Intelligibility of highly predictable Polish target words in sentences presented to Czech readers

Klára Jágrová, Tania Avgustinova

Saarland University, CRC 1102: Information Density and Linguistic Encoding,
Department of Language Science and Technology
[kjagrova|avgustinova}@coli.uni-saarland.de

Abstract. This contribution analyses the role of sentential context in reading intercomprehension both from an information-theoretic and an error-analytical perspective. The assumption is that not only cross-lingual similarity can influence the successful word disambiguation in an unknown but related foreign language, but also that predictability in context contributes to better intelligibility of the target items. Experimental data were gathered for 149 Polish sentences [1] with highly predictable target words in sentence final position presented to Czech readers in a web-based cloze translation task. Psycholinguistic research showed that predictably of words in context correlates with cognitive effort to process the information provided by the word and its surprisal [3]. Our hypothesis is that intelligibility of highly predictable words in sentential context of a related language also correlates with surprisal values obtained from statistical trigram language models. In order to establish a baseline, the individual words were also presented to Czech readers in a context-free translation experiment [4]. For the majority of the target words, an increase in correct translations is observable in context, as opposed to the results obtained without context. The overall correlations with surprisal are low, the highest being the joint surprisal of the Polish stimulus sentence. The error-analysis shows systematic patterns that are at least equally important intercomprehension factors, such as linguistic distance or morphological mismatches.

Keywords: Slavic receptive multilingualism, Czech, Polish, statistical language modeling, context in intercomprehension, reading, surprisal, linguistic distance

1 Introduction

In previous research in cross-lingual intelligibility of written text, the role of linguistic distance (lexical, orthographic, morphological, syntactic, phonetic) was investigated as a predictor for human performance [cf., for instance, 5, 6, 9, 10, 13, 15]. Thus, linguistic distance is supposed to reflect the (dis)similarity of two related codes on the different linguistic levels: the lower the linguistic distance, the more similar and mutually intelligible the two codes should be. Lexical distance is determined as the percentage of non-cognates in a language pair, while orthographic and morphological distances are usually measured as string similarity by means of the Levenshtein distance (LD) [18].

As for the linguistic distance and intelligibility of Polish (PL) sentence material for Czech readers, findings from the literature are summarized in Table 1. Heeringa et al. found that PL is an outlier in terms of orthography among the other Slavic languages spoken in the EU [9]. Jágrová et al. [15] found that in relation to the low lexical distance (10%) between PL and Czech (CS), their orthographic distance (34%) is extraordinarily high when compared to Bulgarian and Russian (RU) that have similar levels of both orthographic (13.5%) and lexical distance (10.5%).

Table 1. PL for Czech readers: comparison of distances and intelligibility (in %) in related research.

Distance	Heeringa et al. [9]	Golubović [5][1]	Jágrová et al. [15]	Jágrová et al. [14]
Lexical	23	17.7	10	12
Orthographic	31	31.7	34	38
Morphological	-	31.4	-	-
Intelligibility	64.29[2]	41.01	-	-

The role of sentential context for the understanding of a particular language Lx, however, was subject to relatively few studies in this research field [14]. Muikku-Werner [19] qualitatively analysed the role of co-text in a study where Finnish students were asked to translate Estonian sentences. She found that the role of neighbourhood density – the number of available similar word forms – changes with words in context, as potential other options have to fit the restricted syntactic frame or be collocated [19, p. 105]. She states that “when recognizing one word, it is sometimes simple to guess the unfamiliar word frequently occurring with it, that is, its collocate. If there are very few alternatives for combination, this limitedness can facilitate an inference of the collocate” [ibid.].

In a study on the disambiguation of cross-Slavic false friends in divergent sentential contexts, Heinz [11] confronted students of different Slavic L2 backgrounds with spoken sentence samples in other Slavic Lx. He points out that the amount of perceived context is decisive for a successful comprehension of Lx stimuli. He also speaks of a negative role that context could play, namely if respondents attempt to formulate a reasonable utterance, they might revise their lexical decision [11], meaning that the target word might be misinterpreted due to misleading or misinterpreted context.

Another concept that is therefore likely to play a role in the intercomprehension of sentences is that of semantic priming [8]. Gulan & Valerjev [7] provide an overview of the types of priming that are identified in psycholinguistic literature (semantic, mediated, form-based, and repetition). The relevant type of priming for the present study appears to be semantic priming with both sub-types – associative and non-associative priming [7, p. 54]. During associative priming, a certain word causes associations of

[1] Data for the written cloze test [5]

[2] Data for the written translation task of the most frequent Ns from the British National Corpus as published in [5, p. 77] on the material of [9]

other words with the reader that might, but do not have to be related in meaning. Typical associations can be *engine–car* or *tree–wood*. A reader then might expect such a target word fitting a prime to occur in the sentence, for instance, at the position of an unfamiliar, unidentifiable word in the Lx. Cases of non-associative priming are words that are usually not mentioned together in such association tasks, but that are "clearly associated in meaning" [ibid.], for instance *to play – to have fun*. Semantic priming, of course, can only work if the prime is correctly recognized as such.

2 Hypothesis

Successful disambiguation of target words in a closely related foreign language relies on both cross-lingual similarity (measurable as linguistic distance) and predictability in sentential context (in terms of surprisal obtained from 3-gram LMs).

In a monolingual setup, it was shown that the more predictable a word is in context, the lower is the cognitive effort to process the information provided by the word – this corresponds to a low surprisal value [3]. On the contrary, words that are unpredictable in context and thus cause greater cognitive effort have higher surprisal values (see 3.2 for details). In the current multilingual setup, target words that have low linguistic distance to the reader's language and are predictable in context are expected to be translated correctly more often than words that are less similar and unpredictable. Since (dis-)similarity is measured by LD and predictability in context is captured by surprisal, the correct answers per target word should correlate with LD and surprisal better than only with LD.

Of course, the amount of correctly perceived sentential context plays a crucial role, too. If the context was not intelligible enough for the reader, then the supportive power of the context in terms of predictability might lose its effect. With a context that is helpful enough, it should be possible to recognize even non-cognates and maybe even false friends in sentences. The effects of semantic priming are not expected to be predictable by the trigram language models (LMs) applied here.

Consequently, the research questions can be formulated as follows:

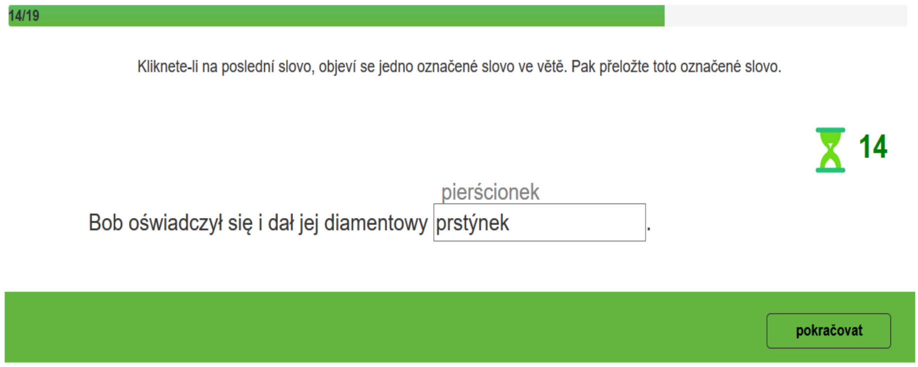
1. Are PL target words more comprehensible for Czech readers when they are presented in a highly predictive sentential context?
2. If so, do surprisal values obtained from trigram LMs correlate with the intelligibility scores of the target words?

3 Method

3.1 Design of the web-based cloze translation experiment

The cloze translation experiments were conducted over an experiment website [4]. After having completed the registration process including sociodemographic data, participants were introduced to the experimental task by a short tutorial video on the website. They were asked to confirm to have understood the task and to set their keyboard to

CS. With each stimulus sentence, they would initially see only the first word of the sentence. They were prompted to click on the word in order to let the next word appear. They were asked to follow this procedure until the end of the sentence. Only after they have clicked on the last word in the sentence, the cloze gap with the target word for translation was displayed. This method ensures that participants read each sentence word by word. There are two separate time limits: one for clicking and reading through the sentence and one for entering the translation of the target word. The latter was automatically set to 20-30 seconds, depending on the length of the sentence. For each target word, data from at least 30 respondents were collected in both conditions.



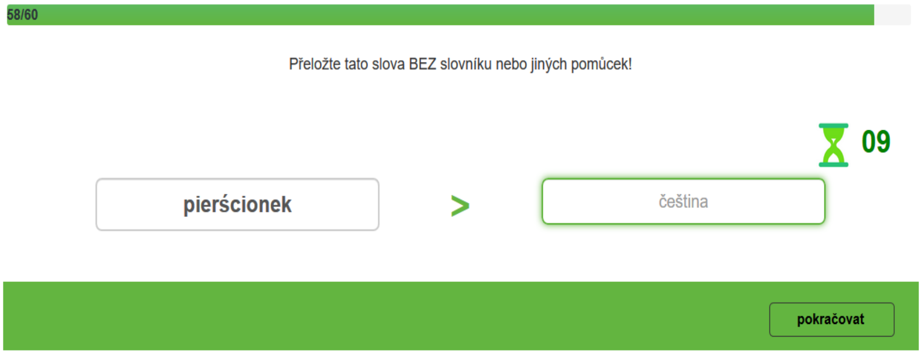
Kliknete-li na poslední slovo, objeví se jedno označené slovo ve větě. Pak přeložte toto označené slovo.

pierścionek

Bob oświadczył się i dał jej diamentowy prstýnek.

Fig. 1. Experimental screen in cloze translation experiments as seen by Czech respondents. The instruction on top says: ‘When you click on the last word, a marked word will appear. Then translate this marked word’. The target word is displayed on top of the frame, the correct CS translation is inside the frame.

As a baseline, the target word forms from the sentences were also presented without context and in their base forms to other Czech respondents over the same experimental website – see Fig. 2. Target words with identical base forms in both Ls were not tested without context. The respondents were asked to translate each presented word with a time constraint of 10 seconds.



Přeložte tato slova BEZ slovníku nebo jiných pomůcek!

Fig. 2. Experimental screen in context-free experiments as seen by Czech respondents. The instruction on top says: 'Translate these words without a dictionary or other aids.' Respondents have 10 seconds time to enter their CS translation.

3.2 Stimuli

In order to use stimuli with predictive context systematically, sentences from a monolingual cloze probability study by Block & Baldwin [1] were adapted. They tested a set of 500 sentences in a cloze completion task where the completion gap was always placed on the last position in each sentence. In addition to the cloze experiments, they validated the sentences in psycholinguistic ERP experiments. Their study resulted in a dataset of 400 high-constraint, high cloze probability sentences. From these sentences, those with the most predictable target words (90–99% cloze probability) were translated into PL for the present study. A colleague and professional translator for PL was asked to translate the sentences in such a manner that the target words remain on the last position in the sentences. The translated sentences are published as a resource in the data supplement.

In the original (American) EN set, there were sentences that contained particular cultural topics and therefore were omitted, which resulted in a set of 149 sentences. Few translations were modified where appropriate, e.g. the original sentence

> *When Colin saw smoke he called 911 to report a fire.* [1]

was modified into

> *Gdy Colin zobaczył dym, zadzwonił do straży pożarnej i zgłosił pożar*.
> 'When Collin saw the smoke he called the fire department and reported a fire.' [1]

The respondents were not informed that the sentential context presented is a helpful, high-constraint context or that the target words should be highly predictable.

Literal Translation for Measuring Linguistic Distance and Surprisal. Linguistic distance and surprisal as predictors of intelligibility were measured for the literal CS translations and for the original PL stimuli. These two measures were applied (i) to the whole sentences, (ii) to the final trigram, (iii) to the final bigram, and (iv) to the target word only. All measures were tested as total and normalized values. The literal CS translations (following the method e.g. in [9]) are meant to as exactly as possible reflect how a Czech would read the PL sentence. To score them with an LM trained on the Czech national corpus (CNC, [17]), it was necessary to ensure that all translated (pseudo) CS word forms can be found in the CNC, because if a form is not found in the training data, the LM would treat it as an OOV (out of vocabulary item). Grammatical forms, phraseological units, and prepositions were kept as in the PL original, e.g. *do* 'to' instead of the correct CS *k* in

Poszła do fryzjera, żeby ufarbować włosy.
'She went to the salon to color her hair.' [1]

which was transformed into

**Zašla do kadeřníka, žeby obarvit vlasy*.

for the calculation. Another example would be *genealogiczne drzewo* 'family tree' that was transformed into *genealogický strom* 'genealogical tree' instead of *rodokmen* 'family tree'. PL words existing in colloquial CS or in CS dialects and reflected in the CNC were also preserved in the literal translations, for instance the conjunction *bo* 'as, since' in

Nie mogła kupić koszulki, **bo** *nie pasowała*.
'She could not buy the shirt because it did not fit.' [1],

which would be *protože* 'because' in a written standard CS translation. PL negations and verb forms in the past tense or in the conditional mood required for their CS correspondences an explicit division of negation particles, verb forms, and auxiliaries. For instance, the negation particle *ne* was separated from CS verbs, and the PL example above was consequently transformed into

**Ne mohla koupit košilku, bo* ne pasovala.

instead of keeping the correct CS negated verb forms *nemohla* '(she) could not' and *nepasovala* '(it) did not fit'. Other examples are verb forms that are reflexive in only one of the languages, for instance, *dołączyła do zespołu* 'she joined the band' is not reflexive in PL, while in CS equivalent *přidala se do kapely* is reflexive. The reflexive pronoun was therefore omitted in the literal CS translation: **přidala do kapely*.

Non-cognates and false friends were replaced by their correct CS translations. The literal translations and their surprisal values and distance measures can be found in the data supplement.

Surprisal is an information-theoretic measure of unpredictability [3]. Statistical LMs inform us about the probability that a certain word w_2 follows a certain other word w_1. For a given word, the surprisal is the negative log-likelihood of encountering this word in its preceding context [3, 14]. It is defined as:

$$\mathit{Surprisal}\ (\mathit{unit}|\mathit{context}) = -\log_2 P(\mathit{unit}|\mathit{context}) \qquad (1)$$

Thus, surprisal reflects frequency and predictability effects in the corpus on which the LM was trained. Fig. 3 illustrates the principle of 3-gram LM counts.

```
Bob oświadczył się
    oświadczył się i
               się i dał
                  i dał jej
                    dał jej diamentowy
                        jej diamentowy pierścionek
                            diamentowy pierścionek .
```

Fig. 3. Example for 3-grams as they could occur in a PL corpus.

The PL stimuli sentences were scored by an LM trained on the PL part of InterCorp [2] and the CS literal translations were scored by an LM trained on the CNC [15]. Both were LMs with Kneser-Ney smoothing [16]. As the 3-gram LMs applied here cannot capture links between items further apart from each other than in a window of three words, the surprisal is expected to predict only such relations that are in direct successive position. Schematic implications such as

> ***Farmer*** *spędził ranek dojąc swoje* ***krowy****.*
> 'The **farmer** spend the morning milking his **cows**.' [1]

or hyponymy such as in

> *Ellen lubi* ***poezję****,* ***malarstwo*** *i inne formy* ***sztuki****.*
> 'Ellen enjoys **poetry**, **painting**, and other forms of **art**.' [1]

are not expected to be predictable with surprisal obtained from the 3-gram LMs.

Linguistic Distance. Orthographic distance was calculated as the Czechoslovak to PL pronunciation-based LD, i.e. always towards the closest CS or Slovak (SK) translation equivalent under the assumption that the Czech readers have receptive skills in SK (cf. method of Vanhove with Germanic distance [20, p. 139]). No costs were charged for the alignment of *w*:*v*, *ł*:*l*, *i*:*y*, *y*:*i*, *ż*:*ž* since their pronunciation is transparent to the readers [12]. If a target word is a non-cognate, its distance is automatically set to 1. Lexical distance is determined by the number of non-cognates per sentence in the language pair.

A separate variable for the category false friends has been added, as false friends can be both cognates and non-cognates (see section 4.2).

4 Results

4.1 Comparison: Target Words With vs. Without Context

The mean intelligibility of target words improved significantly from 49.7% without context to 68% in highly predictive contexts (t(298)=4.39, p<.001). This means that the hypothesis that predictive sentential context contributes to a better intelligibility of highly predictable words in an unknown related language can be confirmed for the scenario PL read by Czech respondents. Fig. 4 contains a trend line at *f*(*x*)=1*x* which divides the data points into those for which intelligibility improved in context (above the line) and those for which intelligibility decreased with the provided context (beneath the line). The points on the line are those for which no difference between the conditions with or without context could be discovered.

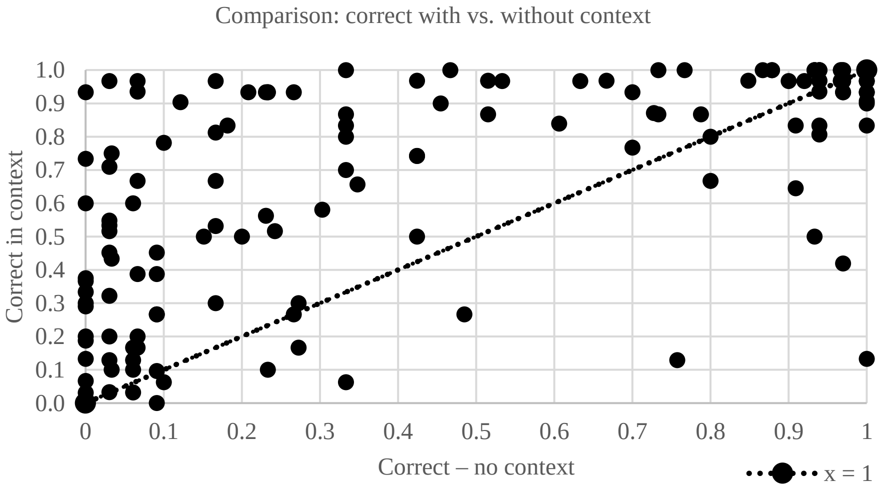


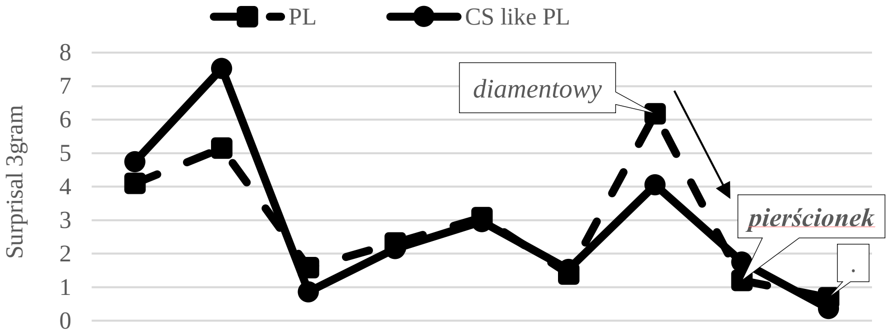
Fig. 4. Comparison of results for target words with vs. without context.

In the condition with context, a correctness rate of 100% could be observed for 26 target words, and 18 other target words were correctly translated by 96.7% of the respondents. In the condition without context, there were only 19 target words with a correctness rate of 100%, and 11 with ≥ 96.7%.

Cases of context-driven decisions are frequently observed in the responses, e.g.

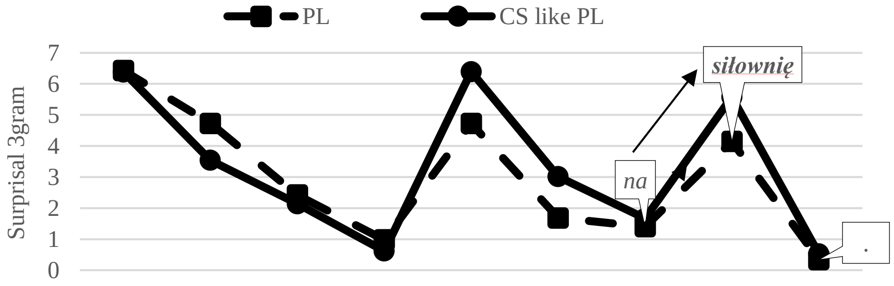
Bob oświadczył się i dał jej diamentowy pierścionek.
'Bob proposed and gave her a diamond ring' [1].

When presented in this sentence, 90% translated the PL target *pierścionek* 'ring' correctly, while in the condition without context only 45.5% gave the correct CS cognate *prstýnek*. The trigram LM confirms that the target *pierścionek* 'ring' is highly predictable after *diamentowy* 'diamond [adjective]' [1], which is indicated by the dropping surprisal curve after *diamentowy* in Fig. 5. The surprisal values are provided in the unit Hart (Hartley).



PL: *Bob oświadczył się i dał jej diamentowy* ***pierścionek****.*
CS: *Bob se zasnoubil a dal jí diamantový* ***prstýnek****.*

Fig. 5. Surprisal graph for the PL sentence *Bob oświadczył się i dał jej diamentowy pierścionek.* 'Bob proposed and gave her a diamond ring' [1].



PL: ***Sportowiec*** *lubi chodzić na podnoszenie ciężarów na* ***siłownię****.*
CS: ***Sportovec*** *rád chodí na vzpírání do* ***posilovny****.*

Fig. 6. Surprisal graph of *Sportowiec lubi chodzić na podnoszenie ciężarów na siłownię*. 'The athlete is enjoying lifting weights at the gym' [1].

In contrast to the sentence in Fig. 5, there is an increase in surprisal in Fig. 6 at the target *siłownię* 'gym [accusative]*'* for the sentence

> *Sportowiec lubi chodzić na podnoszenie ciężarów na siłownię.*
> ‘The sportsman likes to do weightlifting at the gym.’[3]
> ‘The athlete is enjoying lifting weights at the gym.’[4]

In the monolingual cloze completion task [1], 95% of English native speakers provided the response (*gym*), which suggests that the word *athlete* or *sportowiec* ‘athlete’ functions here as a semantic prime. So the higher rate of correct translations in context (58.1% vs. 30.3% without context) might be explained by the thematic association of the target word *siłownię* ‘gym [accusative]’ with the sentence-initial *sportowiec* ‘athlete, sportsman’ rather than with its directly preceding words *ciężarów na* ‘weights [genitive pl.] at’.

Fig. 4 shows an extraordinarily high increase in intelligibility for some targets, mostly for those that can be considered false friends but also have cognate translations (FF-C in 4.2). For example, *znaczek* – CS *známka* ‘stamp’ was frequently mistaken for *znak* or *značka* ‘sign’ (93.9% wrong) when presented without context. In a predictive context, however, it was translated correctly by 71% of the respondents. This was also the case for the target word *wazon* – CS *váza* – ‘vase’ which was mistaken for *vagon* ‘wagon’ (48.5%) without context (only 15.2% correct) and correctly translated by 50% in context. Section 4.2 provides an overview of target categories with examples; for a full list see data supplement.

4.2 Different Categories of Target Words

The intelligibility scores vary with different categories of target words in both conditions, i.e. with and without context – cf. Table 2.

Cognates Identical in Base Form. (C-IB, n=11) This sub-category of cognates includes target words with a base form that is identical in both languages, but not in the inflected forms as presented in the context. For instance, *ryba* ‘fish’ is identical in its base forms in both PL and CS but the PL target *rybę* in accusative differs from its CS correspondence *rybu*.

Real or True Cognates (C, n=89) differ only in orthography and/or in morphological features. We can observe a ceiling effect (maximum scores in both conditions) of target words with very low orthographic distance, such as PL *mokry* and CS *mokrý* ‘wet’ that differ only in diacritics and were translated correctly by all respondents. The same applies to target words with easily identifiable pronunciation, for instance, in *czasu* – CS *času* – ‘time [genitive]’ that was translated correctly by 96.8%. Interestingly, there are target words with a relatively high LD, e.g., PL *obiad* ‘lunch’ with an LD of 40% to the

[3] The PL translator was instructed to keep the target word at the last position in the sentences. Therefore, some translations might vary slightly from their original EN versions (cf. [1]).

[4] original version of the sentence as of [1]

CS *oběd* ‘lunch’, but an intelligibility score of 100% in context (cf. the sentence below) and 93.3% without context.

> *Zrobiła sobie kanapkę i frytki na obiad.*
> ‘She made herself a sandwich and chips for lunch.’ [1]

The intelligibility of true cognates correlates significantly with linguistic distance of the target word (r=.549, p<.001), but not with surprisal (r=.043, p<1).

Cognates in Other Contexts (C-OC, n=3): The items are cognates not in the presented sentence but in other contexts. For instance, PL *szczotka* ‘brush, broom’ can correspond to CS *štětka* ‘brush‘ only in come contexts, e.g. as a brush for shaving, but not as a broom (correct CS *smeták*) for sweeping the floor.

Non-Cognates with Correct Associations (NC-A, n=7): CS translations are not cognates of the PL targets, but they do share some common features. Thus, PL *latawiec* ‘kite’ might be associated with the CS verb *létat* ‘to fly’. Respondents are likely to associate the stimulus with a concept in their language and then come up with the correct CS translation *drak*.

Real Non-Cognates (NC, n=5): Unrelated lexical items that are not expected to be intelligible for the reader without context, e.g., PL *atrament* vs. CS *inkoust* ‘ink’.

False Friends as Cognates: (FF-C, n=15) These items are cognates frequently mistaken for another more similar CS word in at least one of the conditions: with or without context. This is one of four categories of false friends. As a threshold for false friends, the percentage of the particular wrong type of response must have been higher than the sum of no responses and correct responses. In addition, the particular wrong response must have been more frequent than the sum of all other wrong responses.

False Friends that are Cognates in other Contexts: (FF-OC, n=9) These frequently misinterpreted items are cognates in another context than in which they were presented. For example, PL *przebrać* ‘to change clothes’ is frequently mistaken for CS *přebrat* ‘to pick over’, while the correct translation would be *převléct se*.

False Friends with Correct Associations (FF-A, n=5) are words that are frequently mistaken for other more similar CS words which have some common semantic features with a correct cognate translation. Respondents might associate the stimulus with a concept in their language and then come up with the correct translation in context. For instance, PL *drzewo* ‘tree’ is frequently mistaken for CS *dřevo* ‘wood’, which at the same time can lead to a correct association with CS *strom* ‘tree’ in the respective context.

Real False Friends (FF, n=5) are frequently mistaken for another more similar CS word. For instance, PL *gwóźdź* ‘nail’ is frequently mistaken for CS *hvozd* ‘forest’, while the correct translation would be *hřebík*.

Table 2. Intelligibility of target words with vs. without context in the different categories.

	C-IB	C	C-OC	NC-A	NC	FF-C	FF-OC	FF-A	FF
no context	94.5%	65.9%	4.0%	8.7%	6.3%	18.4%	3.83%	3.33%	4.9%
context	81.4%	80.1%	16.6%	49.8%	31.1%	68.6%	19.33%	42.5%	26.3%
t-test		t=3.05		t=5.07	t=1.90	t=5.28	t=2.45	t=2.72	
significance	ns	p<.01	ns	p<.001	p<.05	p<.001	p<.05	p<.05	ns

The differences between the intelligibility of target words with vs. without context are significant for all categories except for cognates identical in their base form, cognates in other context and real false friends. The greatest and highly significant difference between the two conditions was found for target words that are false friends but have cognate translations.

4.3 Analysis of Wrong Responses

The error analysis of responses reveals some features of target words that linguistic distance and surprisal can account for only to a limited extent, if at all:

Differences in Government Pattern. In some sentences, the target words seem to have been more difficult, probably because of differences in government patterns. For instance, the target word *dzień* ‘day’ was translated more often correctly without context (80%) than in context (66.7%) of the sentence

> *Dentysta zaleca myć zęby dwa razy na dzień.*
> ‘The dentist recommends brushing your teeth twice a day.’ [1].

This might be explained by two factors. Firstly, the translation of the PL phrase *na dzień* ‘per day’ is headed by a different preposition in CS – *za den* – or it can be expressed by a single adverb – *denně* ‘daily’. Secondly, and in connection with the first factor, the wrong responses include highly similar words that respondents probably thematically associated with the concept of a dentist from the stimulus sentence: *dáseň* 'gum', *díru* 'hole', or *žížeň* ‘thirst’. Moreover, in CS, these responses occur often together with the preposition *na* ‘on’, e.g. *na dáseň* ‘for (your) gum’, *na žízeň* 'against thirst' and thus might seem perfectly legitimate to the respondents.

Ln Interferences. Effects of another language (Ln) interference occur relatively rarely (with 11 target words) among the responses in context. Out of the 5208 data points for the context condition, 37 responses could be classified as interferences from EN, DE or

SK. One of the few obvious interferences was at the target word *drzwi* ‘doors’ which was translated as EN *drive* by one Czech respondent who indicated to live in Great Britain. Also, *głosu* ‘voice [genitive]’ was translated as *skla* ‘glass [genitive]’ by another respondent living in Great Britian. One respondent translated *biurku* ‘desk [locative]’ as *tužka* ‘biro’, probably due to the similarity of PL *biurko* and EN *biro*. The target word *ból* – CS *bolest* – ‘pain’ was translated as *byl* ‘he was’ by 53.3% of the respondents, probably due to the SK past tense verb form *bol* ‘he was’. Another 6.7% translated *ból* as *míč* ‘ball’, most likely due to the EN *ball*.

One of the responses was most probably a combination of Ln interference and priming: the target word *torcie* ‘cake [locative]’ in the sentence

Jenny zapaliła świeczki na urodzinowym torcie.
‘Jenny lit the candles on the birthday cake.’ [1]

was translated as *svícnu* ‘candlestick’ [genitive/dative/locative] by 16.1% of the respondents. This probably happened though the EN word *torch* and through the successful recognition of *świeczki* ‘candles’ as the CS *svíčky* ‘candles’.

(Perceived) Morphological Mismatches

PL Feminine Accusative Ending –ę in Ns: *swoją rolę* ‘her role [accusative]’ was translated as *role* [nominative singular or nominative/accusative plural] when the correct equivalent would have been *roli* [accusative] in CS. Nevertheless, *role* was counted as a correct answer as the interpretation of the target word as a plural does not harm the overall understanding of the sentence. 26.7% translated the target word *próbę* ‘test, try’ in the sentence

Kim chciała iść na sportownię na kurs na próbę.
‘Kim wanted to give the workout class a try.’ [1]

with words ending with an *-e*, *-é* or *-ě*: *přírodě* [dative of *příroda* ‘nature’], *tance* ‘dances’, *hřiště* ‘sport field, playground’, *sondě* [dative of *sonda* ‘sond’], *laně* [locative of *lano* ‘rope‘], *poprvé* ‘for the first time’, *zkoušce* [dative of *zkouška* ‘test’] for which the correct CS translation would have been *zkoušku* [accusative of *zkouška*].

PL Feminine Instrumental Ending of Ns –ą is apparently mistaken for the regular feminine ending in the nominative or accusative case *–a*. A regular PL-CS correspondence of these endings should be *ą*:*ou,* although other correspondences with PL *–ą* also occur. Typical mistakes were translations of *królową* as *králova* ‘the king’s’, *szczotką* as *šotka* ‘Scottish woman’, *pocztą* as *pocta* ‘honour’.

Verb forms in third person plural, e.g. *kwitną* ‘they bloom’ in which the ending *-ą* would correspond to the CS verb ending *–ou* were also frequently mistaken for a feminine N ending: 13% translated it with a feminine N, e.g., *teplota* ‘temperature’, *květina* ‘flower’ or *kytky* ‘flowers’ [colloquial] instead of *kvetou*.

Words with Different Grammatical Gender. Target words with different grammatical gender were translated less often correctly when presented without any context than in context. There were 11 target words with divergent grammatical gender between PL stimulus and correct CS translation. In all 11 cases, the greatest percentage of the responses is of the same gender as the stimulus in the condition without context. For instance, for the target word *biurko* ‘desk’, respondents have entered a number of neuter Ns, such as *pero* ‘pen’, *pírko* ‘little feather’, *horko* ‘hot weather’. This changed drastically in the condition with context. The percentage of correct responses increased with sentential context in all cases. The difference of correct responses between the two conditions ranges from 3.1% to 73.3% with a mean increase by 28.3% per word pair. The difficulty for the readers here was to consider the possibility that the correct translation might actually be of a different grammatical gender.

Concerning potential misinterpretations of inflectional endings, only the form *napiwku* [genitive] of *napiwek* ‘tip’ that, if not identified correctly as an inanimate masculine genitive form, might easily be misperceived as a feminine accusative form with the inflectional suffix *–u* in CS. Nevertheless, the percentage of feminine responses for the form *napiwku* in context did not increase when compared to the responses for the base form *napiwek*.

Infinitive Verb Forms Mistaken for Ns. A number of respondents apparently perceived the PL infinitive ending *–ć* for a correspondence to the CS nominal masculine agentive suffix *–č,* while the correct PL-CS correspondence for infinitive verb endings would be *ć:t*. The two suffixes (PL infinitive *–ć* and CS derivational *–č*) are indeed phonetically cloze. One of the prominent examples was the target word *bawić* ‘to play’ that was translated as *bavič* ‘entertainer’ by 39.4% when presented without context. Also, other Ns which the respondents most probably associated with the concept of *bavič* were among the responses: *komik* ‘comedian’ and *zábava* ‘amusement’. The verb appeared in two of the sentences, where it was translated as *bavič* significantly less often – 13.3% and 3.2% respectively.

When *padać* was presented without any context, only 62.9% of the respondents translated the target word correctly with its CS cognate *padat*. It was often mistaken for *padák* 'parachute'. When presented in the sentence

> *Zauważyłam, że nie mam parasola, gdy zaczęło padać.*
> ‘I realized I had no umbrella as it began to rain.’ [1],

however, 96.7% translated it correctly as *padat* ‘to fall’ or *pršet* ‘to rain’. The share of infinitive forms mistaken for Ns range from 0 in both conditions to 76.7% for target infinitives without context. On the average, 30.4% of all infinitives without context and only 5.9% infinite verb forms in context were mistaken for Ns.

A complete table with a comparative overview of the target verbs together with the frequencies of their misinterpretations as Ns and correct responses with and without context can be found in the data supplement. As a result of the error analysis, a binary variable for difference in grammatical gender was added in the regression model in order to represent the added difficulty of such target words.

4.4 Correlations and Model

With regard to surprisal, only the surprisal values of the target words and of the whole sentences have a low, but significant correlation with the results obtained in the context condition. The correlation of the CS target words' surprisal and target word intelligibility is only slightly higher than the PL surprisal of the target words (r=.191>r=.186). The correlations of the mean and total surprisal values of the whole sentences with the results in context are only significant in the case of the original PL stimuli sentences, not in the case of their closest CS translations. However, when leaving the cognates out of the correlation analysis, the correlation with the total surprisal of the PL sentence increases to r=.411 (p<.01), even more when correlating only the false friends (all categories) and intelligibility (r=.443, p<.01).

There is a highly significant covariance between the corresponding surprisal measures (for target, bigram, trigram, and sentence) from the two LMs (the CS and the PL one – see 3.2.), the strongest correlation being that of the total surprisal per sentence in both languages (r=.732, p<.001).

For the linguistic distance measures, all correlations are highly significant for the target words in context. The correlations are the highest with the linguistic distance of the target. The longer the involved string of words, the lower the correlation between distance and intelligibility of target words gets: target word > bigram > trigram > sentence. The correlation of intelligibility and linguistic distance is higher for the target words without context (r=.772, p<.001) than in context (r=.680, p<.001).

All lexical distance and false friends variables proved to be highly significant, the strongest correlation being the total lexical distance of the entire sentence (r=.508, p<.001). Both lexical distance and false friends correlate stronger with the results (r=.353 for the category of false friends, p<.001) when they are counted as a total score per sentence than when normalized through the number of words in a sentence. In context, a relatively low, but highly significant correlation was found for the target word having a different gender in the two languages (r=.272, p<.001). Without context, the correlation of grammatical gender and intelligibility is only slightly higher and highly significant (r=.281, p<.001). No correlation was found for the number of words in a sentence. A multiple linear regression with the relevant variables distance of target word, PL sum of surprisal for the sentence, and number of non-cognates per sentence results in a highly significant adjusted R^2=.496 (p<.001), i.e. this model can account for 49.6% of the variance in the data for all sentences.

5 Discussion and Conclusion

When viewing the whole stimulus set, the results show clearly that context helps to correctly identify highly predictable target words in sentential context as opposed to the same words without context. However, the correlations with surprisal are low, the highest being the sum of surprisal of the PL stimulus sentence (not of the closest translation). Other factors appeared to be at least equally important, most of all linguistic distance of the target word and the target word being of a different gender in the two languages.

The error-analytical observations lead to the conclusion that divergent grammatical gender of words in a related foreign language can be strongly misleading and that readers very often tend to choose a translation with the same grammatical gender, especially when there is no sentential context. As soon as sentential context is available, the role of the different grammatical gender loses its dominance. Czech readers proved to be unlikely to identify the PL ending –*ą* as an instrumental marker similar to the CS –*ou*, but often mistook it for a feminine nominal ending. Accordingly, the PL accusative ending –*ę* was frequently mistaken for a plural marker or an ending similar to the CS –*ě* in feminine dative or locative forms or neuter locative forms. It was shown that predictive context helps to correctly identify infinite verb forms in sentences, since they were significantly more often mistaken for Ns when presented without context.

However, individual cases of wrong associations with a thematically dominant concept in the sentences have shown that even understandable high-constraint sentential context can lead to wrong associations and to a lower number of correct responses than without context, even if the target word is a frequent cognate.

An analysis of results for the different lexical categories of target words reveals different levels of importance of the predictors in for these categories. The differences in correct responses between the context and the context-free condition were significant for all categories of target words except for those identical in base forms, cognates in other contexts, and real false friends. The difference between the two conditions was the greatest for false friends that are cognates and for non-cognates that offer possible associations with the correct translations.

For true cognates, no significant correlation between intelligibility and surprisal was found. However, surprisal as a predictor has a much greater impact if target words are non-cognates or false friends than if they are cognates, which suggests that in disambiguation of these, readers rely more on context than on word similarity. The effect of the predictive context seems to be especially striking with non-clear-cut cases of false friends. Since the correlations with linguistic distance are lower for target words in context than without context, the influence of linguistic distance on intelligibility proved to decrease in predictive sentential context. In the final regression model, the total surprisal of the sentence obtained from the PL model has a low, but significant correlation with the results.

References

1. Block, C. K., Baldwin, C. L.: Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. Behavior Research Methods 42(3), 665–670 (2010). https://doi.org/10.3758/BRM.42.3.665
2. Čermák, F., Rosen, A.: The case of InterCorp, a multilingual parallel corpus. International Journal of Corpus Linguistics 13(3), 411–427 (2012)
3. Crocker, M., Demberg V., Teich, E.: Information Density and Linguistic Encoding (IDEAL). In: *Künstliche Intell* 30, 77–81 (2016). doi:10.1007/s13218-015-0391-y
4. Experiment website http://intercomprehension.coli.uni-saarland.de/en/
5. Golubović, J.: "Mutual intelligibility in the Slavic language area." PhD diss., University of Groningen (2016)
6. Golubović, J., Gooskens, C.: Mutual intelligibility between West and South Slavic languages. Russian Linguistics 39, 351–373 (2015)
7. Gulan, T., Valerjev, P.: Semantic and related types of priming as a context in word recognition. Review of Psychology 17(1), 53–58 (2010)
8. Harley, T.: The Psychology of Language – From Data to Theory. 2nd edn. Psychology Press, Hove (2008). http://www.psypress.com/harley
9. Heeringa, W.; Golubovic, 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: Charlotte Gooskens & Renée van Bezooijen (eds.). Phonetics in Europe: Perception and Production, pp. 99–137. Peter Lang, Frankfurt a.M. (2013).
10. Heeringa, W.; Swarte, F.; Schüppert, A; Gooskens, C.: Modeling Intelligibility of Written Germanic Languages: Do We Need to Distinguish Between Orthographic Stem and Affix Variation? Journal of Germanic Linguistics 26(4), 361–394 (2014). DOI: https://doi.org/10.1017/S1470542714000166
11. Heinz, C. 2009. Semantische Disambiguierung von false friends in slavischen L3: die Rolle des Kontexts. ZfSl 54(2), 147–166 (2009).
12. Jágrová, K.: "Reading Polish with Czech Eyes" PhD diss., Saarland, Saarbrücken University (to be published)
13. Jágrová, K.: Processing Effort of Polish NPs for Czech Readers – A+N vs. N+A. In: Guz, W., Szymanek, B. (eds.) Canonical and non-canonical structures in Polish. Studies in Linguistics and Methodology vol. 12, pp. 123–143, Wydawnictwo KUL, Lublin (2018)
14. Jágrová, K., Avgustinova, T., Stenger, I., Fischer, A.: Language Models, Surprisal and Fantasy in Slavic Intercomprehension. Computer Speech and Language 53, 242–275 (2019). https://doi.org/10.1016/j.csl.2018.04.005
15. Jágrová, K., Stenger, I., Marti, R., Avgustinova, T.: Lexical and Orthographic Distances between Czech, Polish, Russian, and Bulgarian – a Comparative Analysis of the Most Frequent Nouns. In: Olomouc Modern Language Series (5) – Proceedings of the Olomouc Linguistics Colloquium, pp. 401–416. Palacký University, Olomouc (2017).
16. Kneser, R., Ney, H.: Improved backing-off for M-gram language modeling. In: International Conference on Acoustics, Speech, and Signal Processing, Detroit, MI vol. 1, pp. 181–184 (1995). doi:10.1109/ICASSP.1995.479394.
17. Křen, M.; Cvrček, V.; Čapka, T.; Čermáková, A.; Hnátková, M.; Chlumská, L.; Jelínek, T.; Kováříková, D.; Petkevič, V.; Procházka, P.; Skoumalová, H.; Škrabal, M.; Truneček, P.; Vondřička, P.: Zasina, A.: SYN2015: reprezentativní korpus psané češtiny. Ústav Českého národního korpusu FF UK, Praha (2015). Accessible: http://www.korpus.cz

18. Levenshtein, V.: Binary codes capable of correcting deletions, insertions, and reversals. Cybernetics and Control Theory (10). 707–710 (1966).
19. Muikku-Werner, P.: Co-text and receptive multilingualism – Finnish students comprehending Estonian. Nordic Journal of Linguistics. 99–113 (2014). doi: http://dx.doi.org./10.12697/jeful.2014.5.305
20. Vanhove, J.: Receptive multilingualism across the lifespan. Cognitive and linguistic factors in cognate guessing. PhD diss, University of Freiburg (2014). https://doc.rero.ch/record/210293/files/VanhoveJ.pdf. Last access 17.01.2019.