Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 53 (2019) 242–275

www.elsevier.com/locate/csl

# Language models, surprisal and fantasy in Slavic intercomprehension ☆

Klára Jágrová[a,b,*], Tania Avgustinova[a,b], Irina Stenger[a,b], Andrea Fischer[a,b]

[a] *Slavic Studies, Saarland University, Saarbrücken, Germany*

[b] *Department of Language Science and Technology, Saarland University, Saarbrücken, Germany*

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

Available online 12 June 2018

**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), avgustinova@coli.uni-saarland.de (T. Avgustinova), 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

role of the sentential context in reading intercomprehension, which we expect to be crucial for successfully decoding the message in the stimulus.

### 2.2. Surprisal as a measure for information density

Psycholinguistic research indicates that the cognitive processing complexity of sentences can be modelled with statistical models. Earlier research (Hale, 2001) explored the use of statistical parsers for this purpose. Also Levy (2008) showed that n-gram models, specifically trigrams, performed well at this task. The employed measure is called surprisal and is defined as:

$$Surprisal(unit|context) = -\log_{10} P(unit|context)$$

For a word, surprisal is the negative log-likelihood of encountering this word in its preceding context. The trigram LMs applied here output surprisal scores in hartley[1] (unit of information). Surprisal is widely used in information-theoretic modelling of human language. Intuitively, it can be thought of as measuring the information content conveyed by a linguistic unit and it appears to scale the cognitive effort required to process this information (Crocker et al., 2015). As an example, consider the following English sentence:

*She went to the shop to buy some apples and _.*

Using our knowledge of the world, we know that *oranges* is a good continuation after *apples and*, while for instance *hexagons* is not. This is reflected well by LMs which would assign a high probability – and hence low surprisal score – to *oranges*, while assigning a low probability – and hence high surprisal score – to the word *hexagons*. If we successively score each word of a sentence given the preceding words, we obtain an information density profile of that sentence. If a word is highly unexpected in its context, it will lead to a peak in information density – a high surprisal score.

Viewed from the decoding perspective, surprisal scores obtained from trigrams correlate very well with e.g. human reading times of texts of various difficulties (Levy, 2008). In reading intercomprehension settings, we view comprehension as a decoding process. In experiments or real-life communications, there is also another perspective – that of encoding. According to the UID (uniform information density) hypothesis (Jaeger, 2010) speakers tend to distribute information uniformly over the duration of an utterance, avoiding peaks and troughs in surprisal. For answers given in sentence translation experiments between related languages, we would intuitively expect that people should prefer those translations of unknown words which are characterized by lower density profiles.

## 3. Hypothesis: processing effort in intercomprehension results from the two orthogonally measurable dimensions, distance and surprisal

According to the aforementioned definition of surprisal, processing difficulty and information content correlate with each other: the higher the surprisal, the higher is the processing difficulty and the higher is the information content of the message. So far, this was proven to be the case in monolingual situations. Regarding processing effort and information density we can conclude that these two only correlate with surprisal in intercomprehension if the code is transparent enough.

In an intercomprehension scenario, however, information content and processing effort of a message are highly dependent on an additional factor: linguistic distance. We expect that this distance, which is also a consequence of the (un-)relatedness of languages, influences the processing of a message even before the context starts to play a role. As soon as the code becomes opaque, there is a loss of information density and an increase in processing effort. In Section 4, we present how readers stick to understandable words in a sentence first and then try to infer the meaning of the remaining sentence that is semantically reasonable to them.

[1] The unit *hartley* is the pendant of the bit; the unit bit uses the binary logarithm to the base 2, while *hartley* uses the common logarithmic base 10.

We make the following assumptions: If a text has low linguistic distance, then transfer of knowledge from a language L1 to an unknown language LX is possible. We can speak of a lexical distance if a text contains non-cognates – words that are not etymologically related to their corresponding translations in the reader's L(s). The recognition of cognates is a prerequisite for successful intercomprehension (cf. Möller and Zeevaert, 2010). However, often etymological correspondences are hardly recognized by the reader because of different spelling or unusual morphological properties. Then we speak of orthographic or morphological distance of cognates respectively, i.e. the difficulty does not lie on the lexical level. If a text is for instance orthographically similar, but lexically distant, this might lead to searching for a way to fill comprehension gaps – usually in the language and grammar repertoire that is available to the reader. In other words, the term *linguistic distance* has either to be further specified by mentioning the level which the distance refers to or to be understood as an overall summary of the distance on all levels (lexis, orthography, morphology, morphosyntax).

Both (correct) inferences and (misleading) interferences from other languages are likely to happen if a text is perceived similar enough. We expect the following interplay of similarity and predictability in intercomprehension: if the encoding of a sentence in LX is very similar or even identical to the reader's L(s), the same processes should apply for the predictability of words given a history as they do in a monolingual situation. If a sentence is linguistically distant, e.g. because of a lack of cognates,[2] processing effort will increase for the readers. As a consequence, readers are expected to fill comprehension gaps with words that make sense to them Fig. 1.

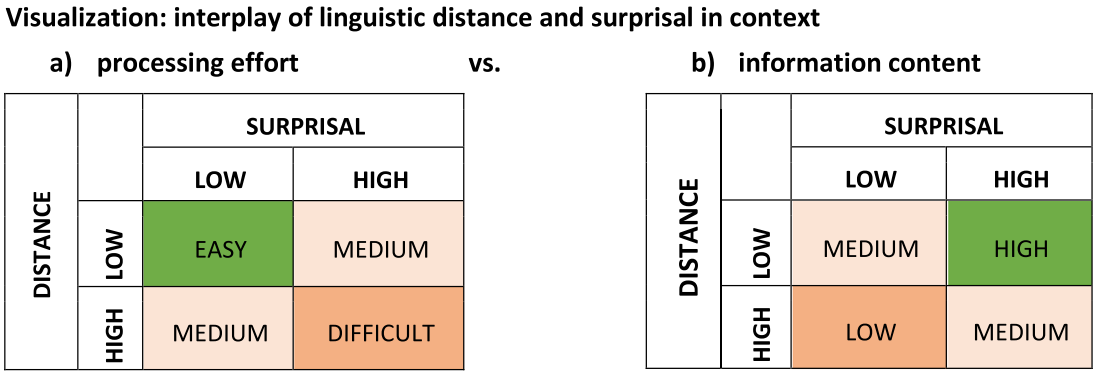



Fig. 1. Expected levels of processing effort (from easy to difficult) and information content resulting from the two separately measurable dimensions linguistic distance and surprisal.

We demonstrate the two distinct dimensions in intercomprehension with an example in this section. In Section 4, we will go further into detail by viewing results from two other experimental settings.

In a small-scale study (cf. Jágrová, 2010), the HR sentence

(i) *Daleko* *je* *kuća* *moja.*
ADV COP n POSS-PRON
*'Far away is the house of mine.'*

was presented to readers of several other Slavic languages. Most of the sentence is expected to be both lexically and orthographically transparent to readers of e.g. Bulgarian (BG), CS, PL, and Russian (RU), with the assumption that Bulgarian and Russian readers are familiar with the Latin script. However, from the morphosyntactic perspective, Russian readers might have difficulties as they do not expect a copula verb here – this is where the sentence is syntactically opaque to Russian readers. RU usually does not use the copula verb in the present tense and indicative mood; forms such as *есть* (est'[3]) 'to be', which would theoretically be the correct translation equivalent for the HR

[2] In this context, we define cognates as historically related words with the same meaning in different languages (cf. for instance Kürschner et al. 2008:86).

[3] Cyrillic is transliterated into Latin script according to the ISO 9:1986 standard throughout this article.

word *je*, are used only if there is an emphasis on the existence of something or somebody (e.g. *У меня есть сестра. U menja est' sestra* 'I have a sister'). Slavic readers expect a noun at the position of *kuća* 'house' because of its feminine ending -a, the subsequent *moja* ('my' [feminine] (possessive pronoun as postmodifier) and the verb, given it is identified as such, preceding the noun. The feminine morphological ending -a of *kuća* is transparent together with its agreement in the possessive pronoun *moja*. The question is: which noun do the Slavic readers expect here and why? And how do Russian readers interpret the copula verb?

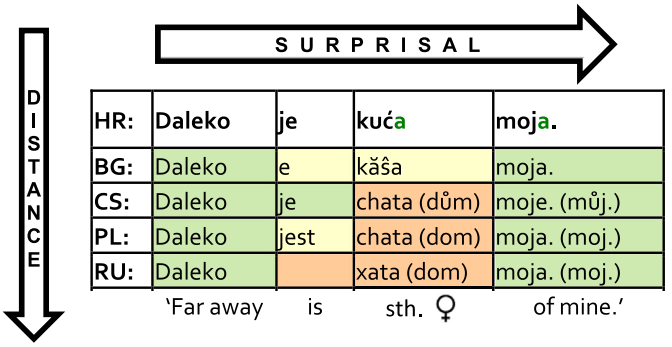



Fig. 2. Visualization of the separately measurable dimensions in intercomprehension: linguistic distance and surprisal.

There are feminine words representing house-like concepts such as *chata* 'cottage' in other Slavic languages. However, their initial letter is not *k*, which might play a crucial role. The following translations were given by the RU respondents ($n = 7$) in this experiment:

(ii) *Далеко же куча моя. (Daleko že kuča moja.)* 'Oh how far away is the stack of mine.'
(iii) *очень далеко (očen' daleko)* 'very far'
(iv) *Далеко есть семья моя. (Daleko jesť sem'ja moja.)* 'Far away is the family of mine.'
(v) *Далеко же киса моя. (Daleko že kisa moja.)* 'Oh how far away is the pussycat of mine.'
(vi) *Далеко ты любимая моя. (Daleko ty ljubimaja moja.)* 'You are far away, love [female] of mine.'
(vii) *Far away is the small stack.* [Answer by an English-Russian bilingual person]
(viii) *Sehr weit entfernt ist meine* . . . 'Very far away is my [feminine] . . .' [Answer by a German-Russian bilingual person]

Why did the Russian respondents choose specifically these translations here for HR *kuća* 'house'? In an attempt to find an explanation for this, we trained a trigram LM with Kneser–Ney smoothing (Kneser and Ney, 1995) on a corpus of RU[4] – the method is further described in Section 4.1. We scored the answers that were given by the participants. The surprisal values of the different translations given (see legend above Figs. 3–8) and possible other translations are visualized in Figs. 3–8. All translation variants of *kuća* are scored in different syntactic frameworks[5] which are indicated by the English translations in the descriptions underneath each diagram. Linearization in the clausal domain in Slavic is syntactically free, i.e. it depends mainly on information structure in terms of topic-focus. The placeholder N in the diagrams stands for the position of the different nouns that *kuća* was translated into. The different RU nouns are given in Latin transliteration and are represented by the different colours in the legend. The data labels in the diagrams are transliterated accordingly. The higher the surprisal score, the more unlikely the word is expected to be in the readers' language.

[4] The RU part within the parallel part of the Russian National Corpus combined with the RU part within the SCD InterCorp of the Czech National Corpus. Corpus size and details are given in Section 4.1.

[5] The HR stimulus sentence has non-standard word order in both HR and RU (cf. surprisal of the RU translation in Fig. 3) where the copula verb is translated correctly. In order to include the role of the divergent word order into the analysis, the translated sentence is scored in all possible RU word order variants in Figs. 3–8.

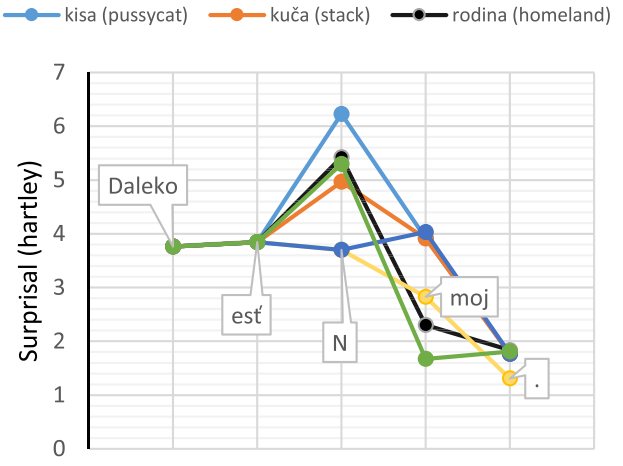



Fig. 3. ‘Far away is N of mine.’ (*je* → copula).

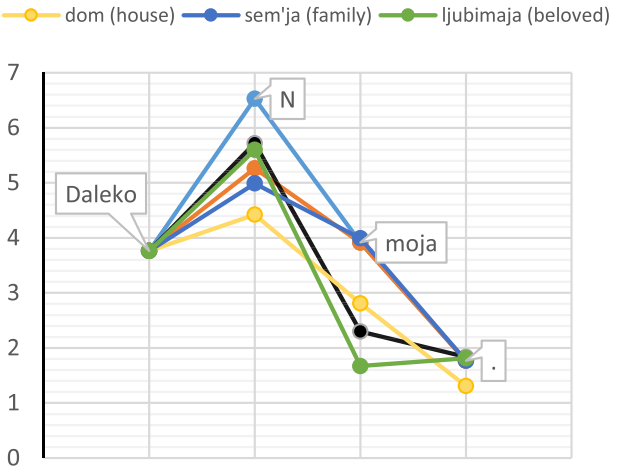



Fig. 4. ‘Far away [is] N of mine.’ [*je* → zero copula].

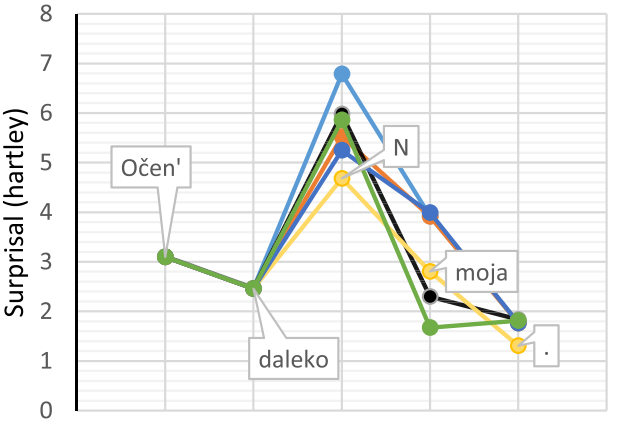



Fig. 5. ‘Very far away [is] N of mine.’ (*je* → *očen’* ‘very’).

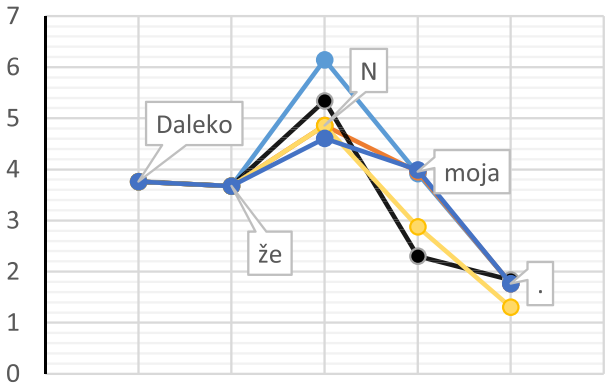



Fig. 6. ‘Oh how far away [is] N of mine.’ (*je* → *že*).

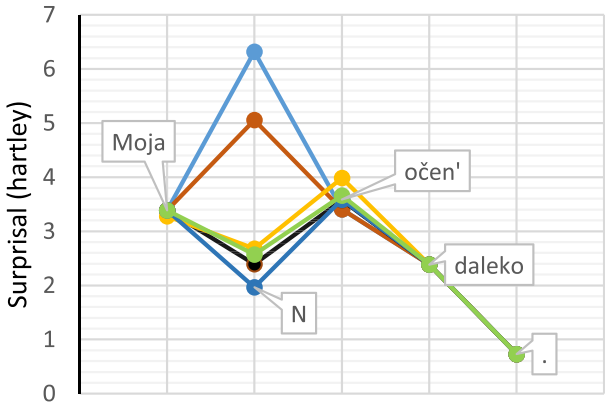



Fig. 7. ‘My N [is] very far away.’ (*je* → *očen’* ‘very’).

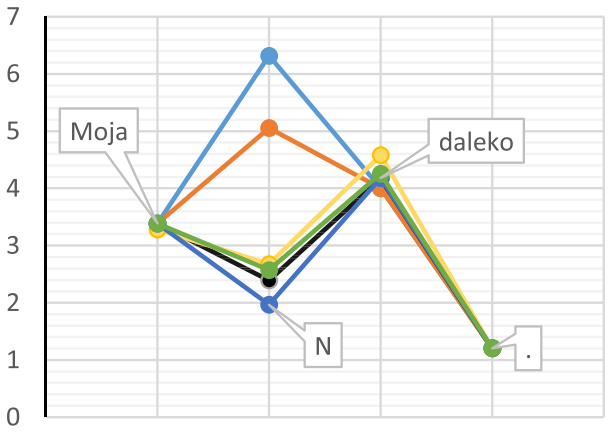



Fig. 8. ‘My N [is] far away.’ (*je* → zero copula).

The highest surprisal value of all the possible translations given for *kuća* in all versions of the sentence with different word order has *киса* (*kisa*) ‘pussycat’. However, this did not prevent the respondent (answer v)) from opting for this word. It is feminine and starts with the letter *k*. In comparison to that, the translation *куча* (*kuča*) ‘heep, stack’ in (ii) and (vii) is slightly less surprising than *киса* (*kisa*) ‘pussycat’ (scores ranging from 6.23 in Fig. 6 to 6.79 hartley in Fig. 5). The translation *семья* (*sem’ja*) ‘family’ given in (iv) is feminine and has the lowest surprisal scores of

all feminine translation options given. The correct translation *дом* (*dom*) 'house' would have a much lower surprisal value than *куча* (*kuča*) 'heep, stack' or *киса* (*kisa*) 'pussycat', and nevertheless, none of the respondents answered 'house'. The reason for this might be that *дом* (*dom*) 'house' is not feminine and does not start with a *k*. Also, none of the 7 respondents translated the unknown lexeme with *родина* (*rodina*) 'homeland' – it is feminine and the context fits well, but again it does not start with a *k*. Therefore, we can assume that the initial letter seems to play a crucial role (cf. Vanhove, 2014) and it overpowers other translations that would fit better into the context.

Likewise, the translations of the Czech readers responding to the same stimulus reveal amusing interpretations:

(ix) *Daleko je domov můj.*[6] 'Far away is the home of mine.' [$n = 4$]
(x) *Jak daleko je můj dům.* 'How far away is my house.'
(xi) *Daleko je láska moje.* 'Far away is the love of mine.'
(xii) *Daleko je má chalupa.* 'Far away is my holiday house.'
(xiii) *Daleko je děvče moje.* 'Far away is the girl of mine.'
(xiv) *Kuča je domov?* '*Kuča* [sic!] means home?'
(xv) *Daleko je vesnice moje.* 'Far away is the village of mine.'
(xvi) *Daleko je chata má.* 'Far away is my cottage.'
(xvii) *Daleko je holka moje*. 'Far away is the girl of mine.'
(xviii) *Daleko je vlast moje.* 'Far away is the homeland of mine.'

And the German respondents with knowledge of at least one Slavic language translated:

(xix) *Weit in meine Küche?* 'Far into my kitchen?'
(xx) *Weit ist meine . . .* 'Far away is my [feminine] . . .'
(xi) *Weit weg ist meine Kutsche.* 'Far away is my carriage.'
(xii) *Weit ist meine Kutsche.* 'Far away is my carriage.'
(xiii) *Weit entfernt ist meine Kundin.* 'Far away is my customer [feminine].'

The answers of the respondents from all three language backgrounds reveal some common features: for *kuća*, all respondents prefer translations of feminine nouns, animate or unanimate, and especially those that have the initial letter *k*, resp. *K*. In reading intercomprehension, readers try to infer the meaning of non-transparent words from the context of the recognized cognates and apparently also from features of the unknown word. This context can be not only semantic, but also syntactic. The latter is the case, for instance, when readers recognize a non-transparent word as a noun. In sentence (i), the context can be assumed to be completely transparent to readers of other Slavic languages. The adverb *daleko* 'far away' and the possessive pronoun *moja* 'my' are fully intelligible – all words are cognates with no or low orthographic distance. The only difficulty encountered here is of lexical nature in *kuća* 'house'. And it can be assumed that the reason why respondents have translated it as *киса* (*kisa*) 'pussycat' are the interferences on the orthographic level, respectively an opacity of grapheme-to-phoneme correspondences. Hence, the HR *kuća* must have been decoded by a Russian reader as if it was written in Cyrillic as follows: the HR grapheme ⟨k⟩ as the RU grapheme ⟨*к*⟩ to the RU phoneme /k/, HR ⟨u⟩ as RU ⟨*и*⟩ to RU /i/, HR ⟨ć⟩ as RU ⟨c⟩ (ignoring the diacritic) to RU /s/, HR ⟨a⟩ as RU ⟨a⟩ to RU /a/: *киса* /'kʲisa/.

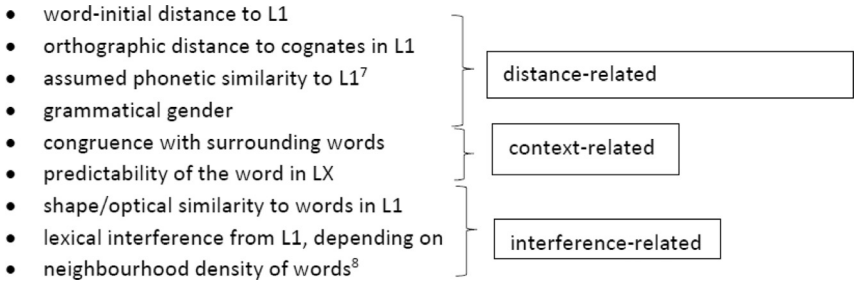
At the same time, there apparently is a script opacity effect in *je* 'is' and *kuća* 'house'. In *je*, the orthographic opacity effect seems to be combined with a lexical interference effect. An interesting observation can be made in the answers of the RU respondents: two of seven translated *je* as *же (že)*, which is probably due to the null form of the copula *be* in RU. In addition to that, readers most probably try to pronounce what they are reading – a phenomenon called *inner speech* in psycholinguistics (Harley, 2007), which results in a realization of the letter *j* as /ʒ/ as it would be pronounced e.g. in French *je* 'I' which again sounds similar to RU *же* (*že*) 'oh'. In those cases in which *kuća* was translated as *куча* (*kuča*) 'heep, stack', the stimulus *kuća* is transparent with regard to orthography of the Latin script. In these cases, the Russian respondents probably tried to pronounce the stimulus, including the transfer processes: HR ⟨k⟩ as HR /k/ to RU ⟨*к*⟩, HR ⟨u⟩ as HR /u/ to RU ⟨*у*⟩, HR ⟨ć⟩ as HR /tɕ/ to RU ⟨ч⟩, HR ⟨a⟩ as HR /a/ to

[6] The phrase *domov můj* is part of the title and the chorus (*Kde domov můj?* 'Where is my home?') of the Czech national anthem and might therefore cause a certain bias compared to the same phrase in other languages.

RU ⟨a⟩, resulting in RU *кучa* (*kuča*). At this point, the RU lexicon seems to interfere: the existence of the word *кучa* (*kuča*) offers the possibility to interpret *kuća* as *кучa* (*kuča*) 'heep, stack'.

Respondents of all language backgrounds identify *kuća* successfully as a feminine noun together with the context of the remaining sentence: *something feminine of mine is far away*. The position of the noun *kuća* is predictable for BG, CS, PL, and RU readers, although *moja kuća* '*my house*' would be a more common formulation and the inversion in the example is of a rather poetic style.

We can conclude that besides sentential context, there are a number of other linguistic factors influencing reading intercomprehension, e.g. (here: L1 represents the reader's native language, or any other language that is dominant during the decoding of the stimulus):




In the decoding process, there appears to be a trade-off between what makes sense and what is (or at least for the reader seems to be) similar to the unknown word. In this contribution, we are viewing the context- and distance-related factors. We do not go into detail about the interference-related factors.

## 4. Estimating stimulus difficulty from linguistic distance and surprisal

In this section, we are using the approach explained in Sections 2.1 and 2.2 to analyse the answers given by Czech respondents in a sentence translation experiment. The experimental setup was the following: Respondents took part in the experiment in pairs. 12 PL sentences were presented to all respondent pairs ($n = 16$) over a computer screen, one for each respondent separately. The computers were placed in different rooms and the respondents communicated over skype (using headsets). After filling in a questionnaire in which they were asked to provide information about their knowledge of and exposure to foreign languages, they were confronted with the task to cooperatively translate 12 PL sentences into CS. They were asked to communicate to their partner every thought about the possible meaning of certain words that they were not sure of. Only one of the respondents was able to enter the translation into the response field at a time, while the partner could see what the writing respondent was typing. The respondents changed turns typing. The whole experiment was set up in a modified design of think-aloud protocols (cf. Ericsson and Simon, 1993). The aim was to record what the respondents were actually thinking when solving the translation task. In pairs, the respondents communicate more openly when solving a task together than when a single respondent was asked to *think aloud* during translation.

The outcomes of the experiments were two kinds: (i) the written translations entered in the solution field and (ii) the audio recordings of the respondents' conversations during the translation task. The audio recordings provide large amount of data about different aspects of the respondents' translation processes and provide explanations as of why they came up with certain solutions.

We can expect that the results obtained from the respondent pairs are somewhat better (more correct answers) than data that would have been obtained from single individuals. Apart from that, there are cases where it is not

[7] as in example (ii) grapheme-phoneme-grapheme transfer, e.g. *kuća* read as /kutʃa/ and associated with *кучa (kuča)* 'heep'
[8] availability of words that have only 1 different letter at any position

trivial to determine if a translation given is correct or not, especially in cases where paraphrases are possible. In the following section, we look at three of the twelve stimuli sentences presented to the respondents,[9] focussing on those situations in which the respondents mentioned that what they are understanding either does or does not make sense. Furthermore, we are looking into what a CS trigram LM can reveal about the contexts, and whether the surprisal values from this CS LM agree with the respondents mentioning doubtful or reasonable context. Hence, this section is about the *decoding* process involved in the translation task.

PL stimuli sentences presented to the Czech respondents (gaps in the cloze translation experiments discussed in section XX are underlined) with their correct CS and EN translations:

(xiv) PL: *Nie widziałam, że jego żona pokazuje ręką, żebyśmy poszli do rektora.*
CS: 'Neviděla jsem, že jeho žena ukazuje rukou, abychom šli k rektorovi.'
EN: 'I did not see that his wife is showing with her hand that we should go to the rector.'
(xv) PL: *Gdyby nie było książek, czytałbym Ci z oczu.*
CS: 'Kdyby nebylo knížek, četl bych Ti z očí.'
EN: 'If there were no books, I would read from your eyes.'
(xvi) PL: *Kupiliśmy nie tylko czerstwy chleb, ale jeszcze gorzej – też stary żółty samochód.*
CS: 'Koupili jsme nejen tvrdý chléb, ale ještě hůř – také staré žluté auto. '
EN: 'Not only did we buy stale bread, but even worse, also an old yellow car. '

### 4.1. Scoring surprisal of stimuli in the translation experiments

In order to determine the surprisal of certain words in context and thus to predict their processing difficulty in the contextual dimension, we trained statistical trigram models with Kneser–Ney smoothing (Kneser and Ney, 1995) on corpora, one for each language under focus. The Kneser–Ney smoothing technique leverages available information from overlapping, smaller n-grams to ensure that surprisal scores computed for unseen word combinations do not turn out extremely high. The training corpora are merged subcorpora of InterCorp (Čermák et al., 2012) and the Russian National Corpus (V. V. Vinogradov Russian Language Institute, 2015). With the PL model, we are able to determine the information density and estimate the processing difficulty of the respective stimuli for a monolingual Polish reader. We train the same type of LM also on a CS corpus which should serve as a representation of a native Czech reader. Then we score the closest CS translation with the help of the CS LM. This provides insight about the processing difficulty of the CS translations for a Czech reader and serves the purpose of evaluating respondents' answers to the stimuli.

Table 1
Overview of the training material for the Kneser–Ney trigram models.

| Language | Corpus | Size (k tokens) |
|---|---|---|
| CS | CS part of the InterCorp merged with the CS part of the parallel part of the Russian National Corpus | 175,190 |
| PL | PL part of the InterCorp merged with the PL part of the parallel part of the Russian National Corpus | 104,713 |
| RU | RU part of the InterCorp merged with the RU part of the parallel part of the Russian National Corpus[a] | 12,860 |

[a] The LM trained on the RU corpus was applied on the sentences in Figs. 2–8.

At this point, surprisal does not inform us about lexical, orthographic or morphological difficulties in this cross-lingual reading situation – this can be done by calculating the linguistic distances on the respective levels of lexis and orthography. The surprisal scores of the PL stimuli sentences are displayed in Figs. 9–11, always in a parallel manner for both languages. Endings such as *-m* in *widziałam* 'I saw' [feminine] (explanation see beneath Table 4) are separated from the suffix in the PL corpus by standard and therefore have to be scored separately. In all three, Figs. 9–11, the translated sentences do not reveal any huge differences in surprisal between the languages, meaning that the predictability of the words in context in both of the languages should be comparable.

[9] In the think-aloud setting the respondents were also presented with modified versions of the PL sentences in which certain PL units were replaced by CS ones. These modified sentences are not subject of the underlying analysis, but will be discussed in a future contribution.

### 4.2. Determining linguistic distance of the stimuli as a measure for similarity

In a first step, we look at the lexical distance of the PL stimuli to their corresponding CS translations. If a PL stimulus word can be translated correctly with a CS cognate, we assign a lexical distance value of 0. If there is no correct cognate translation, a distance value of 1 is assigned. If the PL stimulus word is a false friend in CS, we assign the highest value for lexical distance: 2. The three lexical distance levels are visualized according to their difficulty with green for 0, beige for 1 and red for 2 (Tables 4–6).

In a second step, we calculate the orthographic distance of the cognates (those having lexical distance of 0) to their CS counterparts. The underlying calculation method is the Levenshtein algorithm (cf. Levenshtein, 1966) which aligns consonant and vowel letters of cognates in slots. For every deletion, insertion, or substitution of a letter, a cost of 1 is assigned. For letters that differ only in diacritics, a cost of 0.5 is assigned. If there is more than one possible alignment, the cheapest alignment is chosen. The costs per word pair are summed up and divided by the number of alignment slots, which results in a normalised percentage value for the orthographic distance of two cognates.

The three PL sentences under focus have a lexical distance of 12% and an orthographic distance of 38% towards their closest CS equivalents. Table 3 gives an overview of PL–CS distances measures in previous research: Heeringa et al. (2013) measured the lexical and orthographic distances between the translations of the 100 most frequent nouns of the British National Corpus. In a study which used the same method, but analysed the 100 most frequent nouns extracted from PL and CS corpus-based frequency lists, Jágrová et al. (2016) found a lexical distance of 15% of PL for Czech readers, respectively, 10% for CS for Polish readers and an orthographic distance of 36% of PL for Czech readers, respectively 34% for CS for Polish readers. Golubović (2016) also measured the morphological distance of the Slavic languages spoken in the EU with a result of 31.4% for PL-CS texts.

Table 2
Example for determining orthographic distance of cognates by means of the Levenshtein algorithm.

| # alignment slots | 1 | 2 | 3 | 4 | 5 | 6 | Levenshtein distance to CS |
|---|---|---|---|---|---|---|---|
| PL stimulus word | p | o | s | z | l | i | |
| Aligned with CS cognate | | | š | | l | i | |
| Costs | 1 | 1 | 0.5 | 1 | 0 | 0 | $\sum$ 3.5<br>3.5/6 = 58.33% |

### 4.3. Expected processing difficulty of the stimuli sentences

In Tables 4–6, the overall difficulty estimation process is demonstrated for the three stimuli sentences (xxiv)–(xxvi). This process consists of the two steps described in Sections 4.1 and 4.2.

The lexical and orthographic distances of the PL stimuli words towards the closest[10] CS translations are shown in the rows labelled *Lexical* and *Orth.* The closest CS cognate translations of every word are given in the line labelled *CS*. The expected difficulty based on the predictability of the words is indicated in the lines *Surprisal CS* and *Normalized* and refers to a model of a Czech reader in this situation. For comparison, a good CS translation (not the closest cognate translation) is visualized in Figs. 9–11 above each of the tables. The surprisal scores were obtained from the LMs trained on the PL and CS corpora (cf. Table 1). The trigram LM applied here outputs surprisal on a scale of 0–8 hartley in which a value of 8 hartley represents OOV items (*out of vocabulary words* – words that are not in the corpus). In the row labelled *Normalized,* the surprisal score is normalized to a percentage and, likewise, the expected difficulty is visualized by the colour code. The last row labelled *Assumed difficulty* of Tables 4–6 summarizes the overall predicted processing difficulty of the sentences resulting from both separately measurable dimensions as an average of the linguistic distance and the normalized surprisal score. We categorize the words within the stimuli into three different difficulty levels: green *E* for *easy* ($\leq$0.33), beige *M* for *medium* ($\leq$0.67) and red *D* for *difficult* (>0.67) (cf. colours of the predictions of processing effort in Fig. 1(a) + (b)). Easy words have low lexical and orthographic distance and are predictable in context (low surprisal score). Words with the label *difficult* have high

[10] *Closest* means: if there is a cognate translation of a PL word available, then it is used. If there are more than one cognate translations, the orthographically closest is chosen (by means of Levenshtein distance).

Table 3
Distance of PL for Czech readers: comparison of the distances of the underlying stimuli sentences to stimuli from related research. Lexical distance are the percentage of non-cognates, orthographic distance and morphological distance is measured by Levenshtein edit distance (Levenshtein, 1966).

| | Stimuli sentences | Heeringa et al. (2013) | Golubović (2016) | Jágrová et al. (2016) |
|---|---|---|---|---|
| Lexical | 12 | 23 | 17.7 | 10 |
| Orth | 38 | 31 | 31.7 | 34 |
| Morph | – | – | 31.4 | – |

Table 4
Estimation of the overall processing difficulty of the stimulus sentence (xxiv), resulting from linguistic distance and surprisal.

| PL | Nie | widziała | m | że | jego | żona | pokazuje | ręką | że | byśmy | poszli | do | rektora |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CS | Ne | viděla | jsem | že | jeho | žena | ukazuje | rukou | že | bysme | šli | do | rektora |
| Lexical | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Orth | .33 | .56 | .75 | .25 | .25 | .38 | .25 | .6 | .25 | .3 | .67 | 0 | .33 |
| Surprisal CS | 4.99 | 4.82 | 2.00 | .30 | 2.41 | 1.61 | 4.59 | 2.63 | 1.88 | 3.53 | 2.13 | 1.5 | 6.46 |
| Normalized | .62 | .6 | .25 | .04 | .30 | .20 | .57 | .33 | .24 | .44 | .27 | .19 | .81 |
| Assumed difficulty | M | M | M | E | E | E | M | M | E | M | M | M | M |

orthographic distance or are false friends and are also unpredictable in context (high surprisal score). Those words that are labelled *medium* either have low orthographic distance, but are unpredictable in context, or they are predictable in context, but distant (or have medium values for both distance and surprisal). The colour code in Tables 4–6 follows these difficulty categories throughout all rows.

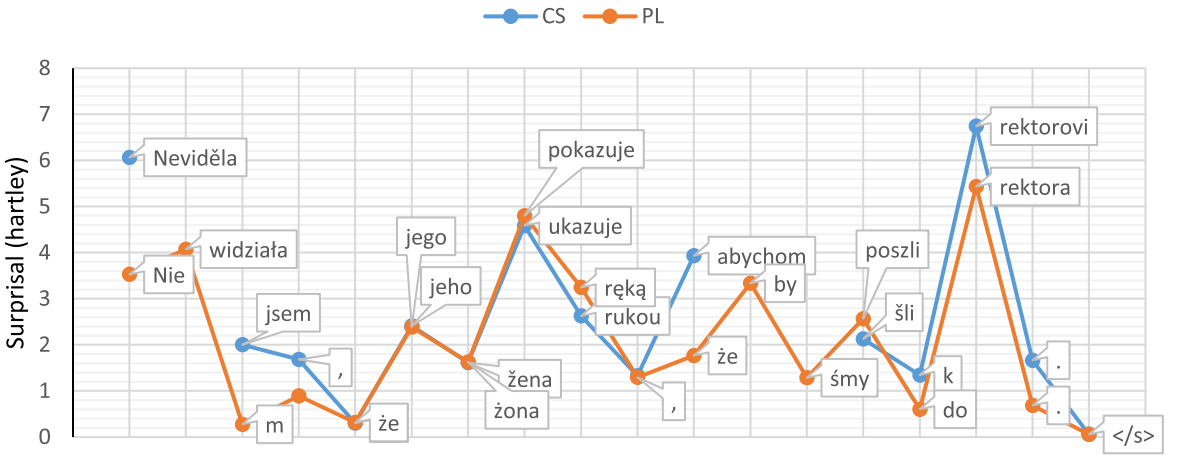



Fig. 9. Surprisal graphs of the PL stimulus *Nie widziałam, że jego żona pokazuje ręką, żebyśmy poszli do rektora.* 'I did not see that his wife is showing with her hand that we should go to the rector.' in comparison with a corresponding correct CS translation.

This PL sentence is expected not to pose any lexical difficulties to Czech readers, except for the difference in the preposition *do* 'to' which also exists in CS, but carries the meaning 'into', leading the readers to expect that a building or something that can be entered would follow the preposition. Given that both prepositions express a direction, a correct understanding of *do* by the Czech readers can be expected. We decided to separate *Nie widziałam* 'I haven't seen' [feminine] into three parts for a more accurate difficulty estimation: the negation *nie* 'no' is a separate word in PL, which in CS is realized in the form of the prefix *ne–* attached to the finite verb in its equivalent *neviděla*. Nevertheless, Czech readers are likely to understand the negation, also because they understand *nie* through their exposure to the identical Slovak *nie* 'no'. The central part *widziała* has a medium orthographic

distance (56%) to its CS equivalent *viděla* '*I/you/she/they saw*'. The past tense particle *−m* is attached directly to the feminine 3rd person ending in PL, while the CS *viděla jsem* 'I saw' [feminine] in turn is realised in two separate words: the finite verb *viděla* in past tense and the auxiliary verb *jsem* in present tense. Therefore, there turns out to be a high orthographic distance between *−m* and *jsem* (75%). The verb *żebyśmy* 'that we should' could be separated and literally transferred into the CS phrase *že bysme* as in Table 4, but in a consecutive sense as it is here, the appropriate written standard translation would be *abychom* (cf. Fig. 9) with the conjunction *aby* 'so that' instead of *že* 'that'. Consequently, *żebyśmy* is expected to be easy (20% and 30% orthographic distance). Other cases of medium orthographic distance would be *żona* 'wife' and *poszli* 'we went'. Viewing the surprisal levels of the CS translation, we observe the highest level for *rektora* 'rector' [genitive/accusative] and medium surprisal values for *ne* + *viděla*, *ukazuje* and *bysme*. Averaging over the difficulties in the two separate dimensions, only *že jeho žena* 'that his wife' is expected to be easily understandable for Czech readers, while the rest of the sentence (except the conjunction *že*) should have medium difficulty.

Table 5
Estimation of the overall processing difficulty for stimulus sentence (xxv), resulting from the two dimensions of linguistic distance and surprisal.

| **PL** | Gdyby | nie | było | książek | czytał | bym | Ci | z | oczu |
|---|---|---|---|---|---|---|---|---|---|
| **CS** | Kdyby | ne | bylo | knížek | četl | bych | Ti | z | očí |
| **Lexical** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Orth** | .2 | .33 | .13 | .42 | .67 | .5 | .5 | 0 | .63 |
| **Surprisal CS** | 4.4 | 2.4 | 5.02 | 5.76 | 4.31 | 4.14 | 1.81 | 3.34 | 2.54 |
| **Normalized** | .55 | .3 | .63 | .72 | .54 | .52 | .27 | .42 | .32 |
| **Assumed difficulty** | M | E | M | M | M | M | M | E | M |

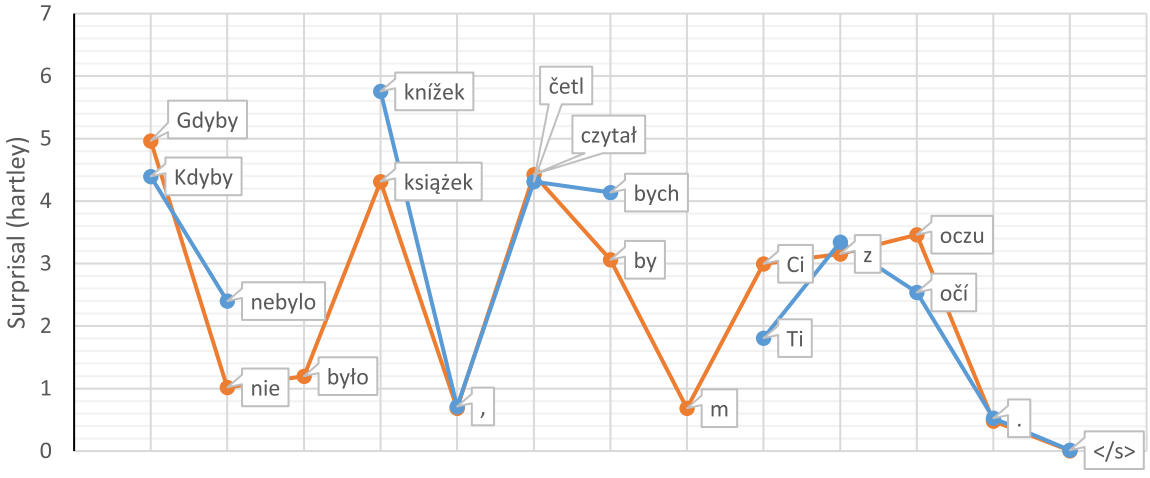



Fig. 10. Surprisal graphs of the PL stimulus *Gdyby nie było książek, czytałbym Ci z oczu.* 'If there were no books, I would read from your eyes.' in comparison with a corresponding correct CS translation.

In this sentence, we again do not encounter any lexical distance between the two languages. We separate the conditional *czytałbym* 'I would read' into *czytał* [finite verb] and *bym* [particle] for an optimal calculation according to its CS equivalent *četl bych* 'I would read' that is realized in two separate words. There is high orthographic distance between *czytał* and *četl* and medium orthographic distance in *książek* 'books' [genitive], *bym* 'I would', *Ci* 'to you' and *oczu* 'eyes'. In the dimension of context, there are only 3 instances with low surprisal: *ne* 'no(t)', *Ti* 'you', and *očí* 'eye. Resulting from the averaged difficulty of the two separate dimensions, we expect medium difficulty for most of the sentence in which only *nie* 'no(t)' and *z* 'from' are expected to be easily intercomprehensible for Czech readers.

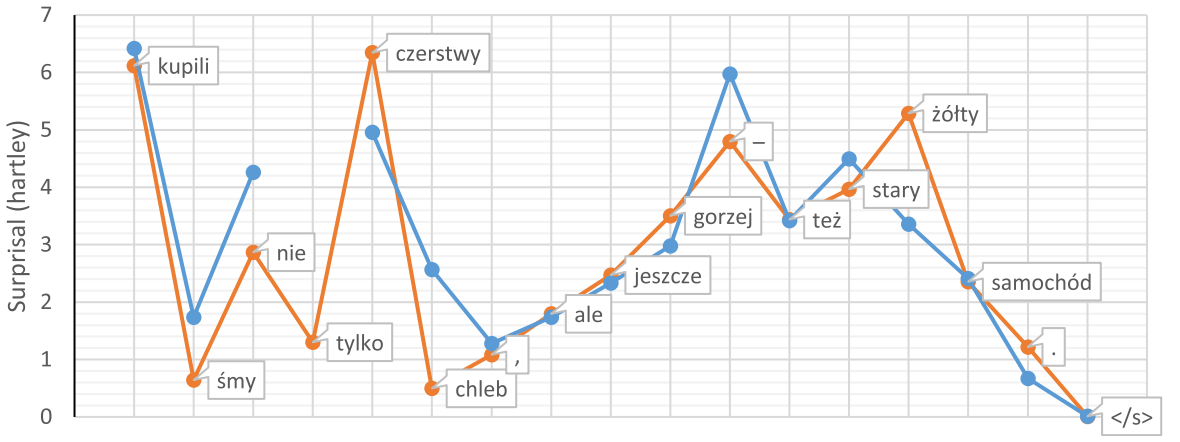



Fig. 11. Surprisal graphs of the PL stimulus *Kupiliśmy nie tylko czerstwy chleb, ale jeszcze gorzej – też stary żółty samochód.* 'Not only did we buy stale bread, but even worse, also an old yellow car.' in comparison with a corresponding correct CS translation.

Table 6
Estimation of the overall processing difficulty for stimulus sentence (xxvi), resulting from the two dimensions of linguistic distance and surprisal.

| PL | kupili | śmy | nie | tylko | czerstwy | chleb | ale | jeszcze | gorzej | też | stary | żółty | samochód |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CS | Koupili | jsme | ne | jen | tvrdý | chléb | ale | ještě | hůře | též | starý | žlutý | auto |
| Lexical | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Orth | .14 | .63 | .33 | Ø | Ø | .1 | 0 | .64 | .75 | .33 | .2 | .67 | Ø |
| Surprisal CS | 6.42 | 1.74 | 4.37 | 2.92 | 5 | 2.62 | 1.73 | 2.34 | 3.56 | 4.29 | 4.18 | 5.02 | 3.22 |
| Normalized | .8 | .22 | .55 | .37 | .63 | .33 | .22 | .29 | .45 | .54 | .52 | .63 | .4 |
| Assumed difficulty | M | M | M | M | D | E | E | M | M | M | M | M | M |

On the lexical level, the words *tylko* 'only', *samochód* 'car' and especially *czestwy* 'stale' are expected to cause difficulties for Czech readers. In the case of *czerstwy*, Czech readers are facing a false friend that might easily be mistaken for its CS homonym *čerstvý* 'fresh', meaning the opposite. This might be considered not only lexically distant, but even misleading. Hence, we are assigning a high difficulty level (2) in the distance dimension here. As for the compound *samochód*, Czech readers will understand *samo* as 'self' and *chód* as 'walker', 'goer' or 'something that walks', resulting in a concept of something moves on its own. We therefore assign a medium difficulty value to the word. There is one instance of high orthographic distance in *gorzej* 'worse' and three instances of medium orthographic distance in *–śmy* [plural marker corresponding to the CS auxiliary *jsme*], *jeszcze* 'even' and *żółty* 'yellow'. The highest surprisal score is assigned by the LM to the sentence onset *Koupili* '[we] bought' which consequently is considered medium difficult in total. Resulting from the medium surprisal score of *tvrdý* 'stale', we assign a high difficulty level to this word, expecting that it is virtually impossible for Czech readers without any knowledge of PL to comprehend it correctly.

## 5. Evidence from think-aloud protocols and results from cloze tests

In this section, we compare our predictions to the translations given by the Czech respondents during the think-aloud protocols and to the results of subsequent web-based cloze tests. As mentioned before, the 12 stimuli sentences were also presented to 23 Czech native speakers in web-based cloze experiments with the task to translate certain words within the sentences that were put in gaps. First, the respondents saw only the first word of the sentence on

their screen. They were prompted to click on the first word in order to make the next word appear. This procedure should ensure that the respondents really do read each word of a sentence, one by one. Only after clicking on the last word in the sentence, the gaps appeared in which the PL word(s) should be substituted by a CS translation. We are also looking at whether the UID hypothesis holds for an intercomprehension scenario in which readers have to fill a comprehension gap. According to the uniform information density (UID) hypothesis of Jaeger (2010), "encoding mechanisms will seek to avoid peaks and troughs in surprisal" (Crocker et al., 2015). It postulates that denser encodings emerge in predictable messages. In this aspect, we also view the productive side of a free translation task. The surprisal scores (obtained from the LMs trained on the PL and CS corpora) of the answers are visualized in Figs. 12–14.

This section presents the written translations of the sentences xxiv–xxvi given by the test subjects during the think-aloud experiments. In the discussion of these, citations from the transcripts of the audio-recordings and results from the cloze experiments are added respectively. Mistakes in the translations are underlined:

(xxiv)

a. *Nevěděla jsem, že jeho žena navrhuje, abychom šli za učitelem.* (ID 16)
'I did not know that his wife is suggesting that we should go see the teacher.'
b. *Nevidím, že jeho žena ukazuje na chlapce, aby šel k řediteli.* (ID 15)
'I do not see that his wife is pointing at the boy that he should go to the headmaster.'
c. *Nevypadá, že jeho žena . . ., měli bychom jít k řediteli.* (ID 14)
'It does not look as if his wife . . ., we should go to the headmaster.'
d. *Nemyslím si, že tudy poteče řeka, měli bychom poslat pro ředitele.* (ID 12)
'I do not think that a river will flow here, we should send for the headmaster.'
e. *Neviděl jsem, že jeho žena ukazuje rukou, že bychom měli jít doprava.* (ID 5)
'I did not see [masculine] that his wife is showing with her hand that we should go right.'
f. *Nepřeji si, aby jeho žena navrhovala, abychom šli za rektorem.* (ID 1)
'I do not wish that his wife suggests that we should go see the rector.'

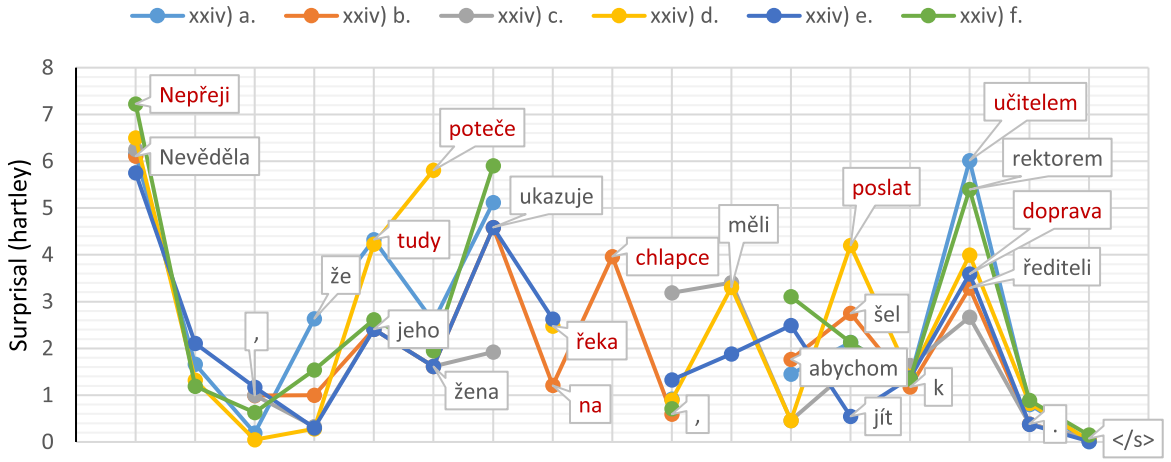



Fig. 12. Surprisal graphs of answers (xxiv) a.–f. Wrong translations of individual words are marked red in the data labels. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

According to the prediction of processing effort in Table 4, we expect medium difficulty (represented by an *M*) for the whole sentence with the exception of *że jego żona* 'that his wife' which was correctly translated by five of the six respondent pairs viewed here. Viewing the written translations xxiv a.–f., we identify three parts of the sentence that seemed to cause the greatest problems: the onset with *nie widziałam* 'I did not see', the end of the subordinate clause with *pokazuje ręką* 'is showing with her hand' and *do rektora* 'to the rector' at the end of the sentence. In the subsequent web-based cloze tests, the gaps were placed on these critical words and phrases. For *nie widziałam*,

there is a variation in the translations regarding gender (masculine in xxiv) e. and 17% in the cloze tests – not only by males), tense (present tense in xxiv) b.–d. and f.; 52% in the cloze tests) and the actual verb (only 17% translated the verb correctly with a form of *to see*). Nevertheless, all six respondent pairs and 83% of the cloze test respondents understood the negation here.

As for *pokazuje ręką*, only the respondent pair xxiv) e. and only 1 of 23 respondents in the cloze tests entered the correct translation. Regarding surprisal, the word *ręką* 'with her hand' is very uninformative here, if not even semantically redundant. The meaning of the sentence would be the same even if this word was omitted: *Nie widziałam, że jego żona pokazuje, żebyśmy poszli do rektora.* 'I did not see that his wife is showing that we should go to the rector'. At the same time, this word is relatively orthographically distant (60%) to its translation *rukou*. In the translations (xxiv) a. and f., *ręką* is ignored by reformulating it together with the preceding word *pokazuje*, e.g. "[Pokazaj] is that something like *show me*? That his ... *showed*."[11] (ID 1). Or the respondents are trying to assign a greater informativity to it which results in misinterpretations (xxiv) b. and d., e.g. "So that's *I'm not saying that his wife is* [pokazuje] ... *instructing*. Well, that's probably *instructing*. *Instructing the hero*.", with *ręką* understood as *reka* 'the hero' resp. *rekovi* 'to the hero' [accusative resp. dative of *rek* 'hero'], which is subsequently turned into the less surprising *chlapovi* 'the guy' [dative]: "Že jeho žena ukazuje chlapovi, aby šel k řediteli." (ID 15) and *na chlapce* 'at the boy' in the translation that was entered (xxiv b.). In the cloze tests, 26% translated *ręką* with *řeka* 'river', 13% with *reka* 'hero', 9% with *říká* 's/he says', and 9% with *na* (*to*) 'at (it)'. The solutions *řeka*, *reka* and *říká* are all orthographically closer to the stimulus than the correct translation *rukou* 'with her hand'. Apparently, most of the cloze test respondents focussed more on the similarity of the stimulus *ręką* to a possible CS word, while only few obviously understood the preceding words and added what would make most sense in context – showing/pointing *at (it)*.

In Fig. 9, we observe a surprisal peak (5.43 hartley in PL, resp. 6.74 hartley in CS) for *rektora* 'rector' [accusative]. The word is considered transparent (actually identical in nominative case), but has a high surprisal score and therefore is expected to have medium difficulty. This is most likely due to the relatively low frequency of this word in both of the corpora. Indeed, only one of six respondent pairs entered the correct translation. Seeing only the written responses from the think-aloud experiments, one would assume that the word is opaque to the readers. In contrast to this, the audio recordings reveal that *rektora* is transparent, but readers do not expect this concept here. The respondents actually were talking about the *rector* (also *rektor* in CS), but most of them did not trust this obviously identical word and were trying to assign a different meaning to it: "Rector, headmaster or something, isn't it? Like, also in our country, is there a rector?" (ID 14). For example, respondent pair 5 moves away from the concept of the rector, ending up with a re-interpretation of the whole phrase: "That's probably not going to be a rector as such. [...] Am I visiting the rector or what? [...] Rect ... recht from German [...] That we should go to the right. That could be it, mhm, something like that. That sounds good" (ID 5). Also respondent pair 1 and 16 distrusted the obvious *rector*, with pair 16 replacing it by the more common *teacher*: "[....] that we should go, but what is rector, right? .] That's not going to be a university rector as for me. [...] something like [...] a teacher?" (ID 16). "[...] if rector is for instance not a headmaster maybe. That's probably not a rector of a uni. [...] What could a rector be, except a rector?" (ID 1). This phenomenon is also reflected by the cloze test results. The respondents were asked to translate the entire NP – *do* and *rektora*. None of the respondents entered the correct translation *k rektorovi* 'to the rector'. 23% simply re-typed the stimulus *do rektora* which would have a different meaning in standard CS (*do* means 'into' in combination with persons). However, the use of *do* in the PL sense might be known from the Moravian dialect in which it has the same meaning as in PL. One respondent entered a wrong preposition: *od rektora* "(away) from the rector". All other answers were, except two (*k doktorovi* 'to the doctor' and *do učitele* 'into the teacher'), things or places instead of persons: *do koryta* 'into the trough', *do potoka* 'into the river', *do vedení* 'to the administration', *do města* 'to the city', *do banky* 'to the bank', *do kostela* 'to church'. This again confirms that the word *rektor*, representing a person, is unlikely to follow the preposition *do* here. Hence, its high surprisal might not only be caused by the fact that the word *rektor* itself is very rare in the CS corpus, but also that it is unlikely to follow the prepostion *do*.

When viewing the encoding perspective of the translations here, we see that the scores for the different translations *učitel* 'teacher', *ředitel* 'headmaster', and *doprava* 'to the right' have lower surprisal scores than the correct translation would have had. We assume that if *rektora* was embedded in a NP with a frequent collocation, such as

*rektora uniwersytetu* ‘rector of the university’ [accusative], its surprisal score would be lower and there would not be that much room for speculation.

In contrast to this, the respondents’ translations for the sentence onset have higher surprisal scores than *neviděla jsem* ‘I didn’t see’ [feminine] has in the correct translation (6.07 and 2.00 hartley). An increase in surprisal is apparent in answer xxiv) d., where *poteče řeka* ‘a river will flow’ exceeds the surprisal score of the actual *jeho žena* ‘his wife’ (2.41 and 1.61 hartley). This might be due to the high similarity of *řeką* ‘hand’ [instrumental case] to *řeka* ‘river’ which seems to be a very dominant factor here. Likewise, this dominance of the orthographic similarity manifests itself also in translation xxiv) b. with *ukazuje na chlapce* ‘[she is] pointing at the boy’, most probably because *řeką* is read and, ignoring the diacritics, pronounced as *reka* ‘hero’ [accusative], resulting in an interpretation that his wife is pointing at a young male person who consequently should go to the headmaster.

(xxv)

a. *Kdyby nebylo knížek, četl bysem si z očí.* (ID 16)
‘If there were no books, I would read from my/people’s eyes.’
b. *Kdyby nebylo slov, četl by mi z očí.* (ID 15)
‘If there were no words, he would read from my eyes.’
c. *Kdyby nebylo knih, sešel by z očí.* (ID 14)
‘If there were no books, he would get out of sight.’
d. *Kdyby nebylo knížek, četli by jsme druhým z očí.* (ID 13)
‘If there were no books, we would read from other peoples’ eyes.’
e. *Kdyby nebylo knih, četl by mi otec.* (ID 7)
‘If there were no books, my father would read to me.’
f. [no answer entered, only recording and transcript available] (ID 3)

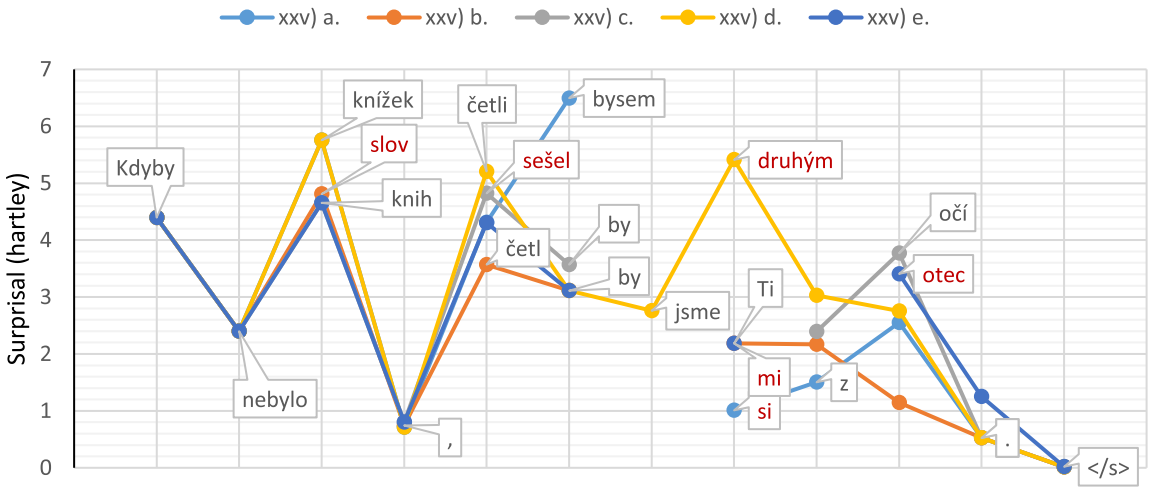



Fig. 13. Surprisal graphs of answers (xxv) a.–e. Wrong translations of individual words are marked red in the data labels. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

As[12] expected in the estimation in Table 5, the negation *nie było* ‘were no’ in the conditional clause is correctly recognized by the respondents in (xxv) a.–f. The respondents performed better than expected at identifying the conditional expression *czytałbym* ‘I would read’ in the main clause as a conditional, but some failed at identifying the correct gender (xxv) b.–e.) and number of the verb (xxv) d.). The respondents in (xxv) c. were apparently misled by

[12] The stimulus was modified here in a way that the *s* in *książek* ‘books’ [genitive] was replaced by an *n* (closer to the CS cognate *knížek*): *Gdyby nie było kniążek, czytałbym Ci z oczu.*

their assumed pronunciation of *czytałbym* as in *schytal* 'would get punished' [colloquial]: "He would get punished. That's like, it just reminds me of some pronunciation, right?" (ID 14). Only after having identified *z očí* as 'from the eyes', they reinterpreted the verb accordingly: "So, from the eyes? He would get out of sight, or not? Aha, could be. Man, I don't know. How about books? Aha, could it be? That would make sense. If there were no books, he would get out of sight? [. . .] That would make sense, wouldn't it?" (ibid.).

In the cloze tests, respondents were asked to translate the whole clause *czytałbym Ci z oczu*. Only 9% of the cloze test respondents entered a correct CS translation of the clause. 26% provided a translation in conditional form, 22% mentioned *oči* 'eyes' and 17% translated a form of *číst* 'to read'.

The peak at *bysem* 'I would' can be explained by the relatively low frequency of this rather Common Czech variant in the corpus – it is most often realized as *bych* 'I would' in standard written CS. There does not seem to be a uniform pattern in the translations given for the main clause. The audio recordings reveal that respondents first try to decode *książek* or the phrase *z oczu* 'from [your] eyes" and only then make up the rest of the phrase around it (cf. transcripts in the appendix). There seem to be two variants: either *otec* 'father' (sentence xxv) e. and transcript ID 3, based on the wrongly assumed pronunciation of *oczu* which is associated with *otcu* – dialectal for *otci* 'to the father') or the correct translation *očí* 'eyes' (sentences xxv a.–d. and f.). Viewing only the two words preceding *oczu*, the CS trigram model indicates that the correct translation for *očí* (genitive plural of *oko* 'eyes') is the 9th most likely word that could fit into this comprehension gap,[13] which again indicates that readers should find it reasonable in context. The translations given in the cloze tests also reveal some cases of apparently wrongly assumed pronunciations of *oczu*: *s ovcí* 'with a sheep', *z octu* 'out of vinegar', and *z ocasu* 'from (my) tail'.

In this stimulus sentence, there is no such case of a highly transparent and at the same time highly surprising word as in (xxiv). Instead, the respondents encountered medium orthographic opacity in *książek* together with a relatively high surprisal of this word, which still resulted in correct translations in (xxv) a. and c.–e. An alternative to the diminutive expression *knížek* 'little books' is the more frequent formulation *knih* 'books' [genitive] (considered correct) which results in somewhat lower surprisal scores as in the translations (xxv) c. and e.

Also, the translation in xxv) b. *slov* 'words' [genitive] has a lower surprisal score than *knížek* 'books' [diminutive], but a slightly higher one than *knih* 'books'. In the recordings of respondent pair 15, we observe that one of the respondents actually had already pronounced the correct translation "If there were no books" (ID 15) right at the beginning. However, both respondents seem to discard this: "Well, books, that's probably not it. [. . .] it makes some logical sense, like, you know, these words." (ibid.). Another remarkable case of discarding a correct translation happened in xxv) f.:

> "If there were no books, I would read from your eyes. But that doesn't make any sense, don't you think? That doesn't make any sense, but . . . what else could it be? [. . .] Well, even about these eyes we don't even know that these are eyes. What would be father? Like, I don't know, this could be . . . well, that would make even less sense. But you see that this could be some form of fath . . . [. . .] Simply if there wasn't this obstacle, I would look directly into your eyes. That sounds better." (ID 3)

Other respondent pairs find the context perfectly reasonable, e.g. "If . . . sure. Books and was reading, that makes sense, so [kšiažek] is . . . [. . .] And eyes, same over here, probably nothing else . . ." (ID 16).

There seems to be some of room for phantasy in the interpretation of *Ci* 'you' [dative], which is successfully recognized as a pronoun by most of the respondents. The capital C might have been a hint, e.g. "And what is this C I there, like? [reading] like we don't have one word. If this [Ci] is, plays some role and also has a capital C. [. . .] Hm, if there were no, [reading]. That sounds like some pronoun [. . .]" (ID 16). In the cloze tests, *Ci* was translated correctly only by those 9% who translated the whole clause correctly. One respondent translated it with *mi* 'me' [dative].

[13] Other more likely PPs according to the LM would be *z nás* 'from us', *z toho* 'from this', *z nich* 'from them', *z nosu* 'from your nose', *z vás* 'from you', *z hlavy* 'from your head', *z pusy* 'from your mouth', and *z obličeje* 'from your face'.

(xxvi)

a. *Nekoupili jsme jenom čerstvý chléb, ale ještě hůř – taky staré žluté auto.* (ID 16)
'Not only did we buy fresh bread, but even worse – an old yellow car.'
b. *Nekoupili jsme pouze čerstvý chléb, ale jeste povoz – též staré žluté kolo.* (ID 15)
'We did not buy only fresh bread, but also a vehicle – an old yellow bicycle.'
c. *Nekoupili jsme dostatečně čerstvý chléb, ale ještě hůře – takový starý žluklý, zkažený.* (ID 14)
'We did not buy sufficiently fresh bread, but even worse – such an old, yellow, rotten one.'
d. *Koupili jsme nejen čerstvý chléb, ale ještě hůř – taky staré zlaté auto.* (ID 10)
'Not only did we buy fresh bread, but even worse – also an old golden car.'
e. *Koupili jsme ne tak starý chléb, ale ještě teplý – taky staré žluté auto.* (ID 8)
'We did not buy such old bread, but it was still warm – also an old yellow car.'
f. *Nejen, že jsme nekoupili čerstvý chléb, ale ještě hůře – také staré [_] auto.* (ID 1)
'Not only did we not buy fresh bread, but even worse – also an old [_] car.'

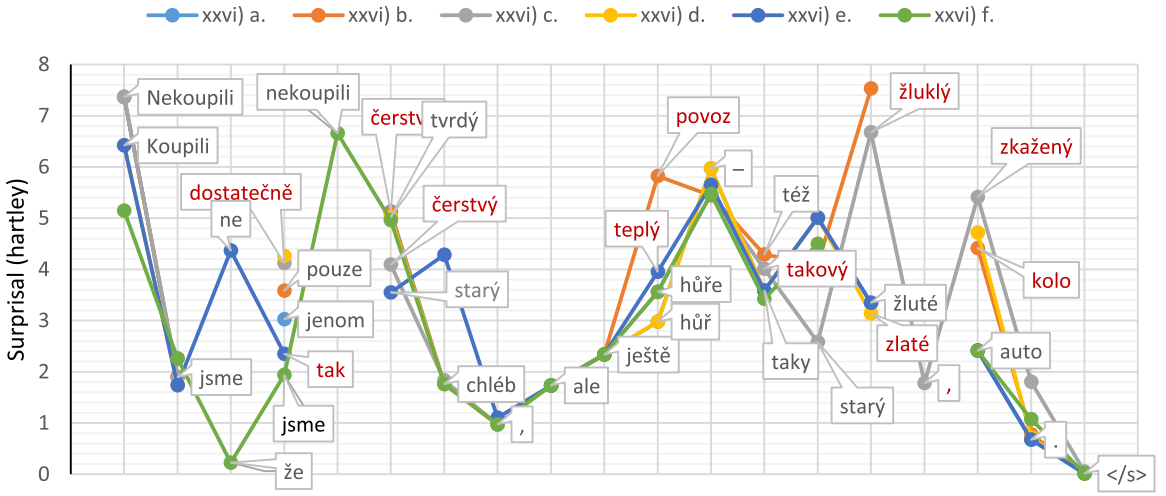



Fig. 14. Surprisal graphs of answers (xxvi) a.–f. Wrong translations of individual words are marked red in the data labels. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Despite the relatively high surprisal at sentence onset, there does not seem to be a problem in understanding the finite verb *Kupiliśmy* 'we bought' together with the negation that was either transferred into CS as a negation of the verb (xxvi a.–c. and f.) or as a negation of *tylko* 'only' that was partly misinterpreted as *tak* 'so' (xxvi e.) or *dostatečně* 'sufficient' (xxvi c.).

In accordance with the estimation in Table 6, the string *chleb, ale* 'bread, but' did not pose any problems to any of the respondents. *Chleb* 'bread' seems to dominate the semantics and the interpretation of the whole sentence strongly. In order to capture the role of this lexeme in this sentence, other models reflecting longer ranges might be more suitable than trigram models.

Contrary to the expectations, also *jeszcze* 'even', *też* 'also' and *stary* 'old' seem easy to decode, as there were no doubts about them uttered by the respondents. Only respondent pair 15 did not type an appropriate translation of *też*, obviously because of a creative interpretation of *samochód* 'car' and *żółty* 'yellow' in connection with *chleb* 'bread': "But even worse, such an old, rancid ... self-goer, well. How do you say that ... we always say that it's already walking. That it's, you know, that it's so old that it started walking already." (ID 15). In this case, the respondents prefer *žluklý* 'rancid' to *žlutý* 'yellow' in order to provide a logical connection to the bread.

As for *samochód* 'car', the surprisal scores of the translations *kolo* 'bike' and *zkažený* 'rotten' are higher than the actual score of the correct translation *auto* 'car' would be. It is probably because of its possible literal translation as *samochod* 'self-goer' that there are basically no limits for the imagination of the respondents (see appendix for numerous examples of intermediate translation variants). According to our expectations, the

difficult stimulus word *czerstwy* 'stale' was identified by only one of six respondent pairs, because one of the respondents was aware of this false friend: "Hey, I actually know this word, dude, because I was once talking to a Polish guy about which words are the same and which are different and he told me directly that *czerstwy* means ... that *czerstwy* simply means the opposite in Polish." (ID 8). All cloze test respondents translated *czerstwy* with its CS false friend *čerstvý* 'fresh'. The other problematic word in this sentence was clearly the non-cognate *samochód*. The translations of the phrase *stary żółty samochód* 'old yellow car' given in the cloze tests confirm the dominant role of *Kupiliśmy* and *chleb* in this sentence. Besides the correct translations of this phrase (26%), 17% were in connection with the topic of grocery shopping. One answer reflects this perfectly: *starý zlatý samoobchod* 'old golden/good old 'self-shop' (lit.)'. The word *samoobchod* seems to combine both *samoobsluha* 'supermarket' and *obchod* 'shop'. Two other respondents entered other food-related variants: *starou žitnou bagetu* 'an old rye baguette' and *starý zlatý samovar* 'an old golden/a good old samovar'. 22% translated *żółtý* 'yellow' as *zlatý* 'golden', which might be explained by the CS collocation *starý zlatý* that can be translated 'old golden' or 'good old'. Consequently, *zlatý* is more likely to follow *starý* than *žlutý*.

## 6. Summary and discussion

We presented a method for estimating the overall processing difficulty of individual words in sentence stimuli resulting from the two orthogonally measurable dimensions of linguistic distance and surprisal in context. We applied the method on PL stimuli sentences and translations of them given by Czech respondents in reading intercomprehension experiments. We analysed the written answers that were given by the respondents as well as the audio recordings of the respondents during a translation task in think-aloud protocol design. Additionally, we evaluated the translations of a number of critical words and phrases within the same stimuli sentences that were gathered in web-based cloze experiments. We compared the predicted difficulties with the experimental results.

Overall, the results show that the predictions do not always agree with the actually observed difficulty of the stimuli. Contrary to our expectations that even absolutely transparent words such as internationalisms would be comprehensible in no matter which context, we discovered that high surprisal scores can ruin the intelligibility advantage that identical words or words with low orthographic distance actually have. The audio recordings bring further insight into the decoding process than if only the written translations of the respondents were considered. Although respondents pronounce the correct translations of words such as *rektor* 'rector' or *auto* 'car', they do not trust these obvious words, because, in their opinion, they do not fit very well in the context of the remaining sentence or they are simply surprising because they are used rarely. Nevertheless, we also observe that readers' opinions about what does and what does not make sense in context can differ and that they do not always agree with the surprisal scores determined with the help of LMs.

However, when viewing orthographically distant words with low surprisal scores, surprisal influences readers' performance only to a point until there are other linguistic features that can have a more powerful influence on understanding, depending on the actual stimulus (e.g. initial letter, neighbourhood density). Knowing how strong the role of context is in these stimuli allows us to draw conclusions about the role of other influencing factors and their possible dominance.

Regarding encoding, our findings suggest that the UID hypothesis does not hold for the translations given by the respondents in the three stimulus sentences: they did not avoid peaks and troughs in surprisal. However, the UID hypothesis refers to communicative situations, which was not the case in our experimental design. There is one observation that can be made regarding the encoding of the translations: for most stimuli with a high surprisal score of 6 hartley or more, respondents showed a tendency to provide translations that would have a lower surprisal score. As stated in the introduction, this contribution does not claim to provide statistically sufficient data on the understanding of PL stimuli sentences by Czech readers. It serves the discussion of certain phenomena influencing intercomprehension in certain stimuli and it is an attempt to use LMs in order to describe the role of context in the stimuli and translations thereof.

The findings of this study are the basis for further research into the topic of reading intercomprehension of sentences and the role of context. A possible alternative would be not to view only the absolute surprisal scores, but the difference in surprisal to the preceding words, especially in frequent collocations, e.g. *chléb* 'bread' is very predictable after *čerstvý* 'fresh' which leads to a decrease in surprisal. Also, the analysis could be repeated with other n-gram models or with LMs other than n-gram models in order to capture longer contextual influences than only trigrams (e.g. neural networks).

## Vitae

Irina Stenger, Andrea Fischer, and Klára Jágrová are doctoral researchers at Saarland University. Tania Avgustinova is a professor of Computational Linguistics at Saarland University. The team works in a DFG-funded project on Intercomprehension and Surprisal in Slavic Intercomprehension (INCOMSLAV) which is part of the Collaborative Research Center Information Density and Linguistic Encoding (SFB 1102). Their research focusses on the processes involved in the intercomprehension of Slavic languages.

## Funding

This study was carried out in the context of a larger research project on mutual intelligibility among Slavic languages, concentrating mainly on BG, CS, PL, and RU. The INCOMSLAV project (Mutual Intelligibility and Surprisal in Slavic Intercomprehension) is part of the CRC 1102 – Information Density and Linguistic Encoding at Saarland University, funded by the DFG.

## Thanks

We wish to thank our colleagues from Charles University in Prague, specifically our colleagues at ÚFAL for providing us the workspaces for the think-aloud experiments and our colleagues Iva Poláčková Šolcová and Ivan Rynda from the Faculty of Humanities as well as Tomáš Svoboda from ČVUT Prague for giving us the possibility to gain student respondents, to whom we also owe special thanks. We also thank the anonymous reviewers for their supportive comments and improvements.

## Appendix

Transcripts of think-aloud protocols for the stimuli sentences (xxiv)–(xxvi).

Table A.1
Transcripts of think-aloud protocols for the stimulus sentence (xxiv). *Nie widziałam, że jego żona pokazuje ręką, żebyśmy poszli do rektora.* 'I did not see that his wife is showing with her hand that we should go to the rector'. The written translations as provided by a particular pair of participants are marked with a grey background colour right next to the participants' ID. The original Czech solution is given in the left and the English translation in the right column. The original CS transcript (left column) is translated into EN (right column). Speculations about the plausibility of words in context and other relevant passages are marked bold. The EN translation is not complete for reasons of effort and relevance. If you wish to add a translation or suggest corrections, please contact the authors at kjagrova@coli.uni-saarland.de.

| ID 1 | Nepřeji si, aby jeho žena navrhovala, abychom šli za rektorem. | I do not wish that his wife suggests that we should go see the rector. |
|---|---|---|
| A | Ně vidzjalam, že jego žona pokazuje rjeka, žebysmi pošli do rektora. To je, ně vidzjalam... | [reading] That's [reading] … |
| B | Tak jego žona, to je jeho žena. | So, [jego żona] is his wife. |
| A | Ty jo, to je fakt, to by mě nenapadlo. | Wow, that's right, I wouldn't have noticed that. |
| B | Ale... žebysme pošli do rektora... pokazuje reka... | But … [reading] … |
| A | Nebo vz... … vzkazuje... … že bysme pošli do rektora. Že bychom zašli za rektorem? No, to teďka, jestli rektor neni třeba ředitel. On to možná nebude rektor jako vejšky. | Or … leaves a message … that we should go see the rector. That we should go see the rector? Well, now, if rector is for instance not a headmaster maybe. That's probably not a rector of a uni. |
| B | Rektor... | Rector … |
| A | Ňje vidžalam... vidžalat, vidžalat... | [reading] |
| B | Vidžjalam... | [reading] |
| A | Vidžjalam, jo... … to je... … jo, kdybych tady měl Google překladač, ty jo. | [reading] yeah … that's … yeah … If I had google translate, man. |
| B | Hm. | Hm. |
| A | Né, to tak neni. | No, that's not. |
| B | .....že jego žona pokazuje reka... … to nevím, jak se ani čte tyhlety písmena v tom reka. | [reading] I don't know, even don't know how to read these letters in this [reka]. |
| A | Ale jako slovensky, žjajam, žjalam, žjalam... … jestli to nebude stejný, podobný. | But like Slovak, [I wish, I wish, I wish] … if that's maybe not the same or similar. |
| B | To může bejt... | That could be … |
| A | No... | Well … |
| B | Jako nepřeju si... | Like, I don't wish that … |
| A | Nepřeji si... | I don't wish … |
| B | ...aby jeho žena... | … that his wife … |

(*continued*)

| | | |
|---|---|---|
| A | ...aby jeho žena pokazuje řika, pokazuje reka, aby jeho... | … that his wife [reading] that his … |
| B | Žebysme, žebyšmi pošli do rektora. | That we, that we [reading]. |
| A | Aby jeho žena jako zařídila něco, že bysme šli za rektorem. | That his wife should like go organize something so that we can go see the rector. |
| B | Hm. | Hm. |
| A | Ale to nepřeju si, to se mi tam líbí, to bych tam dal. | But this I don't wish, I like that there, I would put it there. |
| B | Ně vidžjalam... … aby, že, aby jeho žena pokazuje reka. Nepřeji si, aby jeho žena pokazuje... … pokazuje reka... … Pokazaj, není to něco jako ukaž? Aby jeho... … ukázala. | [reading] that, that, that his wife [reading]. I do not wish that his wife [reading] … [Pokazaj] is that something like show me? That his … showed. |
| A | Na mě to třeba působí tak jako, aby jeho žena doslova jako nepráskala, že bysme za rektorama, nevím. To se mi nezdá. | I have the impression that his wife literally shouldn't go tell on us that we would … to the rectors, I don't know. I find that weird. |
| B | Ne vidžjalam, nje vidzjalam, že jeho žona ukazuje reka, žebyšmi... že bysme šli do rektora... ukazuje řeka... | [reading] |
| A | Hmm... | Hmm... |
| B | Nevíte-li slovo, některé, vyvoďte si je z kontextu anebo hádejte. No, tak... … ukazuje reka... … džjalam, nje vidzjalam, nje vidzjalam... … ukazuje... … abychom šli za rektorem. Asi to bude končit, abychom šli za rektorem. | [reading task: if you don't know a word, derive it from the context or guess.] Well … show [reka] … [reading] that we should go see the rector. That probably ends with that we should go to the rector. |
| A | Hm, asi jo. | Yeah, probably. |
| B | Tak třeba, nepřeju si, aby jeho žena navrhla nebo navrhovala... | So, for example, I do not wish that his wife suggests or suggested … |
| A | Asi jo... | Probably yeah … |
| B | Abychom šli... … nic z hlediska jako smysluplnějšího mě nenapadá. | That we should go … I have no idea what would make more sense here. |
| A | Hm... … asi tak. | Hm, … well, then. |
| B | ..…aby jeho žena navrhovala, abychom šli za rektorem. Takový složitý sou... … trošku složitý souvětí. | … that his wife suggests that we should go to the rector. Such a difficult comp… i little bit difficult compound sentence. |
| A | Hm. | Hm. |
| B | Ne vidzjala... … pošli do rektora, co může bejt rektor, kromě jako rektora? Hm, tak jo. Asi jo. | [reading] … what could a rector be, except a rector? Well, probably yes. |
| A | Hm, asi tak, no. | Well, probably like that, yeah. |
| ID 5 | Neviděl jsem, že jeho žena ukazuje rukou, že bychom měli jít doprava. | I did not see [masculine] that his wife is showing with her hand that we should go right. |
| A | Tak… že bysme šli někam. Do rektora, no. | Ok ... … That we should go somewhere. To the rector, well. |
| B | No, takže, ale prostřede… prostředku je jeho žena, na něco poukazuje. | Well, ok, but in the mid... … In the middle there is his wife, she's pointing at something. |
| A | **To asi nebude rektor jako takovej.** | **That's probably not going to be a rector as such.** |
| B | Do rektora… | Into the rector ..…. |
| A | Hm, hm. | Hm, hm. |
| B | **Pujdu navštívit rektora nebo?** | **Am I visiting the rector or what?** |
| A | Mhm. Vidžielam, vidzielam, vidžiauam. | Mhm. [reading] |
| B | Hm. Tak ně bude určitě zápor. | Hm. Ok, nie is surely a negation. |
| A | Mhm, mhm, to tam není… | Mhm, mhm, that's not there … |
| B | Viděu, vidžau, vidžaua. | [reading] |
| A | Hm, nevi- nevidět. | Hm, not se… not seeing. |
| B | Mm, něvidžaua. | Mm, [reading]. |
| A | Aha. | Aha. |
| B | Nevidí nebo neví. No… | Doesn't see or doesn't know. Well … |
| A | Aha. | Aha. |
| B | Tam bych dal minulý čas, něvidžaual. | I'd put past tense there, [reading]. |
| A | Renkou – rukou třeba? | [Reading] - with her hand maybe? |
| B | Že, jakože nevěděl, že jeho žena pokazuje renkou. | Like, like I didn't know that his wife [reading]. |
| A | Aha. A… | Aha. And … |
| B | Bych… rukou, to rukou by mohlo být, že ukazuje rekou, bychom pošli. | I would … with her hand, that with her hand could be that she's showing with her hand that we should [reading]. |
| A | Mhmm… | Mhmm… |
| B | Co kdybych ho mohl… poslali místo toho… | What if I could … sent him instead of that … |
| A | Jo, jo, jo, určitě. | Yeah, yeah, yeah, sure. |
| B | No tak, tak já napíšu aspoň prostřední část, pokaď že to máme… | Well then, so I'll write down at least the middle part as far as we got it … |
| A | A, aha, jo vidím tam zvláštní znaky, když máš žena třeba, tak mám, tak vidím jinej znak, nebo u… | And, aha, yeah, I'm seeing strange signs when you're writing žena for example, then I got, then I see another sign there, or at … |
| B | Že jeho žena je, ukazuje rukou… Vidíš to, co píšu nebo to mám jenom já? | That his wife is, showing with her hand ... … Do you see what I am writing or do only I have it? |
| A | A… ano, něco takového. Ukazuje rukou, vidziela. | Y... … Yes, something like that. She's showing with ther hand [reading]. |
| B | Dobře. Ukazuje rukou… | Ok. She's showing with her hand … |
| A | Ano. | Yes. |
| B | Že bychom, že byšmi pošli, je to my… | That we should, that we should [reading], that's for me … |
| A | Ne, ne, ne, to, to se mi zdá jako jít. | No, no, no, that, that seems to me like go. |
| B | Že bychom my… poš… pošli, jako pošli. To snad ředitel asi nebude, no. | That we … pass … pass away, like pass away. That's probably not going to be the headmaster, yeah. |
| A | Tak to mam. Mně tam nějakým důvodem naskakuje kostel… | So, I got that. For some reason, a church appears to me there. |
| B | Posli… že bychom šli… a do rektora. | [Reading] … that we should go … and into the rector. |
| A | **ale to je divný.** | **but that's weird.** |

(*continued*)

| | | |
|---|---|---|
| B | Po stranách… ko- kostul nebo něco takového, myslím. Jsme jednou v Orlickejch horách překročili do Polska. Tam byl nějaký ten dřevěný kostul nebo něco takového. | On the sides … ch-church or something like that, I think. Once we crossed the border to Poland in the Orlické hory. There was such a wooden church or something like that. |
| A | Dřevěný kostul. | Wooden church. |
| B | Nevím, jestli jsem to tam měl napsat, že mám takové jazykové zkušenosti. | I don't know if I should have written it down there that I got this language experience. |
| A | To by mohlo být, jo, jo, jo. | That could be, yeah, yeah, yeah. |
| B | Asi zpátky k vážně. Njevidžauam. Je nevěděl jsem? Nevěděl jsem… | But seriously. [reading] is I didn't know? I didn't know … |
| A | Ale zase co, co, že, že jeho žena poukazuje rukou nebo ukazuje rukou? To je… že by. | But again, what, what, that, that his wife is pointing with her hand or showing with her hand? That's … like. |
| B | Že jeho žena, že jeho žena ukazuje rukou. | That his wife, that his wife is showing with her hand. |
| A | A a nebo třeba kyne… jako… Aha. | Or maybe she's waving … like … aha. |
| B | Hm, tak co, takže by to bylo ne- neviděl jsem, že ona, ona ukazuje, kam bychom mohli jít? | Hm, so what, so that would be I didn- I didn't see that she, she's showing where we could go? |
| A | Jo, jo, jo, jo, jo, jo. Ukazuje, že by se… | Yeah, yeah, yeah, yeah, yeah, yeah. She's showing that we … |
| B | Tedy ona navádí a určuje směr. A já jsem ne- neviděl… | Meaning she's guiding and determining the direction. And i didn- didn't know … |
| A | Aha… a třeba doprava. | Aha … maybe to the right. |
| B | Neviděl jsem, že ona mi ukazuje cestu. Neviděl jsem, že jeho žena ukazuje rukou, že bychom měli jít? By bylo potom česky… | I didn't see that she's showing me the way. I didn't see that his wife is showing with her hand that we should go? That would be Czech … |
| A | Nebo nebo… no, no, no. Nebo rektum jako pravý. | Or or … well, well, well. Or rectum like right. |
| B | Žě bychom mě- měli jít. Rekt… recht z němčiny, by bylo. | That we sh- should go. Rect … recht from German, would be. |
| A | Jo, jo, jo. | Yeah, yeah, yeah. |
| B | Abychom měli jít doprava. To by mohlo být, mhm, něco takovýho. To zní dobře. | That we should go to the right. That could be it, mhm, something like that. That sounds good. |
| A | Aha. Ano, tak to už je skoro stejně, že bychom měli jít a abychom šli. | Aha. Yes, so that is almost the same that we should go and that we better went. |
| B | Že, neviděl jsem, že jeho žena ukazuje rukou, že bychom měli jít doprava. Nebo aby, abychom šli doprava. Mhm, asi jo nebo významově určitě. | Like, I didn't see that his wife is showing with her hand that we should go to the right. Or that, that we better went to the right. Mhm, probably yeah, but meaning-wise certainly. |
| A | Asi ne, mně se to docela líbí. | Probably yeah, I like it pretty much. |
| B | Dobře, padesát sekund… | Ok, fifty seconds … |
| A | Jo. | Yeah. |
| B | Budem ještě něco měnit? Takže… mně taky. | Should we change anything? So … me too. |
| A | Já taky… á souhlasím s překladem, já to tady mám taky. | Me too … and I agree with the translation, I got this here, too. |
| B | Dávám souhlasit s překladem. | I click agree with the translation. |
| **ID 14** | Nevypadá, že jeho žena ..., …, měli bychom jít k řediteli. | 'It does not look as if his wife ..., …, we should go to the headmaster. |
| B | Nje... … vid... … co? Njevidziavam, že jego žona pokazuje, co? | [reading] what? [reading] what? |
| A | Reka. | The hero. |
| B | Reka, žebysmy pošli do rektora. Cože? | [reading] what? |
| A | Nevypadá, třeba? | It doesn't look like, maybe? |
| B | Že jeho žena... | That his wife … |
| A | Nevypadá, že jeho žena... | It doesn't look like his wife … |
| B | Ježiš, že bysme zašli, že by, že by šli pro doktora? | Jesus, that we should go see, that, that they should call the doctor? |
| A | Doktora, myslíš? | The doctor, do you think so? |
| B | Nevypadá, že jeho žena pokazuje reka. | It doesn't look like his wife is [reading: showing the hero.] |
| A | Nemusíš to psát rovnou? | Don't you have to write it down straight away? |
| B | Takže ne, nevypadá? | So, it doesn't, doesn't look like? |
| A | Mhm, když, tak to upravíme. | Mhm, or else we can still correct it. |
| B | Nevypadá... | It doesn't look like … |
| A | Ty píšeš i háčky? | Are you writing with diacritics? |
| B | ...že jeho, mhm, žena něco... | ... That his, mhm, wife something ... |
| A | A co třeba něco, že poukazuje, řeka, jako jestli to není v tom smyslu, že něco říká. Poukazuje, říká... … To je blbost. Víš, jak to myslím? | And what if something, that she [reading], like if it's not like in the sense that she's saying something. [reading], she says … That's nonsense. You know what I mean? |
| B | Aha, je to oddělené čárkama. Víš, že je to, že jeho žena poukazuje reka, čárka, že bysme pošli do rektora. | Aha, this is separated by commas. You know that this is that his wife is pointing at the hero, comma, that we should [reading]. |
| A | No, to jo, ale nevypadá a teď to myslím, to poukazuje reka. | Well, yeah, but it doesn't look like and now I mean, that pointing at the hero. |
| B | No... | Yeah … |
| A | Tak jestli to není jako něco říká. Tak tam máš stejně jako čárku před že. | If that isn't like she's saying something. Anyways, there is a comma before [že]. |
| B | Jo, jo, jo, jo, já jsem myslela, že myslíš, že to navazuje na to, že bysme... | Yeah, yeah, yeah, yeah, I was thinking that you think that this relates to that that we should … |
| A | Ne, ne. Že by měli jít k rektorovi. Do rektora. | No, no. That we should go to the rector. [reading] |
| B | Rektor, ředitel nebo něco, ne? Jako, že i u nás, ne, že je rektor jako? | Rector, headmaster or something, isn´t it? Like, also in our country, is there a rector? |
| A | To nevím. | That I don't know. |
| B | Aha. | Aha. |
| A | Ježiš, Maria... | Jesus, Maria. |
| B | Že si spíš myslím, že je to něco, jako že je... Jako že mi přijde, že nějaká prostě špatná, že mi to furt evokuje takovou naléhavost a... | I rather think that this is something … like I think that there is something wrong, it still evokes some urgency to me and … |

(*continued*)

| | | |
|---|---|---|
| A | Mně zas ta druhá půlka věty, jako že bysme měli něco. Jestli jít k řediteli třeba. | As for me, the second half of the sentence is like that we should do something. Maybe go to the headmaster. |
| B | Tak měli bysme jít k řediteli, za tu čárku? | So, we should go to the headmaster after the comma? |
| A | Třeba. | For example. |
| B | Tak. A že jeho žena pokazuje... | Ok. And that his wife is [reading]. |
| A | Že mu žena přikazuje? | That his wife is telling him to? |
| B | Ale co mu přikazuje? | But what is she telling him to? |
| [...][…] | | |
| ID 15 | Nevidím, že jeho žena ukazuje na chlapce, aby šel k řediteli. | I do not see that his wife is pointing at the boy that he should go to the headmaster. |
| A | Jo. | |
| A | To první bude ne. | |
| A | Nevidziela, tak to bude jako, že neříkám, že jeho žena… no, zona. | |
| A | Hej, hej, tak tam bude, ne… něco a pak to další je, že jeho žena, ne? Určitě. | |
| A | A tak pak tam je, že bysme něco poslali do rektora. Hm, no, no, ale… No ale šli jsme do rektora. | So there it says that we should send something to the rector. Hm, yeah, yeah, but … but we went into the rector. |
| A | Nebude? | Not? |
| A | Nevidzjilam, hm. Pokazuje, to bude jakože ukazuje, poukazuje. Máš tam chybu, za tím ne je otazník. Aha, okej. | [Nevidzjilam], hm. [Pokazuje] that's like she's showing, pointing at. You got a mistake there, you got a question mark after the ne. Aha, ok. |
| A | Ne… | Ne… |
| B | **No, to nedává smysl.** | **Well, that doesn't make any sense.** |
| A | No právě, nevidzjelam. Mhm, tak to je úplně konec asi. | Exactly, [nevidzjelam]. Mhm, so that's it, absolutely. |
| A | Že jeho… pokazuje. | That his … [pokazuje]. |
| A | Mhm, tak, hm, něco tam napiš. | |
| A | Že jeho… no jako to zona bude podle mě určitě žena. Ještě jak tam máš tu, tu tečku nad tím. | That his … well, this zona as for me is surely wife. Surely because there is this, this dot on top. |
| A | Neříkám… | |
| A | Hej, dvě minuty a půl skoro. Tak to bude, že neříkám, že jeho žena pokazuje… přikazuje. No to bude možná přikazuje. Přikazuje rekovi. | Hey, almost two and a half minutes. So that's I'm not saying that his wife is [pokazuje] … instructing. Well, that's probably instructing. Instructing the hero. |
| B | No, to ukazuje, bych dal ukazuje. | |
| A | Hej, no to by možná… jo, tak to bude ukazuje. | |
| A | Hm, jakože neříkám, že jeho žena ukazuje, abysme šli k rektorovi, k řediteli. No, tak ale už to trošku dává smysl, neříkám, že jeho žena… Hm, přikazuje… to je… možná přikazuje. | |
| B | Vidím, že jeho žena ukazuje… | |
| A | Hm, jo. To bychom tam máš na konci n. | |
| A | Super. | |
| A | …že ukazuje… sakra. A to, nebude to nějakej člověk, jakože aby, že, že jeho žena ukazuje někomu, aby šel k řediteli. Jako, že… no chlape… chlapovi. Že jeho žena ukazuje chlapovi, aby šel k řediteli. Super. | |
| A | Souhlasím s překladem. | |
| A | To je jedno, hej teďka budu psát já, to bude zase konec. | |
| A | No, deset vteřin. | |
| A | Hm, čtyři. Mhm. | |
| A | Hm. | |
| ID 16 | Nevěděla jsem, ze jeho žena navrhuje, abychom šli za učitelem. | I did not know that his wife is suggesting that we should go see the teacher. |
| A | Ně vidzalam, že jego zona pokazuje reka, žebysmi… | [reading] |
| B | Tak ně vidzialam, ne- neviděla? Neviděla nebo nevěděla? Nevěděla jsem, že jeho žena… | So, [reading] I didn't see? Didn't see or didn't know? I didn't know that his wife … |
| A | Že jeho žena… | That his wife … |
| B | Takže… se někdo asi směje… nevěděla jsem… | Ok … somebody is laughing … I didn't know … |
| A | Jo, to je moje žena. | Ah, this is my wife. |
| B | .....že jeho, že jeho žena. | … that his wife, that his wife. |
| A | Máš tam otazníky? | Do you have question marks there? |
| B | Emm, jo. Akorát to bude teda bez interpunkce, ne, bez diakritiky… | Emm, yeah. But this is going to be without interpunction, no, without diacritics … |
| A | Jó, to nevadí. | Ok, no problem. |
| B | Neviděla jsem, že jeho žena pokazuje reka? | I didn't see that his wife [reading]? |
| A | Jó, to vůbec nevím, co je. | Oh, I have no idea at all what that is. |
| B | Ty jo. | Man. |
| A | Pokazuje reka. | [reading] |
| B | .....pošli do rektora… Pokazuje reka… | … [reading] … |

(*continued*)

| | | |
|---|---|---|
| A | Tady máme spolubydlícího, ten mi může poradit, ten byl teďka v Anglii. Ten se naučil polsky. | We got a flatmate here, he can give some advice, he's been to England and learned Polish there. |
| B | Tak to já jsem se ještě polsky nenaučila… | Well I haven't learned Polish yet. |
| A | Pokazuje reka, ty jo, co to může být. | [reading] man, what could that be? |
| B | Pokazuje, tak, no, bude to ukazuje? Pokazuje… | [reading], well, that is she is showing? [reading] … |
| A | Žebysmi pošli do rektora. Že bysme šli, ale teďka co je rektor, že jo. | [reading] that we should go, but now what is rector, right? |
| B | Hm… | Hm… |
| A | Že to asi nebude jako rektor na univerzitě, podle mě. | **That's not going to be a university rector, as for me.** |
| B | To je nějaký jako, no… Jako učitel? Třeba teďka zase z… nevím, třeba jako mentor je učitel, tak rektor by taky mohl… | That's something like, well … like a teacher? Maybe like, now … I don't know, like, a mentor is a teacher, so rector could also be … |
| A | No jasně, no. | Yeah, sure, yeah. |
| B | …v polštině. Tak, já nevím. Nevěděla jsem, že jeho žena ukazuje reka, ty jo, to vůbec netuším. | … in Polish. Well, I don't know. I didn't know that his wife is presenting the hero, man, I have no idea. |
| A | Ukazuje reka… Tak tam napíšem, že to nevíme. No nebo prostě na to asi nepřijdem, že, když to nevíme. Já nevím jako, mně to nic jako… | Is showing the hero … So we'll write down that we don't know. Or simply we don't have a clue, right, if we don't know that. I don't know like, to me that doesn't … |
| B | No, necháme tam tu první část té věty a pak jako to vytečkujeme nebo jak to... | Well, let's leave the first part there and then we can put dots there or like that … |
| A | Jó, jó, vytečkuj ty dvě, že jeho žena. Vytečkuj ty dvě slova, jak kdyby… Napíšeme… | Yeah, yeah, put some dots there that his wife. Put dots there instead of the two words as if … we will write … |
| B | Tečka, tečka, tečka… že bychom šli… | Dot, dot, dot … that we should go … |
| A | Pošli? Jako že by… jó, tak a to je asi, že bysme tam šli, no. | [reading - could be CS they passed away or imperative send sth.] That's like … ah, so that's probably that we should go there, yeah. |
| B | Že bychom šli, no. Tak pošli, že bychom pošli to snad nebude. Že bychom šli a do, no. | That we should go, yeah. We passed away, that's not going to be that we passed away. That we should go and into, yeah. |
| A | Do rektora… tak to vůbec nevím. | [reading CS Into the rektor] that I absolutely don't know. |
| B | Tak to bude v na... … že jo, po bude za. | So that's in, on … right? That is to. |
| A | Jako do je za. | Like, into is to? |
| B | Jako za někým, bych si tipla. | Like to somebody I would say. |
| A | Jo, jo, jo, jo. | Yeah, yeah, yeah, yeah. |
| B | Že bychom šli za… tak já nevím, dáme tam fakt něco, tak střelíme nebo to fakt necháme spíš takhle? | That we should go to … well, I don't know, let's put something there, either we guess it or we just leave it like that? |
| A | Bylo tam napsaný, počkej, v těch instrukcích je napsaný, že si to máme… vyvoďte si ho z kontextu nebo hádejte. Aha, tak hádejte… | It said, wait, in the instruction it says that we should … infer it from the context or guess. Aha, guess … |
| B | No... | Well … |
| A | Tak vlastně bysme měli hádat… Tak jako aby to dávalo smysl. Nevěděli js… nevěděl jsem, že jeho žena… | Well, actually we should guess … So that it makes sense. We didn't kno… I didn't know that his wife … |
| B | Jo, pokazuje reka. | Yeah, [reading which could mean: is showing the hero.] |
| A | Jako navrhuje, navrhuje… | Like suggesting, suggesting … |
| B | Navrhuje, dobře. Dáme navrhuje. | Is suggesting, ok. Let's put suggesting there. |
| A | Jakože, víš co, přikazuje, poukazuje, navrhuje, to by jako teoreticky… | Like, you know, she's ordering, pointing at, suggesting, that would theoretically … |
| B | Ano, navrhuje, že bychom šli za… | Yes, suggesting that we should go to … |
| A | Že bychom šli… a není to abychom? | That we should go … and isn't is that [correcting subordinate conjunction]? |
| B | Abychom šli… jo, navrhuje, jo, jo… abychom. | That we should go … yes, suggesting that, yes, yes … that we should. |
| A | Pak to budem mít správně, ale to vůbec neznamená, že té polštině rozumíme, že jo. To je by chance jenom. | Then this should be correct, but that doesn't mean that we understand Polish, right? That's just by chance [speaking English]. |
| B | Abychom šli za… | That we should go to … |
| A | Za učitelem. | To the teacher. |
| B | Jo, tak dáme učitele? | Yeah, so should we put the teacher there? |
| A | Jo, jo… dej učitele, dej učitele. | Yeah, yeah … put the teacher there, put the teacher there. |
| B | Za učitelem, no, já myslím, že nic lepšího už nevymyslíme. Tak jo, takže souhlasíme? | To the teacher, well, I think that we won't come up with anything better. Ok, so do we agree? |
| A | Jasný… | Sure … |
| B | Dobře, souhlasíme. Ježiš, to je… | OK, we agree. Jesus, this is … |
| A | To je zajímavý a už to začíná, jako ten první byl takovej lehčí, ale teďka už jako přituhuje. | That's interesting and it's about to start, like, the first one was a bit easier, but now it's getting harder. |
| B | Ano, tam se to dalo hodně odvodit od toho kontextu, že jo. Tady, vlastně jenom u jedný věty, už to tak jednoduchý není. | Yes, there it was possible to infer a lot from the context, right. Here, actually just with this one sentence, it's not that easy anymore. |

Table A.2
Transcripts of think-aloud protocols for the stimulus sentence (xxv). *Gdyby nie było książek, czytałbym Ci z oczu.* 'If there were no books, I would read from your eyes'. The written translations as provided by a particular pair of participants are marked with a grey background colour right next to the participants' ID. The original Czech solution is given in the left and the English translation in the right column. The original CS transcript (left column) is translated into EN (right column). Speculations about the plausibility of words in context and other relevant passages are marked bold. The EN translation is not complete for reasons of effort and relevance. If you wish to add a translation or suggest corrections, please contact the authors at kjagrova@coli.uni-saarland.de.

| | | |
|---|---|---|
| **ID 3** | [no answer written down] | |
| A | Jak bys to přečet? Já bych to přečet asi jakože, kdyby nje- by- lo, lo ks, ksiašek, čital bych či z oču. Asi že, víš co? Tak pokud ty kšiašky nebo co. . . jsou, by mohly být knížky, myslíš, že to je? | How would you read that? I would probably read that like if [reading]. Like, you know, if these [kšiašky] or what . . . are, could be books, do you think, that's . . .? |
| B | Mohlo by to být, ale zatím bych to nechal stranou. Může to být v podstatě cokoli. Začněme tím, co víme. Takže kdy. . . kdyby ně bol, bylo. . . Tak to. . . | Could be, but I would leave that aside for now. That could basically be anything. Let's start with what we know. So, if . . . if [reading] . . . then . . . |
| A | No, tak to je určitě, že kdyby nebylo. | Well, that must, if there were no . . . |
| B | Ano. A tedy z oču, tak, tak to vypadá jako, jako z očí. | Yes. And then [oczu], so, so that looks like, like out of the eyes. |
| A | Jako z očí. | Like out of the eyes. |
| B | Ano. | Yes. |
| A | Četl bych ti z očí. | I would read from your eyes. |
| B | Mhm, takže kdyby nebylo knížek, četl bych ti z očí. Já, já teda nevím. | [correct solution found at this point: " Kdyby nebylo knížek, četl bych ti z očí." [If there were no books, I would read from your eyes.]] I, I don't know. |
| A | Ale to nedává žádnej smysl, nemyslíš? | But that doesn't make any sense, don't you think? |
| B | To nedává žádnej smysl, ale. . . co by to ještě mohlo být? Kšiažek nebo ksia. . . | That doesn't make any sense, but . . . what else could it be? [reading] or [reading] ... |
| A | No i ty oči ani nevíme, že jsou oči. | Well, even about these eyes we don't even know that these are eyes. |
| B | Jak by se asi řekl otec? Jako, já nevím, může to být. . . no, ale to by dávalo ještě menší smysl. Ale rozumíš, že by to mohl být nějakej tvar ot. . . | What would be father? Like, I don't know, this could be . . . well, that would make even less sense. But you see that this could be some form of fath . . . |
| A | Víš, co je, víš co je taky zajímavé? Že to Ci je velkým písmenem. | You know what's, you know what's interesting, too? That Ci has a capital letter. |
| B | Může, může to být jako Ti, jo jako, že velké t při ot. . . | That could, that could be like Ti [You], ah ok, with capital T with fath ... |
| A | Jo, jo, jo, jakože velké zájmeno, hm. | Yeah, yeah, yeah, like a capitalized pronoun, hm. |
| B | Tak, že by to dávalo smysl. Kdyby nebylo kšažek, čital bych, by či z oču. | So that would make sense. If there were no [reading], I would read from [reading]. |
| A | Anebo. . . | Or . . . |
| B | Mhm? | Mhm? |
| A | Nebo prostě kdyby nebyla nějaká překážka, tak bych ti viděl přímo do očí. | Simply if there wasn't this obstacle, I would look directly into your eyes. |
| B | To by mohlo být spíš. Nicméně je to velký ti. . . velký ti, ale. . . | That sounds better. However there is this capital Ti . . . capital Ti, but ... |
| A | No protože, jakože mnohem větší tip je podle mě, že ty kšiažky jsou knížky. | Well, because, a much bigger tipp as for me is that these kšiažky [reading] are books. |
| B | Mhm, máš pravdu. | Mhm, you're right. |
| A | Protože. . . nějaké fonetické podobnosti a ale vyznámově to nedává žádný smysl. | Because . . . there are some phonetic details and but meaning-wise it doesn't make any sense. |
| B | Mhm, ano, ano s tím, s tím souhlasím. A můžeme říct vždycky ještě, že jsme to mysleli metaforicky ty překážky. Ale obecně asi půjde o nějaký problém, no. Tak nebýt překážek, čet, četl bych ti z očí. Hm. | Mhm, yes, yes with that I, I agree with that. And we can always say that we meant these obstacles metaphorically. But in general this is probably about some problem, right. So, if there weren't these obstacles, I would read, read from your eyes. Hm. |
| A | No ona jako ta psaná polština je mnohem těžší na rozumění, než, než když člověk ví, jak to zní. | Well, written Polish is much more difficult to understand than, than if one knows how it sounds. |
| B | Je to asi nejlepší, co zatím máme, co ty na to? | That's probably the best we have so far, what do you think? |
| A | Dobře, dobře, kdo z nás dvou píše? | Ok, ok, who of us is writing? |
| B | Vypršel nám čas. Dobře ale ten, kdo to bude poslouchat. . . | We ran out of time. Ok, but then, the one who will listen to this . . . |
| **ID 7** | Kdyby nebylo knih, sešel by z očí. | If there were no books, my father would read to me. |
| A | Kdyby ňje. . . ňjebylo kniažek, czytalbym Ci z ocsa, hm. Kdyby nebylo knížek, cz. . . | |
| B | Četl bys. . . | |
| A | Tak piš. | |
| B | . . .si z oca. Nevím, co je oca. | |
| A | Otec. Kdyby. . . napiš. Czytalbym, czytalbym Ci z otca. Kdyby nebylo. . . knížek. . . czytalbym. . . | |
| B | Spíš četl by sis. . . | |
| A | Četl by mi otec, asi. Ci z oczu. . . Kdyby ňjebylo, czytalbym sis. . . asi jo, souhlasim, jo? | |
| B | Hm. | |
| **ID 13** | Kdyby nebylo knih, sešel by z očí. | If there were no books, he would get out of sight. |
| B | Kdyby nebylo knížek. . . | |
| A | Ci z ocu. . . četli bysme. . . | |
| B | Četli. . . | |
| A | To je četli, podle mě, četli bys. . . | |
| B | No, já bych taky řekla, ale nevím, co je to poslední slovo. | |
| A | Os, otsu. . . mozku. | |
| B | Z mozku? | |

(*continued*)

| | | |
|---|---|---|
| A | Četli by. . . | |
| B | Kdyby nebylo knížek, četli by jsme. . . četli bysme si z očí. | |
| A | No. Škoda, že nemáme i kameru. | |
| B | Ty jo. . . | |
| A | Četli bysme z o. . . to by mohlo být to očí. Četli bysme z očí, jakoby někomu z očí. | |
| B | Asi jo, jako, vidím ti to na očích. | |
| A | Jo, to by mohlo být. | |
| B | Myslíš? | |
| A | No, napiš to tam. | |
| B | Tak četli by jsme druhým z očí. | |
| A | Jo. Máš tam oší. | |
| B | Ne, mám tam očí. | |
| A | Tak souhlasit? | |
| B | Mhm. | |
| A | Ty máš vždycky takové jednoduché a já mám takové těžké. | |
| B | Vždyť je děláme společně. | |
| A | Mně hrozně padaj ty sluchátka. | |
| **ID 14** | Kdyby nebylo knih, sešel by z očí. | If there were no books, we would get out of sight. |
| B | Kdyby ně bylo kšiažek, čitalbym Či z oču... . . . čitalbym Ci z oču? | [B trying to read the stimulus] |
| A | Kdyby něčeho nebylo, tak by to schytal. | If something wasnt there, he would be punished for it. |
| B | Hm, jo, to je možné, no. Čitalbym Ci z oču. | Well, yeah, that's possible, well. [reading] |
| A | To je divný, že tam je to velký C, viď? | It's weird that there is a capital C, isn't it? |
| B | Mhm, fmf. Kdyby nebylo, fmf. | Mhm, fmf. If there was no, fmf. |
| B | No... . . . Kdyby nebylo... . . . hm, či já vim. Schy... . . . jak jsi říkala, schytal? | [Well . . . If there was no . . . hm, who knows. [reading] . . . how did you say, get punished for it? |
| A | Schytal by. To je takový to akorát připomíná, že jo, nějakou výslovnost. | He would get punished. That's like, it just reminds me of some pronunciation, right? |
| B | Z oču... . . . tak z očí. | [reading] . . . so, from the eyes? |
| A | Sešel by z očí, ne? | He would get out of sight, or not? |
| B | Aha, že by? | Aha, could be. |
| A | Ty jo, nevím. Třeba knížek? | Man, I don't know. How about books? |
| B | Aha, že by? To by dávalo smysl. | **Aha, could it be? That would make sense.** |
| B | Kdyby nebylo knih, sešel by z očí? | If there were no books, he would get out of sight? |
| A | Mhm. | Mhm. |
| B | **To by i dávalo smysl, že?** | **That would make sense, wouldn't it?** |
| A | Právě. | Sure. |
| B | A sešel by nebo sešel bys? | And he would get out of sight or you would? |
| A | Já myslím, že takhle to je dobrý. | I think it's alright like that. |
| B | Sešel by. | He would get out of sight. |
| A | Hm. | Hm. |
| B | No, souhlas. | Yeah, agree. |
| ID 15 | Kdyby nebylo slov, četl by mi z očí. | If there were no words, he would read from my eyes. |
| A | Kdyby nebylo knížek. . . | If there were no books . . . |
| B | Četl bych ci z ocu.(?) | |
| A | Četl by mi z očí. | He would read from my eyes. |
| A | Kdyby nebylo. . . no ocu budou oči. Cz je č. | If there were no . . . well, ocu is eyes. Cz is č. |
| B | No, to jo, no. | Yeah, true, well. |
| A | Kdyby nebylo ksizek. . . | If there were no [ksizek] . . . |
| A | Hm. | Hm. |
| A | To je jedno, tak to dopiš. Kdyby nebylo ksi. . . ksiazek, hm. | Doesn't matter, just continue writing. If there were no [ksi . . . ksiazek], hm. |
| A | No, knížky, to asi nebude ono. Ks. . . žek. | Well, books, that's probably not it. [Ks . . . žek.] |
| A | Hm, kdyby nebylo. . . no. Kdyby nebylo čeho? | |
| A | Kdyby nebylo. . . ty jo, co to znamená? | |
| A | Hm. | |
| A | Ksiazek. . . hm. Kdyby nebylo. . . no, ty jo. Četl by z očí. Hm. | |
| A | To fakt nevím. Dvě minuty, hm. | |
| A | Hm, tak třeba to bude, viď. No, takhle bych to nechala. Souhlasím s překladem. | |
| A | Fakt. | |
| A | Nemůžu najít ten kurzor, takže asi ne. . . jo, dobrý. | |
| A | No, tak přijdeme na to, co je ksizek? Asi ne, no. Souhlasím s překladem. | |
| A | Jako dává to logicky smysl, jakože víš co, ty slova. Nemusí bejt úplně jakoby podobný těm našem. | Like, it makes some logical sense, like, you know, these words. They don't have to be totally similar to ours. |
| A | Hm. | |
| A | No některý, ale některý taky nebudou. | |
| A | No jako jo, ale tak Slováci maj taky plno slov, který vůbec příbuzný češtině nejsou. Jakože většina jo, ale. | |
| **ID 16** | Kdyby nebylo knížek, četl bysem si z očí. | If there were no books, I would read from my/people's eyes. |
| B | Tak. . . Kdyby ně bylo kšiasek, čital-bym či z oču. Tak kdyby nebylo něčeho. . . | So . . . [reading] So, if there was something missing . . . |
| A | Knížek, knížek. | Books, books. |

(*continued*)

| | | |
|---|---|---|
| B | Knížek, aha, knížek. . . četl by jsem. . . Kdyby nebylo knížek. . . | Books, aha, books . . . I would read . . . If there were no books . . . |
| A | Četl bysem z očí? | I would read from the eyes? |
| B | Četl by. . . to je jako takový. . . | I would read . . . that's so . . . |
| A | Jo, jo, jo. . . | Yeah, yeah, yeah . . . |
| B | Moudro, jo. Četl. . . čítal bysem, jo. . . četl bys... . . . jo, já myslím, že jo. . . bysem z očí. | Some wisdom, yeah. I would read . . . would read, yeah . . . you would read . . . yeah, I think, that yes . . . I would read from the eyes. |
| A | Jo. | Yeah. |
| B | Jo, četl bysem. . . | Yeah, I would read . . . |
| A | Jo, já souhlasím. | Yeah, I agree. |
| B | Tak. . . | Ok . . . |
| A | Tady není jako moc co vymýšlet. Kdyby ně bylo ksiažek, čítal bysem z. . . | There is not much to make up. [reading] |
| B | Kdyby. . . jasně. Knížky a čítal, to dává smysl, takže kšiažek budou. . . | If . . . sure. Books and was reading, that makes sense, so [kšiažek] is . . . |
| A | A oči, jako taky, asi nic jinýho jako asi tam. . . | And eyes, same over here, probably nothing else ... |
| B | Kdyby nebylo knížek, četl bysem z očí. Jo. . . takže tak, jo? | If there were no books, I would read from the eyes. Yeah . . . so, like that, yeah? |
| A | Jo. | Yeah. |
| B | Jo? | Yeah? |
| A | Jo, jo, jo. | Yeah, yeah, yeah. |
| B | Ještě chvilku. Ještě máme chvilku, tak možná. . . ještě chvilku to. . . | Just a moment. We got some time left, so maybe . . . just a moment . . . |
| A | A co je tam to Cé í, jako? Čitalbym Cé í z očí, že tam vlastně jako kdyby nemáme jedno slovo. Jestli to Ci je, hraje nějakou roli a je to ještě velkým. | And what is this C I there, like? [reading] like we don't have one word. If this [Ci] is, plays some role and also has a capital C. |
| B | Čitylbym je. . . | [reading] is |
| A | Je to velkým. Ale. . . | It's with a capital . . . but . . . |
| B | Ano. . . | Yes . . . |
| A | A nevím, proč je to velkým. | And I don't know why it has a capital letter. |
| B | Hm, kdyby ne, ně bylo ksiožek, čitalbym Ci z očí. To zní jak nějaký zájmeno, ty jo, kdyby to bylo. | Hm, if there were no, [reading]. That sounds like some pronoun, man, if that was . . . |
| A | Četl by jsem si z očí? | I would read from my eyes? |
| B | Četl by jsem si z očí? Hm, ano, ano. . . já myslím, no. . . to, to pořadí slov ve větě bude podobný jako, jako v češtině. | I would read from my eyes? Hm, yes, yes . . . I think, well . . . that, that word order in the sentence is similar to that in Czech. |
| A | No jasně, no. | Sure, yeah. |
| B | Čitalbym, četl bysem si z očí. | [reading], I would read from my eyes. |
| A | Jo a bysem je dohromady. | Yeah and that [bysem] is one word. |
| B | Já jsem se na to ptala, jsem si nebyla jistá. | I asked, I wasn't sure. |
| A | Jo, jo. | Yeah, yeah. |
| B | Četl bysem si z očí. | I would read from my eyes. |
| A | Bacha, přehazuje se ti to, máš tam nastavený automatický opravy, že jo. | Watch out, your spellchecker is on, right. |
| B | Ano, ano. . . bysem. . . | Yes, yes . . . I would . . . |
| A | Nebo napiš bych si a. . . | Or write down I would . . . |
| B | Bysem, bych si z očí. | I would, I would from my eyes. |
| A | Jo. . . | Yeah . . . |
| B | Jo, souhlasíme? | Yeah, do we agree? |
| A | Jo, já už mám. Tak dál. Tak to bylo takový celkem jednoduchý, ty jo. | Yeah, I've already clicked. Let's go on. So that was pretty easy overall, man. |
| B | Nó, ono, no. . . no, se to tak zdálo alespoň. Ono to může být úplně jinak ve výsledku. . . neboli úplně blbě. | Well, that, well, . . . well, at least it seemed so. It could be totally different as a result . . . or totally wrong. |

Table A.3
Transcript of the think-aloud protocols for the stimulus sentence (xxvi). *Kupiliśmy nie tylko czerstwy chleb, ale jeszcze gorzej – też stary żółty samochód.* 'Not only did we buy stale bread, but even worse – also an old yellow car.' The written translations as provided by a particular pair of participants are marked with a grey background colour right next to the participants' ID. The original Czech solution is given in the left and the English translation in the right column. The original CS transcript (left column) is translated into EN (right column). Speculations about the plausibility of words in context and other relevant passages are marked bold. The EN translation is not complete for reasons of effort and relevance. If you wish to add a translation or suggest corrections, please contact the authors at kjagrova@coli.uni-saarland.de.

| **ID 1** | Nejen, že jsem nekoupili česrtvý chléb, ale ještě hůře - také staré auto | Not only did we not buy fresh bread, but even worse – also an old car |
|---|---|---|
| A | [...] […] Nekoupili jsme tak čerstvej chleba, ale ještě je dobrej, ten starej žolty samochód. Hm, starej se sám... … sám. Takže bych to viděl tak, že jako nekoupili jsme... … nekoupili jsme, nekoupili jsme příliš... … příliš starý chleba, ne, příliš čerstvý chleba... … eště čerstvý chléb, ale... … ještě gorzej... … ale ještě horší? | We did not buy such fresh bread, but it´s still good, this old yellow self-goer. Hm, care for yourself … yourself. So, I would suggest that kind of we did not buy … we did not buy, we did not buy too … too old bread, no, too fresh bread … still fresh bread, but … even worse … but even worse? |
| B | Ale ještě hůř... | But even worse |
| A | Ale, ale... | But, but ... |
| B | **To je, todle je zvláštní...** | **That´s, that´s weird ...** |
| A | No..… | Yepp. |
| B | Já bych řekl, že to je něco jako koupili jsme ani ne, buďto nejen, anebo ani ne tak čerstvý chléb... … chléb, ale hůř, starý žlutý automobil. | I would say that this is something like we bought not so, either not only or not such fresh bread ... … bread, but worse, an old yellow car. |
| A | Hm, no. | Hm, yeah. |

(*continued*)

Table A.3 *(Continued)*

| | | |
|---|---|---|
| B | Tak samochód je auto, pokud se nepletu. Žlutý, žlotý... | Well, samochód is a car, if I´m not mistaken. Yellow, žlotý... |
| A | Jak jsto říkal, prosím tě? Koupili jsme ten... | How did you say? We bought this … |
| B | Koupili jsme ani ne tak čerstvý chléb... nje tylko čerstvy chléb... ale, ale ještě hůř. | We bought a not so fresh bread … nje tylko čerstvy chléb … but, but even worse. |
| A | Ale ještě hůře... | But even worse … |
| B | Jo, jako ve smyslu nejen, že jsme nekoupili chleba, ale ještě hůř... | Yep, like in the sense of not only did we buy bread, but even worse … |
| A | No... Ale ještě hůře. Ten starý, tež... | Yeah … but even worse. This old, also … |
| B | Staré žluté auto. | An old yellow car. |
| A | No, to vůbec nemá smysl. Koupili jsme ani ne tak čerstvý... | Well, that makes no sense at all. We bought a not so fresh … |
| B | Pošleš, pošleš pitomce koupit chleba a von přitáhne auto, tak moc je pitomej. ...nje tylko čerstvý chleb, ale ješče gorzej. To je divná věta anebo... gorzej není jako hůř... ale ještě gorzej... jako ještě horký, ale to asi ne. Jakože ještě horký... Hm, to tam nedává... | You send, you send an idiot out to buy bread and he brings a car, so stupid is he. … Not only fresh bread, but even worse. That's a weird sentence or … gorzej is not worse … but even worse … like still hot, but probably not. Like still hot … hm, that does not … |
| A | **To je blbost...** | **That's nonsense** |
| B | Jakože je tak čerstvej, že je ještě horkej, ale to tam fakt nesedí to auto. | Like not so fresh, but still hot, but that car really doesn't fit there. |
| A | No. Anebo to bude, nekoupili jsme tak čerstvý chléb, ale ještě horší, taky starý auto, no. | Well. Or it will be: we didn't buy such fresh bread, but even worse, also an old car, well. |
| B | Možná jo. | Maybe yes. |
| A | **To njak nedává smysl.** | **That somehow doesn't make sense.** |
| B | Ono to možná... … to je ještě jako ve smyslu, nejen že vůbec nekoupil chleba, ale ještě koupil nějakej starej šrot. | Maybe this is … maybe like in the sense of not that he did not only buy bread at all, but also bought some old trash. |
| A | To by možná šlo, no. | That could be it. |
| B | .....smi ně tylko... … Nje tylko, nevím co je. Takže koupili jsme ne tak čerstvý chléb... … Anebo... … ne, ty jo. | [reading] I don't know what [nie tylko] is. So, we didn't buy such fresh bread … or … no, man. |
| A | No, tak v pohodě... | Ok, that's fine … |
| B | Nebo třeba... … nje tylko, jestli to neni jako něskoro slovensky, že to je jako malo chleba. Koupili jsme málo, málo čerstvýho chleba. Já nevím. | Or maybe … [reading] could be like něskoro in Slovak, like too little bread. We bought too little, too little fresh bread. I don't know. |
| A | | Hm. |
| B | Tak, tak třeba... Nejen že jsme nekou... že, že jsme nekoupili čerstvý chléb, ale hůře... | So, so maybe … Not only did we … that that we didn't buy fresh bread, but worse … |
| A | Hm, asi jo. Nevim. | Hm, probably yeah. I don't know. |
| B | .....ště hůře. | … ven worse |
| A | Také starý automobil? | Also an old automobile? |
| B | Také staré auto. Počkej anebo nebo to znamená ne tylko čerstvý jako nepříliš čerstvý, jako starý chléb. Nejen, že jsme nekoupili... … Zas já nevím, proč by bylo ale... … no, to je odporovací. | Also an old car. Wait, or does it mean not only fresh like not fresh enough, like stale bread? Not only did we buy … Again I don't know, why this should, ... … but … well, this is adversative. |
| A | No. | Yepp. |
| ID 8 | Koupili jsme ne tak starý chléb, ale ještě teplý - taky staré žluté auto. | We did not buy such old bread, but it was still warm – also an old yellow car. |
| A | Aha. | Aha. |
| B | Ku- kupilismi ňje tylko čerstvý chleb, ješče gořej… | [reading] |
| A | Gořčej podle mě. | Gořčej in my opinion. |
| B | Ne, to je, to je ř. Rž je ř, ale g se nečte jako ř, ne? Ne, g se čte jako h? | No, that is, that is ř. Rž is ř, but g is not read like ř, is it? No, g is read like h? |
| A | No, gořej prostě. | Well, gořej simply. |
| B | Gořej to je starý… | Gořej that is old … |
| A | Tež, to je ž, myslím. | Tež, that's a ž, I think. |
| B | Jo. | |
| A | Mhm. | |
| B | Tež starý žoltý samochód. | |
| A | Žon- žonltý, mysím. Ne, žontý. | |
| B | Žo- žo- žontý, žontý, okej. | |
| A | To… nebo počkat, ne. Žontý? Nevím. Nevím teďka. | |
| B | Já… ts… nepleť sem francouzštinu. | |
| A | Hej, tohle slovo zrovna vím, kámo, protože my jsme se jednou bavili to, s jedním Polákem a jakože jaké slova máme různé a on přímo říkal, že čerstvý znamená u nich… že čerstvý znamená u nich prostě opak. | Hey, I actually know this word, dude, because I was once talking to a Polish guy about which words are the same and which are different and he told me directly that czerstwy means … that czerstwy simply means the opposite in Polish. |
| B | Aha, fakt? | Aha, really? |
| A | Že čerstvý není čerstvý, ale starý, jo? | Like, čerstvý is not fresh, but old, right? |
| B | No tak to, to je docela dobrý. | |
| A | Takže, takže koupili jsme ne- jako ne tylko, tak to bude asi ne tak… | |
| B | Nebo ne to… ne tolik, ne tak starý chléb. | |
| A | …ne tak starý chléb. | |
| B | No. Takže kou… pili jsme… | |
| A | Ale ješče goř, to zní jako… | |
| B | …ne tak st… starý chléb… Gořej, gořej, co by mohlo bejt… | |
| A | Buď to bude horký… ješče bude ještě, prostě. | |
| B | No, jasně, mm, ale ještě… Ne tak starý a ještě horký? | |
| A | Ty jo, to zní divně. | |
| B | Ale jako, já, já bych taky řekl, že gořej bude něco jako horký nebo… jako víš co, je to od slova hořet, tak… | |

(*continued*)

Table A.3 *(Continued)*

| | | |
|---|---|---|
| A | No, asi. | |
| B | Též starý… žoltý, žol… ž. | |
| A | Samochod je auto, to je jasný. | |
| B | Samochod je… jo samochod je auto, aha, to mi nedošlo. | |
| A | Hm, no… Žoltý je asi žlutý, možná. | |
| B | No, jasně. | |
| A | Anebo… počkej žolty. | |
| B | No, jasně. T- t- takže, takže starý žlutý… taky starý žlutý auto? | |
| A | No. | |
| B | Ale ještě, ještě… | |
| A | Počkej ale nebo koupili jsme… | |
| B | Nemo- nemo- nemohlo by to bejt místo, místo… | |
| A | Anebo to souvisí s… ano. | |
| B | Počkej, že co souvisí? Nebo to? | |
| A | Že to ale ještě gořej souvisí s tím sa- samochodem, když je tam to… nebo nevím, teďka. | |
| B | Nemohlo by bejt, nemohlo by bejt gořej hoř- hořký? To nedává moc smysl, ale jako podle mě by mohlo bejt. | |
| A | Nemi… buď to bude… Nevím, proč je tam ta pomlčka teda potom, moc to nechápu. | |
| B | Jakože koupili jsme… koupili jsme starej chleba a taky žlutý auto. | |
| A | Asi jo. | |
| B | To je divný, ne to je divný. | |
| A | No, asi jo, dobre no, to asi nemá, no, dobre, tak… ale ještě… | |
| B | Jako dobře, já si… Jseš si jistej, že samochod je auto teda… | |
| A | Jo, jsem… | |
| B | A teďkon to slovo též. | |
| A | Kupilismi… | |
| B | Jsme si jistý, že je to též nebo že by, že je to taky? Jako jestli to není něco jinýho, že jo. | |
| A | To je, že určitě taky. | |
| B | Jo? | |
| A | Mhm. | |
| B | Takže ještě, ještě teplý. | |
| A | Nebo myslím teda. | |
| B | Takže jako asi se shodnem, takhle? Mm, ale, ale ještě teplý. Taky starý žlotý, žloté auto. | |
| A | Staré žloté auto. | |
| B | Sorry. | |
| A | Starý žluté… Počkej, počkej… ale… ne tak. Mmm, asi, no… | |
| B | Mm, jako… mně to nic jinýho neříká než tohle, ale jako smysl to moc nedává. | |
| A | Mhm. | |
| B | Ale, třeba, třeba koupili žlutý auto, starý, jo. Asi… jako, třeba jo. Asi, asi jako, souhlasím s překladem. | |
| A | Ale ještě gořej, tež starý žlutý auto. | |
| B | Jako já už tam asi nic jako ne to… | |
| A | Jo, asi, asi bych to tak viděl, no. Počkej, ještě, koupili jsme ne tak… jo, jo, dobrá, no. Nezdá se mi ta formulace věty, ale zřejmě to tak bude. Že, chápeš, neřekneš koupili jsme ne tak starý chléb, ale on byl ve skutečnosti ještě teplý, což prostě nedává… jako logicky nějak. No, to je jedno, okej, souhlasím. | |
| B | Ale tak jako, mělo by tam bejt jako, jako že… koupili jsme čerstvý chléb? | |
| A | Už nic nevymyslíme teďka, dobrý. | |
| B | Ne, tam je jako fakt ten spor, jakoby, že, že nebyl tak starý, ale ještě byl teplý. Jako je tam ten spor, jako. To ten spor tam je. Nebo jako něco minimálně v tom smyslu. Ale jako víc, víc toho asi ne to… ne- nevyplodím. | No, there is really this contradiction, like, that it was not that old, but still warm. Yeah, there is this contradiction, like. It's there. Or like something at least in the sense of this. But like we're not going to do more, we're not going to … bread another solution. |
| A | Jo, ano, no, jo, dobre. | |
| ID 10 | Koupili jsme nejen čerstvý chléb, ale ještě hůř - taky staré zlaté auto. | Not only did we buy fresh bread, but even worse – also an old golden car. |
| B | No, tak to už je jiná. Tak to bude čerstvý chléb. | |
| A | Takže… No a já bych řekla, že jako koupili jsme nejen… | |
| B | Jo, jo, koupili jsme nejen čerstvý chléb, ale ješče, ješče, ještě… | |
| A | Nejdéle… Tam na té klávesnici nemám ty háčky. | |
| B | Já je, já je tu taky nemám právě. Koupili jsme nejen čerstvý chléb, ale gorzej… | |
| A | Ale ještě a co by to tak mohlo být… | |
| B | To tez bude říež, teda taky. | |
| A | Mhm. | |
| B | Starý bude asi taky starý. | |
| A | Hm. | |
| B | Ž… žoltý… | |
| A | Tak… | |

(*continued*)

Table A.3 *(Continued)*

| | | |
|---|---|---|
| B | Žlutý? | |
| A | No… | |
| B | Zlatý, nevím. | |
| A | Hmm. | |
| B | Samochod. | |
| A | Hm… tak sakra, kde to je to í/ý? | |
| B | Mhm. | |
| A | Tak dáme žlutý. | |
| B | Jo, jo. | |
| A | Nebo zlatý? Já nevím, co by tam jako z toho šlo vymyslet. | |
| B | My nevíme, co to je to samochod. | |
| A | Samochod, hmm, tak to mě nenapadá nic. | |
| B | Jak, jakou oni používají měnu, ti Poláci? Neni zolty jako zlatý, fakt? | |
| A | No, tak jo, tak zkusíme zlatý. | |
| B | Ale samochod, ty jo. Koupili jsme nejen čerstvý chléb, ale ještě… taky starý… | |
| A | Co by to mohlo být? | What could that be; |
| B | A to a to gorzej? | And this [gorzej]? |
| A | No, tak to taky ne, no, nevím. No, že by to bylo hůř? Já bych zkusila hůř. | Well, also not, well, I don't know. Well, could that be worse? I would try worse. |
| B | Tak jo. | Ok. |
| A | Nic lepšího mě nenapadá. Zlotý, sakra, co by to mohlo být? Tak nevím, auto. | I have no better idea than that. [zlotý], damn, what could that be? Well, I don't know, car. |
| B | Dobře. | |
| A | Tak to, to přepíšu teda tady ještě. Tak, dobře. Odklikávám. | |
| B | Taktéž. | |
| **ID 14** | Nekoupili jsme dostatečně čerstvý chléb, ale ještě hůře - takový starý žluklý, zkažený. | We did not buy sufficiently fresh bread, but even worse – such an old, yellow, rotten one. |
| A | Kupilismy ně tylko čerstvý chleb, ale ješče gorzej... horkej, ne? Též stary žólty samochód. Nekoupili jsme takový čerstvý chléb... | [reading] … a hot one, right? Also an old [reading]. We didn't buy such fresh bread … |
| B | Mhm, mhm. | Mhm, mhm. |
| A | Takový? | Such? |
| B | Mhm... nebo tolko, jako že by tylko, ne takový, ale nekoupili jsme tolko čerstvý chléb... | Mhm … or [tolko], like [tylko], not such, but we didn't buy [tolko] fresh bread … |
| A | Jako že toliko, jako tolik. Nekoupili jsme tolik. | Like so much, like so much. We didn't buy so much. |
| B | Jako že ne ve smyslu tolik, ale že nebyl tak čerstvý. Že nekoupili jsme až tak čerstvý chleba. | Like, not in the sense of so much, but that it wasn't that fresh. Like, we didn't buy such really fresh bread. |
| A | Ale ješče gorzej. | [reading] |
| B | Ale ještě hůř, takový starý, žluklý... … samochod, no. Jak se to říká, že... … u nás se též říká, že už to chodí. Že je to, víš co, že je to tak staré, až už to chodí. | But even worse, such an old, rancid … self-goer, well. How do you say that … we use to say that it's already walking. That it's, you know, that it's so old that it started walking already. |
| A | Žluklý. | Rancid. |
| B | No, já nevím, jak to přeložit. Jako, že je to... | Well, I don't know how to translate that. Like it's … |
| A | Nekoupili jsme tolik čerstvý chléb... | We didn't buy such fresh bread … |
| B | Možná až tak čerstvý chléb. | Maybe not so really fresh bread. |
| A | Dostatečně? | Enough? |
| B | Nebo dostatečně čerstvý chléb. | Or fresh enough bread. |
| A | Jo? | Yeah? |
| B | To je ono. | That's it. |
| A | Ještě horší... | Even worse … |
| B | Ještě hůře. | Even worse … |
| A | Vůbec nevím, ha. Vidím tam totéž? | I don't know at all, ha. Do I see the same? |
| B | Takový. Víš co, takové, nekoupili jsme vůbec jako čerstvý chleba, ale ke všemu prostě ještě takový hnusný, žlutý, jo, sám chodí. | Such a ... … You know what, such, we didn't buy fresh bread at all, but above all simply also such a disgusting, yellow one, yeah, it walks on its own. |
| A | Starý. | An old one. |
| B | Já nevím jak, hm... | I don't know how, hm …. |
| A | Takový starý, tvrdý třeba? Pochodující. Takový starý. | Such an old, stale one maybe? A marching one. Such an old one. |
| B | Já bych řekla, že žluklý. | I would say it's rancid. |
| A | To úplně odhadujem, prostě. Jak může bejt chleba žluklej? | We're totally guessing, like. How can bread be rancid? |
| B | Jó? No tak jako, že se zapaří, a tak jako, že zesmrádne... Takový... | Yeah? Well like … if it starts sweating and like … starts smelling … such a ... |
| A | A to už jenom fakt jakoby hádáme. | And now we're really just guessing. |
| B | No a to samochod, to bych prostě řekla, že je to to, no, nevim jako. Prostě, až chodím. Já nevím, jak se, jak tomu říkat. | And that self-goer, I would simply say that this is this, well, I don't know, like. Simply, started walking. I don't know how, how to call it. |
| A | Kazící se prostě? Zkažený? | Simply decaying? Rotten? |
| B | Zkažený? | Rotten? |
| A | Mhm. | Mhm. |
| B | Takovou starou, žluklou zkaženinu. | Such an old, rancid rotten thing. |
| A | Já bych to nechala takhle. To už, stejně nevíme, jak to je správně. Ne? | I would leave it like that. We're, we don't know anyway how it's supposed to be, do we? |

(*continued*)

Table A.3 *(Continued)*

| | | |
|---|---|---|
| B | Hm. | Hm. |
| A | .....a je to zbytečný čtení. | … that's useless reading. |
| B | To je vtipné. Co tam dál máme? Ta polština je strašně hezký jazyk, takový vtipný. | That's hilarious. What else have we got? Polish is such a beautiful language, so funny. |
| A | To jo, mhm. | Oh yeah, mhm. |
| ID 15 | Nekoupili jsme pouze cerstvy chleb, ale jeste povoz - tez stare zlute kolo. | We did not buy only fresh bread, but also a vehicle – an old yellow bicycle. |
| A | Kupilismi… kupilismi ňje tolko cerstvy chleb, ale jesce gorzej – tez stary zolky samochod. Takže to bude koupili jsme… nekoupili jsme tak čerstvý chléb… | [reading] So that's we bought … we didn't buy such fresh bread … |
| A | Ne tolik. | Not so much. |
| A | Koupili jsme… | We bought … |
| B | Celý bych to přeložil, jakoby, že jsme nekoupili… | I would translated the whole thing, like, that we didn't buy. |
| A | Mhm. | Mhm. |
| A | Nekoupili jsme… Ale ještě gorzej. | We didn't buy … But even [gorzej]. |
| A | Nebude to nějakej sýr? | Isn't that some sort of cheese? |
| A | Já nevím, třeba si udělá sám. Ne, tak počkej. Ještě něco… taky starý zolty, může bejt co? Nebude to žlutý? Samochod bude jakože… Ale ještě… Tez bude též, ne? | I don't know, maybe he does it on his own. No, wait. One more thing … also an old [zolty], what can that be? Isn't that yellow? [Samochod] is like … but also … [tez] is also, right? |
| A | No, no, no nebo nějakej komentář k tomu. | Yeah, yeah, yeah or some kind of a comment to that. |
| A | Starý bude starý. | [Stary] is old. |
| A | Samorost. | [self-grower or] driftwood. |
| A | Zjoltý, žlutá to nebude. | [reading] that's not yellow. |
| A | Nebo ne? | Or not? |
| B | Já bych tam dal žlutý. | I would put yellow there. |
| A | Tak jo, takže starý žlutý samostroj. | Ok, so an old yellow [self-machine]. |
| B | Jestli tam nemůže bejt nějaký auto… | Could that be some car … |
| A | Áh, to bude kolo. Staré žluté kolo. | Ah, that's a bike. An old yellow bike. |
| A | A tak to, to je… na jazykovej… to nemá logickej původ. | Ah ok, that's … language-wise … that has no logical reason. |
| A | Podle mě to bude to kolo. | I think that's a bike. |
| A | No, nevím jako. Jestli to nebude to kolo. | I don't know, like. If that's not a bike. |
| A | Hej ale gorzej to podle mě… víš co, to je takový jako… jako nějakej vůz nebo tak. Ale ještě… Též staré žluté kolo… Koupili jsme chleba a kolo. Ty kráso… Ale ještě… hm… Ty jo, co to může bejt, goj… | Hey and [gorzej] as for me … you know, that's like … some vehicle or so. But also … also an old yellow bike … We bought bread and a bike. Oh man … and also … hm … man, what can that be, goj … |
| B | To může bejt něco, po čem se… | That could be something after which you … |
| A | Jako projímadlo? | Like a laxative? |
| B | No, třeba to bude nějakej druh sýru. To jsme říkali, jestli to není… | Well, maybe that's some sort of cheese. We said that maybe it's … |
| A | No, jakože, asi by bylo logický, aby to bylo jídlo, no. Když jdu koupit chleba, tak asi nekoupí kolo. Ale tak, máme minutu a půl, musíme tam něco napsat. | Well, it's logical that it might be food, well. If I go and buy bread, then he's not going to buy a bike. Anyway, we got 1 and a half minute, we have to write something down. |
| A | Hm. | Hm. |
| A | Hej tak já tam napíšu sýr. Ale ještě sýr, taky staré žluté kolo. Hej tak tam napíšu povoz. No, ale ještě povoz. No, jako nedává to moc smysl, ale… | Hey, so, I'm going to write cheese there. But also cheese, also an old yellow bike. Hey, so I'm going to write vehicle. Well, but also a vehicle. Well, dosn't make much sense, but … |
| A | Souhlasím. | |
| A | No, fakt souhlasím. Hm, teďka píšeš ty, co? | |
| A | Ne. | |
| A | No, to by mě zajímalo, co tady to fakt znamená. | |
| A | Mhm. Teď. | |
| A | A máš krátkou. | |
| **ID 16** | Nekoupili jsme jenom čerstvý chléb, ale ještě hůř - taky staré žluté auto. | Not only did we buy fresh bread, but even worse – an old yellow car. |
| B | Kupilismi ně tylko čerstvý chl… čerstvý chléb, ale ještě gorzej, též starý žoltý samochod. Tak samochod vím, že je auto. | [reading] I know that samochod means car. |
| A | Jó, to vím taky. | Yes, I know that, too. |
| B | Super. Žoltý bude asi žlutý, na konci teda – žlutý auto. Nebo, že jo, no, to nemusí být, ale. | Super. [reading] is probably yellow, at the end I mean, yellow car. Or, you know, well, it doesn't have to be, but … |
| A | **Ale pro čerstvý chléb, jako co to znamená?** | **But why fresh bread, like, what's that supposed to mean?** |
| B | **To jo, jakou to má souvislost s autem?** | **Yeah, how does that relate to a car?** |
| A | Koupili jsme… | We bought … |
| B | Tylko bude právě, ne? Ne, tam je, kupilis… kupilismi ně tylko. Tak to bude, no… | [reading] is right now, isn't it? No, there is [reading] … so that will be, well … |
| A | Ně tylko čerstvý chléb… ale ještě gorzej, ještě gorzej je ještě hůř? | [reading] even [gorzej] is even worse? |
| B | Ale ještě gorzej, ty jo, to vůbec ne- nemám, nemám ponětí, gorzej, co může, co může znamenat. | But even worse, man, I don't ha-have, have any idea, [gorzej], what that could, what that could mean. |
| A | Též starý žol- žolty samochód. Koupi… koupili… ně tylko… | [reading] |
| B | Tak koupili jsme, no… já… tylko. Já myslím, že tylko je právě, ne? Jako právě teď nebo… | So, we bought, well … I … [tylko]. I think that it's right now, isn't it? Like right now or … |
| A | To vůbec nevím, já vůbec nevím. Já rozumím fakt jenom čerstvý chléb, ješče a samochód a žoltý je prostě žlutý a starý je starý, že jo, takže. | I have no idea, I have no idea at all. I really understand just fresh bread, even and self-goer and [žoltý] is yellow and [stary] is old, right, so … |
| B | Ano, no… Tak to jsme na tom úplně stejně. | Yes, well … well, same over here. |

(*continued*)

Table A.3 *(Continued)*

| | | |
|---|---|---|
| A | Koupili jsme ně tylko, já nevím, co je to ně tylko, tak, tak, tak to zkus nějak tipnout nebo… | We bought [reading], I don't know what this [nie tylko] is, so, so, so try to gues somehow … |
| B | No tak mě napadlo teda to… koup… no, ne, právě, to moc nedává smysl, ale… | Well, it came to my mind that … bou … well, no, right now, that doesn't make much sense, but … |
| A | Co to je, opak, že jo, právě, že jo, jako? | What is it, the opposite, right, right now, right, like? |
| B | Anebo jestli třeba to, že za tím koupili vlastně je to ně, jestli to znamená právě, že nekoupili. Jako třeba ve slovenštině, kdy já som není, jako já nejsem, tak... … že to neguje… | Or if it's maybe that after this [we bought] there is this [nie], if it means namely that they we didn't buy. Like for instance in Slovak, [kdy já som není], like I'm not, so … that negates … |
| A | Jo, že to takhle funguje. | Oh, that it works like that? |
| B | …že až vlastně za tím slovesem, že jo. | … like only after the verb, right. |
| A | No, takže by to bylo jako: nekoupili jsme? | Well, so this should be like: we didn't buy? |
| B | Hm, já si myslím, že by to mohlo být nekoupili jsme čerstvý chléb. | Hm, I think that this could be we didn't buy fresh bread. |
| A | Nekoupili jsme… | We didn't buy … |
| B | Protože ve slovenštině… ve slovenštině se říká, já som není, je já nejsem. Takže jestli kupili sme ně, tak možná to bude nekoupili jsme. | Because in Slovak … in Slovak they say [já som není], that's I'm not. So if [kupili sme nie], that could be we didn't buy. |
| A | A to ještě gorzej je teda ještě hůř? | And even [gorzej] is even worse? |
| B | No, asi to tam dej. Jinak… | Yeah, probably, put it there. Or else … |
| A | To též je co, jako taky? Že též je taky? | That [też] is what, like also? [też] is also? |
| B | Též, jestli to bude jako též, též starý žlutý samochod. | [też], if it's also, it will be like also, also a yellow self-goer. |
| A | Já bych tam dal starý… starý žlutý auto jenom bych tam dal. Jako nekoupili jsme čerstvý chléb, ale ještě hůř, starý žlutý auto. | I would put old there … old yellow car I would put there only. Like we didn't buy fresh bread, but even worse, an old yellow car. |
| B | No, no, souhlasím. | Yeah, yeah, I agree. |
| A | Protože jako, nevím co… Jo počkej, nekoupili jsme jenom… A co kdyby to tylko bylo jenom? | Because like, I don't know … Wait, we didn't buy only … and what if this [tylko] is only? |
| B | Jenom, aha, jo, jo, jo. Nekou… jo, to dává smysl, jasně. Nekoupili jsme jenom… jo, no vidíš, takže to bude ten zápor. Ne- nekoupili jsme jenom čerstvý chléb, ale, ale ještě… | Only, aha, yeah, yeah, yeah. We didn … yeah, that makes sense, sure. We didn't buy only … yeah, you see, so it will be this negation. We didn - didn't buy only fresh bread, but, but even … |
| A | **No, teďka to dává smysl.** | **Well, now it makes sense.** |
| B | Jo... | Yeah … |
| A | Jo? | Yeah? |
| B | Supr, dobře, dobře ty, no. | Super, well, well done, man. |
| A | Dohromady to dáme, hele, ale knížku bysme si asi nepřečetli. Možná tak za půl roku, jako. | We can do it together, see, but we probably wouldn't be able to read a book. Maybe after half a year, like. |
| B | Tak já už bych to auto mrazila(?) před dvouma minutama, ty jo. | So I would have frozen (?) that car already two minutes ago. |

## References


Čermák, F., Rosen, A., 2012. The case of InterCorp, a multilingual parallel corpus. Int. J. Corpus Linguist. 13 (3), 411–427.

Crocker, M., Demberg, V., Teich, E., 2015. Information density and linguistic encoding (IDEAL). Künstliche Intell 30 (2016), 77–81. doi: 10.1007/s13218-015-0391-y.

Doyé, P., 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

Ericsson, K., Simon, H., 1993. Protocol Analysis: Verbal Reports as Data. second ed. MIT Press, Boston ISBN 0-262-05029-3.

Golubović, J., Gooskens, C., 2015. Mutual intelligibility between West and South Slavic languages. Russian Linguist. 39, 351–373.

Golubović, J., 2016. Mutual intelligibility in the Slavic language area. University of Groningen.

Gooskens, C., 2013. Experimental methods for measuring intelligibility of closely related language varieties. In: Bayley, Robert, Cameron, Richard, Lucas, Ceil (Eds.), Handbook of Sociolinguistics. Oxford University Press, Oxford, pp. 95–213.

Hale, J., 2001. A probabilistic earley parser as a psycholinguistic model. In: Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies. Pittsburgh, Pennsylvania. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1–8 http://dx.doi.org/10.3115/1073336.1073357 .

Harley, T., 2007. The Psychology of Language − From Data to Theory. Psychology Press. 2008 http://www.psypress.com/harley .

Heeringa, W., Golubović, J., Gooskens, C., Schüppert, A., Swarte, F., Voigt, S., 2013. Lexical and orthographic distances between Germanic, Romance and Slavic languages and their relationship to geographic distance. In: Gooskens, C., Bezoijen, van, Hg., R. (Eds.), Phonetics in Europe: Perception and Production. Peter Lang, Frankfurt a. M, pp. 99–137.

ISO 9:1986. Documentation − transliteration of Slavic Cyrillic characters into Latin characters. Documentation and Information. (ISO Standards Handbook 1). third ed. ISO, Genève. 1988.

Jaeger, T.F., 2010. Redundancy and reduction: speakers manage syntactic information density. Cogn. Psychol. 61, 23–62. doi: 10.1016/j.cogpsych.2010.02.002.

Jágrová, K., 2010. Russisch-Deutsch-Tschechische Interferenzen. State examination thesis TU Dresden.

Jágrová, K., Stenger, I., Marti, R., Avgustinova, T., 2016. Lexical and Orthographic Distances Between Czech, Polish, Russian, and Bulgarian − A Comparative Analysis of the Most Frequent Nouns, In: Language Use and Linguistic Structure. Palacký University, Olomouc 2017, 401−416.

Keller, F., 2010. Cognitively plausible models of human language processing. In: Proceedings of the ACL 2010 Conference Short Papers (ACLShort '10). Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 60–67.

Kneser, R., Ney, H., 1995. Improved backing-off for M-gram language modeling. In: Procedings of the International Conference on Acoustics, Speech, and Signal Processing, Detroit, MIvol. 1, pp. 181–184. doi: 10.1109/ICASSP.1995.479394. 1995.

Kürschner, S., van Bezooijen, R., Gooskens, C., 2008. Linguistic determinants of the intelligibility of Swedish words among Danes. Int. J. Hum. Arts Comput. 2 (1/2), 83–100.

Levy, R., 2008. Expectation-based syntactic comprehension. Cognition 106 (3), 1126–1177.

Levenshtein, V.I., 1966. Binary codes capable of correcting deletions, insertions, and reversals. Sov. Phys. Dokl. 10 (8), 707–710.

Möller, R., Zeevaert, L., 2010. "Da denke ich spontan an Tafel" – Zur Worterkennung in verwandten germanischen Sprachen. http://www.dgff.de/de/zff/zff-artikel-index/zff-artikel-detail/artikel/da-denke-ich-spontan-an-tafel-zur-worterkennung-in-verwandten-germanischen-sprachen.html?tx_ttnews%5BbackPid%5D=906&cHash=4176411eecfab086c7cf429baa911140 .

Nábělková, M., 2007. Closely-related languages in contact: Czech, Slovak, "Czechoslovak". Int. J. Sociol. Lang. 183, 53–73.

Schüppert, A., Haug Hilton, N., Gooskens, C., 2016. Why is Danish so difficult to understand for fellow Scandinavians? Speech Commun. 79, 47–60.

Vanhove, J., 2014. Receptive multilingualism across the lifespan. Cognitive and linguistic factors in cognate guessing. University of Freiburg.

V.V. Vinogradov Russian Language Institute of RAS: The Database of the Parallel Corpora within the Russian National Corpus (Parallel corpora of Slavic languages). Granting of the right to use data base (non-exclusive license), Moscow city, November 18, 2015