/2019/online/

I. Stenger, T. Avgustinova, R. Marti (2017) Levenshtein distance and word adaptation surprisal as methods of measuring mutual intelligibility in reading comprehension of Slavic languages. Computational Linguistics and Intellectual Technologies: International Conference 'Dialogue 2017' Proceedings. Issue 16(23), vol. 1, 304–317.

I. Stenger, K. Jágrová, T. Avgustinova (2020) The INCOMSLAV Platform: Experimental Website with Integrated Methods for Measuring Linguistic Distances and Asymmetries in Receptive Multilingualism. In J.Fiumara, C.Cieri, M. Liberman, C. Callison-Burch (Eds.), LREC 2020 Workshop Language Resources and Evaluation Conference 11-16 May 2020, Citizen Linguistics in Language Resource Development (CLLRD 2020), Proceedings, pp. 40–48

I. Stenger, K. Jágrová, A. Fischer, T. Avgustinova (2020): "Reading Polish with Czech Eyes" or "How Russian Can a Bulgarian Text Be?": Orthographic Differences as an Experimental Variable in Slavic Intercomprehension. In T.Radeva-Bork, P.Kosta (Eds.), Current developments in Slavic Linguistics. Twenty years after (based on selected papers from FDSL 11). Peter Lang, 483-500 (preprint, link to publication)

I. Stenger, K. Jágrová, A. Fischer, T. Avgustinova, D. Klakow, R. Marti (2017) Modeling the impact of orthographic coding on Czech–Polish and Bulgarian–Russian reading intercomprehension. Nordic Journal of Linguistics 40(2), 175–199.

A. E. Townsend, L. A. Janda (1996) Common and comparative Slavic: phonology and inflection with special attention to Russian, Polish, Czech, Serbo-Croatioan, Bulgarian. Slavica Publ.

J. Vanhove (2014) Receptive multilingualism across the lifespan. Cognitive and linguistic factors in cognate guessing. PhD thesis. University of Fribourg (Switzerland).