Processing effort of Polish NPs for Czech readers – A+N vs. N+A

Klára Jágrová
Saarland University
kjagrova@coli.uni-saarland.de

Abstract: This contribution compares the impact of two canonical grammatical features of Polish – the pre- and postmodification of nouns by adjectives – on the intelligibility of Polish for Czech readers. In contrast to Polish, the postmodification of nouns by adjectives is considered archaic or literary language in Czech, but possible also in general. Consequently, we postulate that post-nominal adjectives in NPs cause additional processing effort for Czech readers when they attempt to read and understand Polish. We correlated linguistic distance, surprisal scores obtained from 3gram language models, and overall difficulty scores with the results of a free translation experiment of Polish NPs in the A+N and N+A condition and found a moderate correlation between these predictors and processing time in the most representative subset of the data that was gathered. We also found that Czech readers were able to translate more words correctly in the A+N condition than in the N+A condition. This study is part of a greater research interest on the intelligibility of Polish for Czech readers.

Keywords: reading intercomprehension, Czech, adjective placement in Polish, processing difficulty, noun phrases

1. Introduction

1.1. Adjectival modification in PL

There have been several contributions discussing the systematic distinction between the adjective+noun (A+N) and noun+adjective (N+A) linearization in Polish (PL). For instance, Cetnarowska (2013) presents a description of adjectival modification in PL drawing back on the representational theory formulated by Bouchard (1998, 2002) about the location and interpretation of adjectives (As) in French and English. She states that "the most common position of classifying modifiers in Polish is the post-head position" and that "the classifying post-head adjectives are subsective" (Cetnarowska 2013: 19). We aim at investigating if NPs with N+A linearization in PL are more difficult to understand for Czech readers who are trying to understand PL than NPs with A+N linearization, which is the typical one in Czech (CS). Cetnarowska, Pysz and Trugman (2011) observe "a slight difference in the interpretation of A+N and N+A units containing classifying adjectives in Polish since the A+N phrases are perceived as less formal while N+A units are typical of scientific discourse" (Cetnarowska 2013: 20). In CS, the usual linearization is A+N, and N+A is likewise considered very formal, archaic or literary style, but it is also possible, for instance, in biological terms such as *šalvěj lékařská* 'Salvia officinalis'. Nevertheless, N+A linearization of NPs occurs much less in CS – see section 3.2. for further details on the typicality of the two linearizations.

1.2. Statistical language models (LMs) as predictors of processing difficulty

In a monolingual situation, statistical language models (LMs) inform about the predictability of words given a certain history of words. Consequently, LMs can inform about the probability that a certain A follows a certain N and a certain N follows a certain A. In psycholinguistic research, processing effort in monolingual readingsituations was measured in terms of event-related potentials (ERPs) or by reading time of stimuli. Roger Levy (2008) showed that n-gram LMs, specifically trigrams, performed well at predicting the processing difficulty which was measured by reading times of texts of various difficulties. The employed measure is called surprisal and is defined as:

$$Surprisal\ (unit|\text{context}) = -\log_2\ P(unit|\text{context})$$

Surprisal is widely used in information-theoretic modelling of human language. Surprisal reflects frequency and predictability effects in 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). For a word, surprisal is the negative log-likelihood of encountering this word in its preceding context.

Using our knowledge of the world, we know that *dom* 'house' is a predictable continuation after *biały* 'white', while for instance *sześciokąt* 'hexagon' is not. This is reflected well by a LM trained on a corpus of PL which assigns a high probability – and hence low surprisal score (1.12 hartley) – to *dom* after *biały*, while assigning a low probability – and hence a high surprisal score (7.02 hartley) – to the word *sześciokąt* after *biały*.

If we accordingly score both words in the NPs, we obtain a total surprisal score for both words of the NP. We obtain a total suprisal score of 3.05 hartley for *biały dom* 'white house' (1.93 + 1.12) and 11.19 hartley for *biały sześciokąt* 'white hexagon' (4.18 + 7.02). Thus, if an N is highly unexpected after a certain A, it will lead to a high total suprisal score of the NP.

1.3. Intelligibility of Polish for Czech readers

The phenomenon of intercomprehension reveals a robust human ability to understand an unknown language, without being able to use it actively, i.e. for speaking or writing. This scenario works more or less well, depending on the language-reader combination. Several linguistic and extra-linguistic factors may influence the successful disambiguation of unfamiliar linguistic code by a reader. Golubović (2016) measured the linguistic distances (for definitions see section 3.1.) between the Slavic languages spoken in the European Union and found that PL is an outlier in terms of orthography, having the greatest orthographic distance to the other five Slavic EU languages. PL has an orthographic distance of 31.7 % and a lexical distance of 17.7 % if read by Czech readers (Golubović 2016: 47–49). This finding was confirmed by Jágrová et al. (2017) who found that, in comparison to other language combinations, a large discrepancy prevails in the CS-PL pair with regard to their low lexical distance on the one hand (12% on average) and their high orthographic distance (34.5% on average) on the other hand, while the lexical and orthographic distances of other Slavic language combinations do not differ to such an extent (Jágrová et al. 2017: 411-13). This suggests that orthography alone might crucially impair the intelligibility of PL for Czech readers.

The role of context in intercomprehension, however, has been subject to very few studies (e.g. Heinz 2008), although context does play an essential role for successful intelligibility. As far as we know, no intercomprehension study has attempted to capture the role of phrasal context as a measurable variable.

In the following section, we introduce the hypothesis of the two dimensions influencing processing difficulty in intercomprehension: linguistic distance and surprisal in context. We postulate that the difficulty caused by context in intercomprehension can be measured by surprisal. We then explain the expected interaction of factors on the two dimensions, resulting in a predicted overall processing difficulty of the NP stimuli. We introduce our stimuli in section 3, the experiments in section 4 and we finally present the experimental results in relation to our hypothesis in section 5.

2. Hypothesis

Two Dimensions of processing difficulty in intercomprehension: linguistic distance and surprisal

The basic assumption is that processing difficulty in intercomprehension results from factors on two orthogonal dimensions: **(i) linguistic distance** and **(ii) suprisal in context**. Stimuli that have a high linguistic distance with regard to the reader and are at the same time relatively unpredictable in context are expected to cause the greatest difficulties to the reader. If a stimulus is linguistically close or identical to the reader's language, the same processes should apply for the predictability of words given a history as they do in a monolingual situation.

The underlying hypothesis is that the unexpectedness of the post-nominal attributes in a NP will cause greater processing effort for CS readers when trying to intercomprehend PL. It is expected that the greater processing effort manifests itself in higher response times as well as in a lower percentage of correct translations.

3. Stimuli

In order to determine if this hypothesis holds, web-based free translation experiments are conducted, where 109 different PL NPs are presented to Czech readers in two different conditions: 109 NPs with A+N and the same NPs in N+A linearization. The total of 218 NPs is divided into blocks of 4x36 plus 2x37, so that when participants start an experiment on the website, a block containing 36 or 37 NPs will be presented to them. Only after having finished the first block, the next block of NPs will be offered to the participants. The participants could decide if they do one, two or three experimental blocks. The stimuli blocks were arranged in such a way that each NP was presented to a reader in only one of the two conditions. Within a stimuli block, the number of NPs from each condition was evenly distributed. The NPs of a block were presented automatically in random order.

The NPs were compiled from the most frequent Ns and As (Ns were manually matched with suitable As) from a readily available list of the most frequent PL lemmas[1] (Broda and Piasecki 2016). An expert native speaker of PL was consulted to look over the stimuli and to provide all possible correct translations for each NP, especially with regard to differences in meaning between the two linearizations. In addition to the stimuli compiled from the most frequent lemmas, 9 NPs that were part of sentence stimuli in a previous sentence translation experiment (Jágrová forthcoming) were included into the stimulus set for future comparison.

3.1. Predictions of processing difficulty resulting from linguistic distance

In research on intercomprehension, linguistic distance was traditionally used as a predictor for the mutual intelligibility of closely related languages (cf. Gooskens 2013). Linguistic distance was usually measured with regard to lexis, orthography, morphology or phonology separately. With an increase in linguistic distance there is an increase in difficulty for the reader of a related language, which results in lower mutual intelligibility of two related languages. If a text has low linguistic distance, then transfer of knowledge from a language L1 to an unknown language LX is possible.

3.1.1. Lexical distance

We speak of lexical distance when words in an unknown language LX cannot be translated with cognates in the reader's language. We define cognates as words that are etymologically related and are recognizable as such. We do not distinguish between etymologically related or loan words, as long as the words share a meaning in at least one possible context (cf. Jágrová et al. 2017). In a first step, we look at the lexical distance of the PL stimuli to their corresponding closest CS translations. If a PL stimulus word can be translated correctly with a CS cognate, we assign a lexical distance value of 0. The cognate translations do not have to be ideal translations. Instead, they represent a L1 word that facilitates the comprehension of a word in LX. For instance, if a Czech native speaker reads *naturalny* 'natural', (s)he will most probably associate it with *naturální* 'natural' and only then with the more frequent Czech translation equivalent *přírodní* 'natural'. If there is no correct cognate translation, a distance value of 1 is assigned. For instance, there is no CS cognate translation for the PL adjective *poszczególne* 'individual', hence we translate it with *jednotlivé* 'individual' and assign a lexical distance value of 1. If the PL stimulus word is a false friend in CS, we assign the highest value for lexical distance: 2. Previousstudies on lexical distance (e.g. Gooskens et al. 2013) treated false friends like other non-cognates. This would mean to also assign a distance value of 1 to them. In a regression analysis of the experimental results we found that predictors calculated with a lexical distance score of 2 for false friends correlate better with processing times of the NPs than predictors calculated with a distance score of 1 for false friends.

Viewed separately, only 12 of the Ns and 16 of the As in the stimuli do not have cognate translations in CS. We view the stimuli NPs as units in which each component (A and N) contributes equally to the overall lexical distance of the stimulus. In the stimuli set, there are 82 NPs consisting of 2 cognates (lexical distance is 0), 16 that are combinations of a cognate and a non-cognate (lexical distance is 0.5), 9 NPs with a lexical distance of 1 that consist either of two non-cognates or of a false friend and a cognate, and 2 NPs that are combinations of two false friends (*ostatni okres* 'last period' and *kolejny raz* 'another time', having a lexical distance of 2 and being false friends to *ostatní okres* 'other district' and *kolejní ráz* 'rail character'). The lexical distance values are represented in the table in the appendix in the column labelled *Lex*.

3.1.2. Orthographic distance – Levenshtein distance (LD)

Even if words in a related language are cognates, they can be difficult to identify for CS readers, most probably because of their relatively high[2] orthographic distance. Then we speak of orthographic distance of cognates, i.e. there is no lexical distance, but orthographic distance.

[1] Compiled from a corpus of 1.8 billion tokens, including IPI PAN, Korpus Rzeczpospolitej, the PL Wikipedia (backup version of 2010) as well as an extensive collection of online documents (Grupa Technologii Językowych G4.19 Politechniki Wrocławskiej 2016).

[2] Compared to other language pairs, e.g. Russian and Bulgarian that have only 13% orthographic distance (cf. Jágrová et al. 2017: 413).

# slots	1	2	3	4	5	6	7	8	LD
PL stimulus	c	z	ł	o	w	i	e	k	
closest CS cognate	č		l	o	v		ě	k	
Costs	0.5	1	0.5	0	1	1	0.5	0	4.5/8 = 0.5625

Tab. 1) Example for the calculation of the Levenshtein distance of a cognate pair.

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. 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. This value is represented in the table in the appendix in the column labelled *Orth.*

The average orthographic distance within the stimuli NPs is 40 % for the As and 33 % for the Ns. The distances range from 0 for a number of identical Ns (e.g. *rada*, *projekt*, *kraj*, *grupa*, *firma*) and such minimally distant As that differ only in diacritics (e.g. *podobny – podobný* with a distance of 0.0714 %) to such distant cognates pairs as *mężczyzna – muž* (0.8333 %).

3.2. Predicting processing difficulty with surprisal

Surprisal scores can be interpreted as a measure for the typicality of certain constructions. We expect to get reliable information of about how likely one word anteceeds another one from surprisal scores. By combining the Ns and the As, a limited context is created which might influence the performance of Czech readers trying to understand the NPs compared to a situation in which they would see only individual Ns or As separately. For instance, if a Czech reader tries to translate the PL stimulus *pewna ręka* 'steady hand', the A *pewna* might be helpful in finding out that *ręka* is not *řeka* 'river', but *ruka* 'hand' in CS, because *pevná řeka* 'steady river' would not make much sense.

We trained statistical trigram models with Kneser-Ney smoothing on both the stimulus and the readers' language – one on the PL part of SCD InterCorp (size: 118.651.918 words) and one on the Czech National Corpus (SYN version 5, released in 2015, 4.599.643.984 words). 3gram models iterate through a corpus and count the ocurrences of all combinations of three consecutive words in a corpus. With the help of the surprisal scores from the CS LM, we can not only estimate which word order is more likely, we can also estimate how likely particular Ns are after particular As, respectively how likely particular As are after particular Ns.

The lower the surprisal score of a NP, the more expectable it should be for a reader. If we compare the surprisal scores obtained from a PL LM, we observe that 73 of the 109 NPs are more likely to appear in the A+N order than in the N+A order. The differences between the surprisal scores of the A+N and the N+A conditions range from -4 (surprisal(*ciężka głowa*) – surprisal(*głowa ciężka*) = 6.7 – 10.7 = –4), suggesting a preference for the A+N order, to +4.27 (surprisal(*zmianowa praca*) – surprisal(*praca zmianowa*) = 11.96 – 7.69 = 4.27), suggesting a preference for the N+A order. The mean difference between the two conditions in the stimuli set is 0.48 (SE = 1.74).

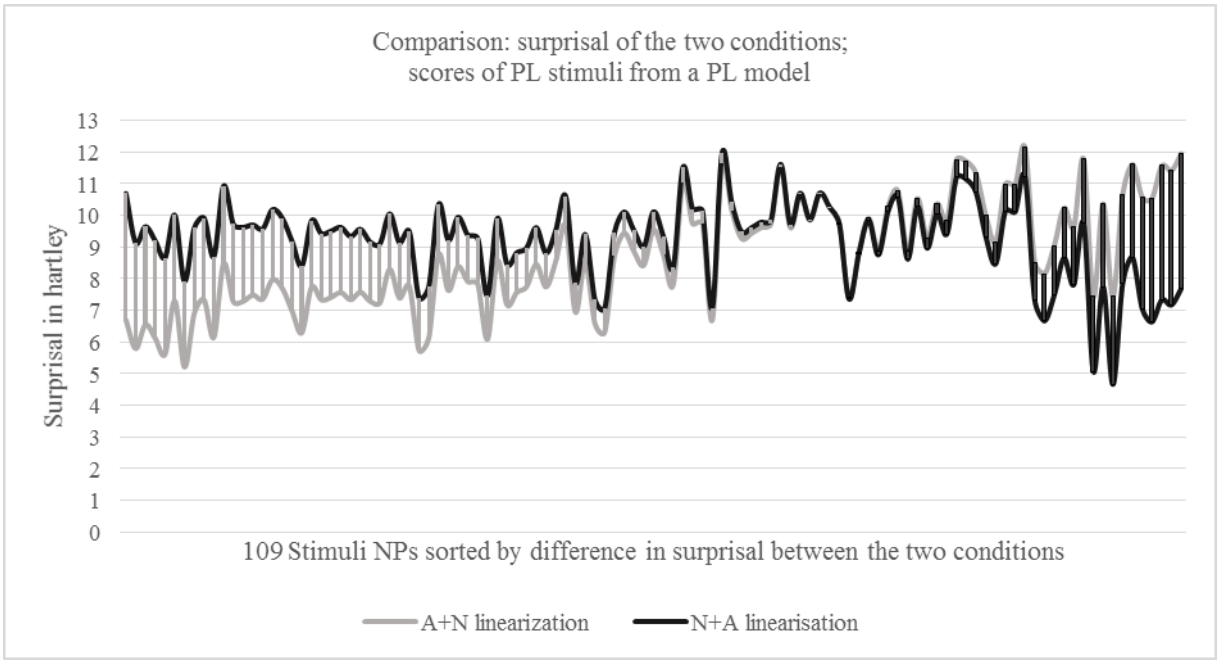

Fig. 1) Comparison of surprisal scores of the PL stimuli, obtained from a PL LM. The NPs are sorted by difference between surprisal scores, starting from *głowa ciężka* 'heavy head' to the left, where the N+A linearization has the biggest difference in surprisal compared to the same NP in its more typical A+N linearization – *ciężka głowa*. In contrast, the NPs *praca zmianowa* and *zmianowa praca* 'shift work' have the greatest difference of all stimuli NPs here, but with the N+A condition being more typical – the last data pair to the right.

Regardless of the different linearization in PL, the CS translations of both NPs with the most extreme differences in typicality would be in A+N linearization: *těžká hlava* 'heavy head' and *směnová práce* 'shift work' or *práce na směny* (literally 'work in shifts'). This is the case for virtually all CS translations of the stimuli NPs. For some stimuli NPs, alternative CS translations are possible with genitive constructions instead of the A. If informants have entered an alternative translation with a genitive construction, it was counted as correctly translated, because we assume that the sense was understood. The same applies for cases such as *śmierć przyszła / przyszła śmierć* 'future death', where the adjective *przyszła* 'future' might as well be a verb '[fem.] came' in the past tense in PL. Thus, if the Czech informants entered *přišla smrt* 'death came', it was counted as correct.

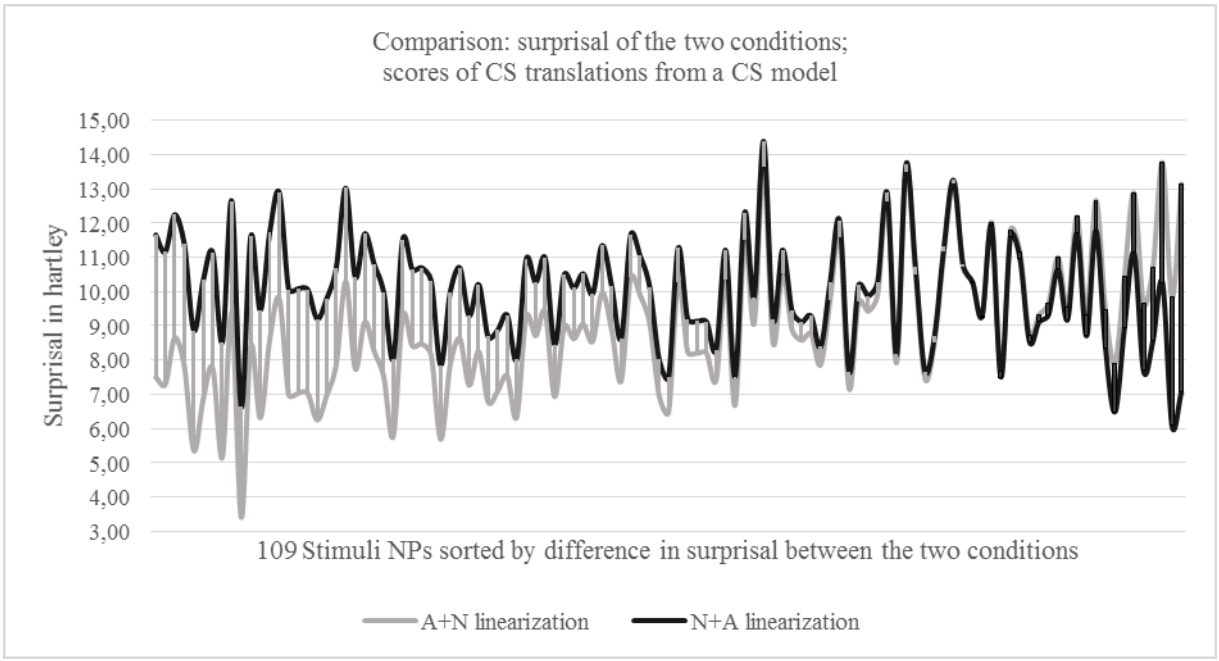

Fig. 2) Comparison of surprisal scores of the closest CS translations of the PL stimuli, obtained from a CS LM. The NPs are sorted by difference of surprisal scores. The two rightmost NP pairs are *plná hodina* / *hodina plná*[3] '(a) full hour' and *zevní dělání* / *dělání zevní* 'external making' (closest translation of the orginal PL *zewnętrzne działanie* 'external activity'), for which the N+A linearization has the biggest difference in surprisal compared to the same NP with A+N linearization in the stimuli set. On the contrary, *privátní pokoj* vs. *pokoj privátní* 'private room' have the greatest difference in surprisal with a preference for the A+N linearization and are the leftmost NP pair in the graph.

The surprisal graphs in Fig. 2 display higher surprisal scores for most of the NPs in the N+A linearization, confirming our intuition that A+N is more usual in CS (lower surprisal scores). Only 15 of the 109 NPs in the N+A condition have lower surprisal scores, meaning that these bigrams occur more often in N+A than in the A+N linearization. Compared to Fig. 1), the surprisal values of the CS NPs are higher in general. This is due to the fact that the scores were extracted for the closest CS cognate translations, i.e. those that readers will use as a 'bridge' for understanding. For instance, we expect that when reading *środowisko naturalne*'natural environment', CS readers will draw back to *prostředí naturální* with a relatively high surprisal score (9.96 hartley) before they would enter the more common translation *životní prostředí*, for which the surprisal score would be lower (6.47 hartley). The A *naturální* is relatively infrequent in CS and hence less predictable, which leads to this higher surprisal score.

3.3. Predicting the overall processing difficulty of the NPs

In a first step, we summarize lexical and orthographic distance in an aggregated distance measure that should represent the processing difficulty on the cross-lingual dimension. This is done in such a way that we treat non-cognates (units with a lexical distance of 1) like items with an orthographic distance value of 100 %. In a next step, we include the second dimension – the dimension of predictability in context (from the reader's perspective) – represented by the sum of the surprisal scores of a NP. As a result of the difficulties that can be expected on both dimensions, we calculate an overall processing difficulty (henceforth referred to as "overall difficulty").
For instance, the overall difficulty of the NP *silna kobieta* and *kobieta silna* 'strong woman' is calculated as follows: *kobieta* and *žena* are non-cognates and make up ½ of the NP, which leads to a distance of 50 % already. Comparing *silna* and *silná* that have an orthographic distance of 10 %, we add 10 % of the remaining half of the NP to the 50 %. The aggregate distance score of *silna kobieta* and *silná žena* therefore is 55 %. This value is multiplied by the sum of the surprisal score of the two words for each of the two linearisations, leading to an estimated overall difficulty score of 3.72 for the A+N condition and 4.77 for the N+A condition. Hence we predict that the N+A condition will be more difficult for CS readers.

PL stimulus	**in relation to CS**		**Lex 1**	**Lex 2**	**Orth 1**	**Orth 2**	**Dist ø**	**Surp 1**	**Surp 2**	**Surp ∑**	**Difficulty**
silna kobieta	silná	žena	0	1	0.1	(1)	0.55	4.67	2.11	6.78	3.729
kobieta silna	žena	silná	1	0	(1)	0.1		4.2	4.49	8.68	4.774

Tab. 2) Example for the calculation of the predicted overall difficulty for the stimulus *silna kobieta* and *kobieta silna* for a Czech reader.

[3] The lower suprisal of *hodina plná* as opposed to *plná hodina* might be because the bigram *hodina plná* is frequently followed by a genitive NP, e.g. *hodina plná radosti* 'an hour full of joy' and therefore might occur more often in the corpus than *plná hodina*.

We expect that the somewhat higher predicted difficulty score for *kobieta silna* will manifest itself in less correct translations and/or higher processing times than for *silna kobieta*.

The following table shows the means of the possible predictors of processing difficulty for Czech readers when deciphering the PL NPs in both conditions.

Data of all NPs (n=109)	A+N condition (n=1293)	N+A condition (n=1296)
Mean lexical distance per NP	20.18 %	
Mean lexical distance As	21.1 %	
Mean lexical distance Ns	19.27 %	
Mean orthographic distance per NP	35.62 %	
Mean orthographic distance As	39.82 %	
Mean orthographic distance Ns	33.08 %	
Mean aggregate distance per NP	52.27 %	
Mean surprisal per NP	9.46 hartley	10.02 hartley
Mean calculated difficulty per NP	4.66	5.14

Tab. 3) Comparison of linguistic distance and surprisal scores together with the expected mean difficulty for the A+N vs. the N+A condition.

4. Experiments

The web-based experiments were carried out over the freely accessible website http://www.intercomprehension.coli.uni-saarland.de. After registering with a user account, informants are first asked to enter empirical information and data on their language background and skills. The informants are automatically assigned experiments in a language, depending on their language background (native language).

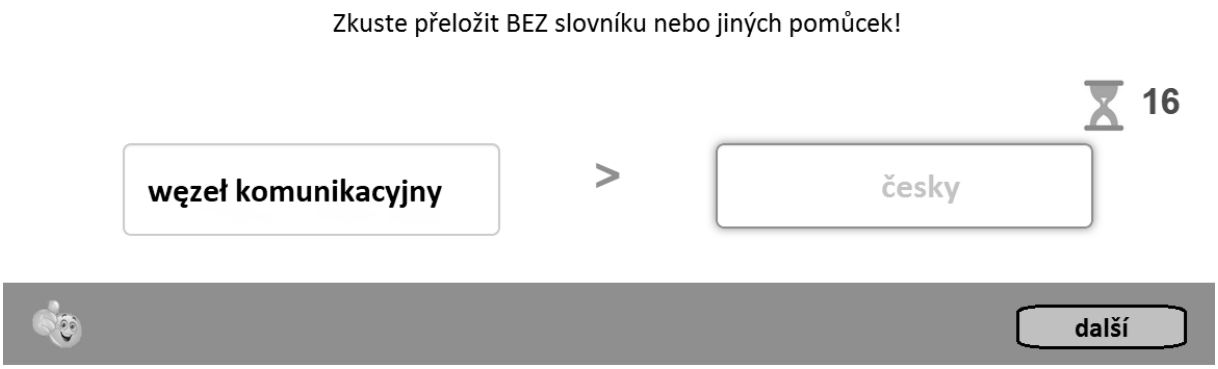

Fig. 3): Experimental screen. The prompt on top says 'Try to translate this WITHOUT a dictionary or the internet!'. The correct translation of the stimulus NP *węzeł komunikacyjny* 'transport hub' would be *komunikační uzel* in CS. The emoticon shows thumbs up, because the previous NP was translated correctly (feedback given for previous stimulus).

For each stimulus NP presented, informants have 20 seconds time to enter their translation, represented by the timer in the upper right corner. The system saves anything that was entered by an informant, regardless of whether an informant confirms the translation by pressing the return key (or clicking *další* 'next') or not. Immediate feedback is given in the shape of an emoticon at the left bottom of the page – a thumbs up for a correct translation or a sad face for a wrong or missing translation. There is a tolerance for lower/upper case and diacritical signs, i.e. if translations were entered without diacritics, but are correct otherwise, informants get positive feedback. Although the CS translations are optimal in A+N linearization, it was counted correct if participants entered the translation in N+A linearization. At the end of each experimental block, participants were presented their results on a brief statistics page, displaying the number of correct translations, total time for the block and average time per stimulus. All the results that are evaluated in this contribution have been checked manually for correctness – if participants have entered a correct solution that was not fed to the website beforehand, it was subsequently counted as correct.

5. Experimental results

Initial hesitation time (before typing), time spent typing, submission hesitation time (time between the last keystroke and pressing the *enter* or the *next* button) and total time spent on the stimulus (sum of the three other sub-measures) is recorded for each translation. For practical reasons, only the total time spent on the stimulus (henceforth referred to as 'processing time') is evaluated in this contribution. The experimental results are compared between the two conditions. The data sizes are not evenly distributed over the different NPs due to the different blocks of NPs that participants were assigned, meaning that within the data collected, the number of translations per NP ranges from 3 up to 17 translations in each of the conditions.

Data of all NPs (n=109)	A+N condition (n=1293)	N+A condition (n=1296)
Correctly translated NPs	49.5 %	41.63 %
Correctly translated: only As	66.51 %	61.6 %
Correctly translated: only Ns	63.57 %	61.99 %
Mean processing time of all NPs	10080 ms	10037 ms
Mean processing time of correctly translated NPs	8534 ms	8430 ms

Tab. 4) Comparison of experimental results: correctly translated stimuli and their mean processing time in the A+N vs. the N+A condition.

When comparing the number of completely correctly translated NPs (both words correct) in the two conditions, more correct translations can be observed for the NPs with A+N linearization (49.5 %) than for those with N+A linearization (41.63 %), confirming the hypothesis that difficulty in the A+N linearization is lower for Czech readers. The mean total processing times of correctly translated NPs differ only to a minimal extent between the two conditions: 8534 ms (SD = 4004.01 ms) in the A+N condition vs. 8430 ms (SD = 4022.8 ms) in the N+A condition and cannot be considered significant (t(638) = .5123).

5.1. Correlation with predictions

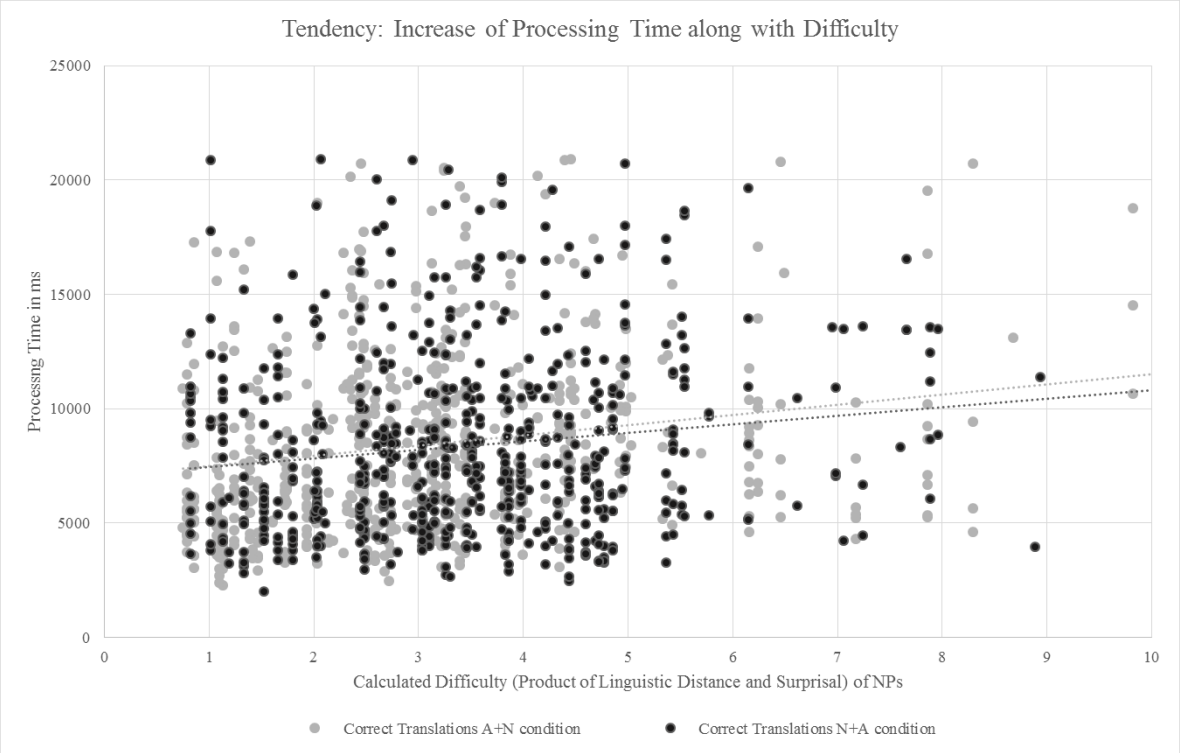

Fig. 4): Processing time in ms for all correct translations in relation to the calculated overall difficulty. Comparison between the A+N (light grey) and N+A (dark grey) condition. Outliers with a difficulty of more than 10 are not displayed for reasons of visualisation.

The graph in Fig. 4) shows the processing times for the correctly translated NPs in relation to the predicted overall difficulty of the NPs. As stated above, the diagram reflects somewhat lower surprisal scores for the A+N condition in general. The tendency of an increase in processing time with growing difficulty is observable, although only very vague ($r(615) = .194$ for A+N and $r(638) = .259$ for N+A).

In order to avoid the influence of the different data sizes for each of the NPs, we pick out the 30 most representative NPs from the data collected – those for which at least 10 translations were collected for each condition. As explained in section 3.1.1. and 3.1.2., all of the NPs have the same linguistic distance in both conditions and differ only in word order and consequently in their surprisal scores. We hypothesized that if the different linearization does have an influence on successful intelligibility, the results of the two conditions should differ in relation to their surprisal. This moderate tendency is displayed in Fig. 5).

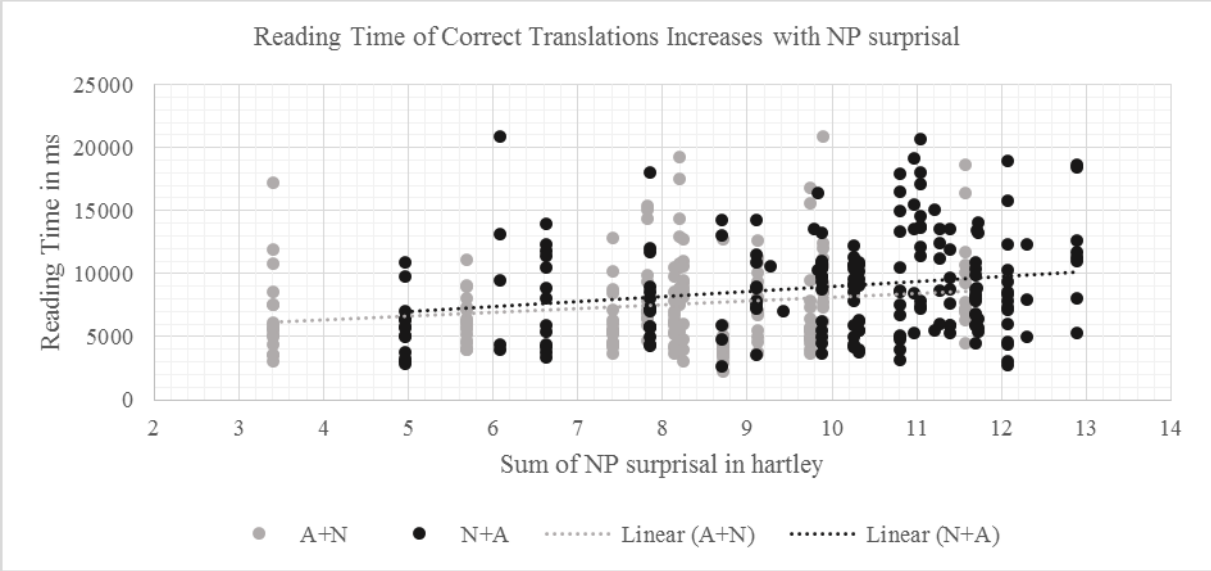

Fig. 5): Processing times of the correct translations of the most representative 30 NPs from the dataset in relation to the NP surprisal for the A+N (light grey) and the N+A condition (dark grey).

Similarly to the data shown in Fig. 4), there is a tendency observable among the most representative NPs: the higher the surprisal, the higher is the processing time. However, this tendency is relatively vague ($r(179) = .169$ for A+N and $r(198) = .218$ for N+A) .

Results of the most representative NPs (n=30)	A+N condition (n=428)	N+A condition (n=482)
Correctly translated NPs	42.06 %	41.08 %
Correctly translated: only As	42.06 %	40.87 %
Correctly translated: only Ns	44.63 %	41.7 %
Mean processing time of all NPs	10202 ms	10366 ms
Mean processing time of correctly translated NPs	7585 ms	8936 ms
Mean surprisal per NP	8.74 hartley	10.04 hartley
Mean aggregate distance of NPs	52 %	
Calculated mean difficulty of NPs	4.80	5.42

Tab. 5) Comparison of experimental results: correctly translated stimuli and their mean processing time in the A+N vs. the N+A condition.

When comparing the number of completely correctly translated NPs (both words correct) in the two conditions on the set of the 30 most representative NPs, the difference between the two conditions becomes only minimal. The mean processing times of the correct translations also differ somewhat between the two conditions, with a slight advantage for the A+N condition.

Correlation of processing time and calculated overall processing difficulty

We postulated that the overall processing difficulty of the NPs for Czech readers is the product of linguistic distance and surprisal. According to this hypothesis, there should be a correlation between the actual intelligibility scores of the stimuli and their difficulty. We analysed the correlations (Person's r) of processing times and the possible predictors: lexical distance of NPs, orthographic distance of NPs, total distance of NPs, surprisal of NPs in both conditions, and overall difficulty of NPs in both conditions (resulting from the four other predictors). Although the correlations between the processing time and the possible predictors are not high, it could be observed that overall difficulty is the best predictor, having a correlation of r = .259 (p < .001) for the A+N condition and .194 (p < .001) for the N+A condition. All correlations are given in Tab. 6).

	Lex. dist. NPs	Orth. dist. NPs	Aggr. dist. NPs	Surprisal NPs	Difficulty NPs
Processing time A+N	.190 (p < .001)	.140 (p < .001)	.245 (p < .001)	.121 (p = .002)	.259 (p < .001)
Processing time N+A	.118 (p = .003)	.096 (p = .017)	.157 (p < .001)	.105 (p = .009)	.194 (p < .001)

Tab. 6) Correlations of processing times in both conditions with the predictors lexical distance of NPs, orthographic distance of NPs, aggregate distance of NPs, surprisal of NPs in both conditions, and difficulty of NPs in both conditions.

For reasons of precision, it turned out to be reasonable to also evaluate correctly translated As and Ns separately. The results are displayed in Fig. 6) and 7).

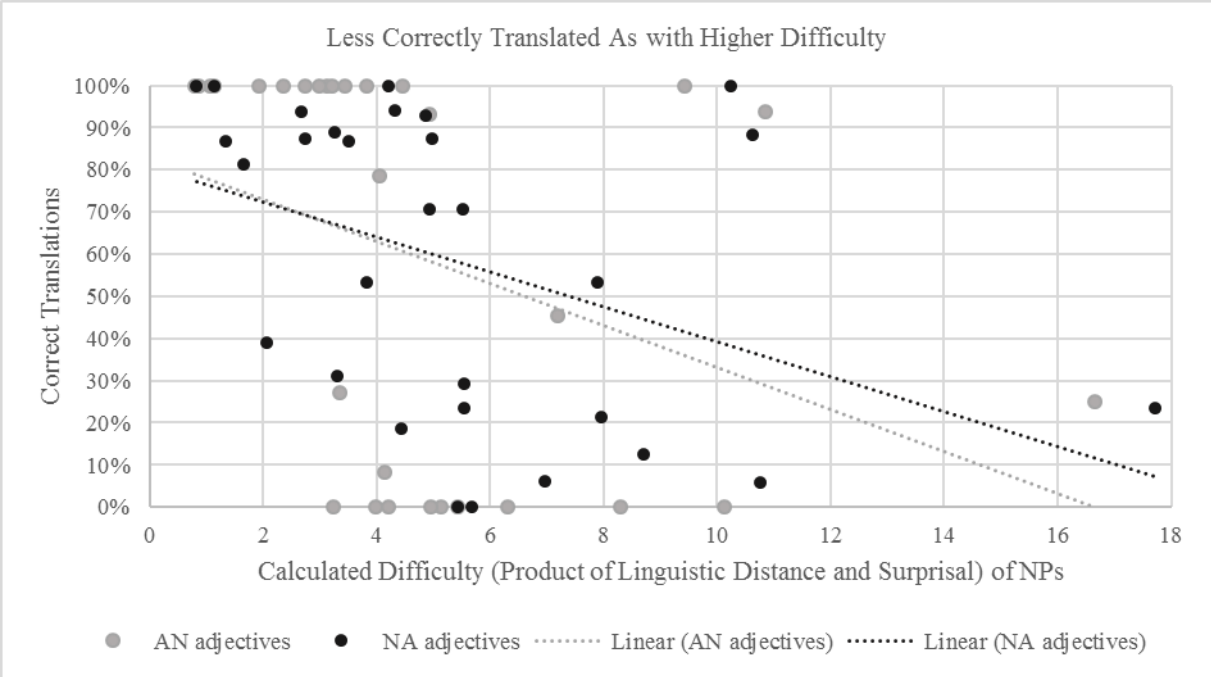

Fig. 6): Viewing As separately: Percentage of correctly translated As from 30 NPs with the most representative datasets. The percentages are set into relation with their overall difficulty. The overall difficulty is the product of linguistic distance of the As and the surprisal scores of the NPs.

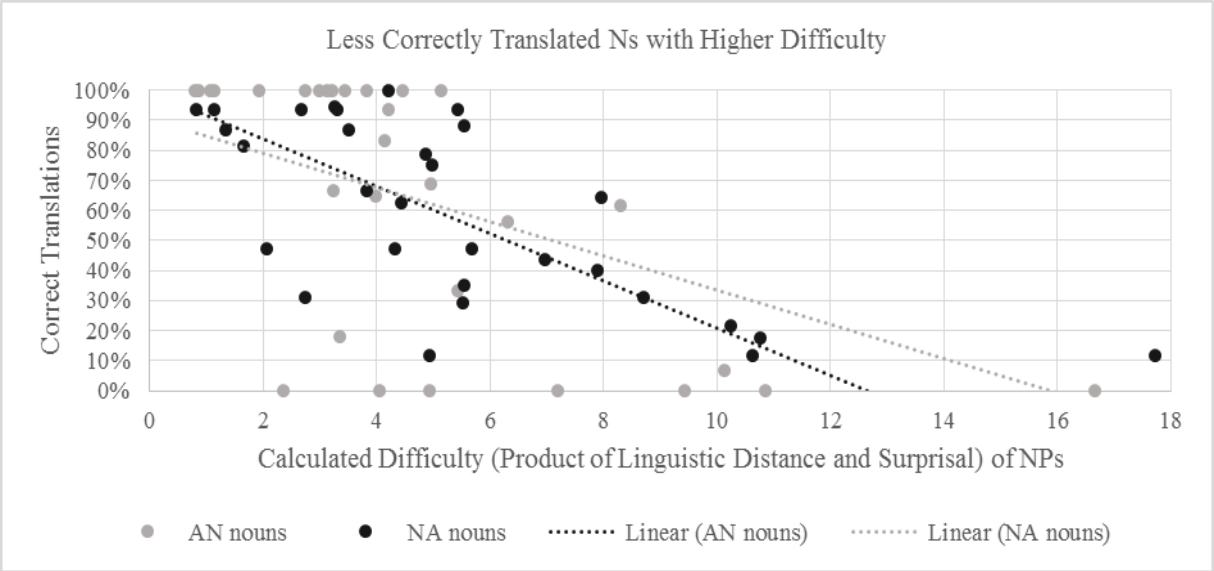

Fig. 7): Viewing Ns separately: Percentage of correctly translated Ns from 30 NPs with the most representative datasets. The percentages are set into relation with the expected overall difficulty. The overall difficulty is the product of the linguistic distance of the Ns and the surprisal scores of the NPs.

Many NPs are combinations of a relatively easy to understand word and another difficult word. Not distinguishing between Ns and As would make the evaluation more imprecise. When viewing As and Ns separately, the difference between the two conditions decreases with more correct translations for the A+N condition (66.51 % As and 63.57 % Ns) than for the N+A condition (61.6 % As and 61.99% Ns). The fact that Ns were translated correctly more often than As is not surprising, considering the fact that linguistic distance of the Ns is lower than that of the As (see section 3). The data in Fig. 6) and 7) show a comparison of As (Fig. 6) and Ns (Fig. 7) between the two conditions in relation to the estimated overall difficulty. Indeed, the data display the tendency that both surprisal and linguistic distance have an influence on the intelligibility, as the intelligibility scores decrease with a higher overall difficulty value that summarizes surprisal and linguistic distance.

6. Conclusion

The hypothesis that the N+A linearization in PL NPs causes an additional processing difficulty for Czech readers in an intercomprehension scenario could partly be confirmed in this experiment. When comparing the correct translations given by the informants, the two conditions differ with slightly better results for the A+N condition (49.5 % correctly translated NPs) than for the N+A condition (41.63 % correctly translated NPs). The hypothesis could not be confirmed in terms of processing times – those are slightly, but not significantly, higher for the A+N condition both for all translations (difference between means: 43 ms) and for only the correct translations (difference between means: 104 ms). However, it has to be kept in mind that processing time is only evaluated for the correctly translated stimuli andinformants might have taken longer to think about and enter a correct answer rather than entering a random wrong answer more quickly or no answer at all. The hypothesis that processing time is higher for the N+A condition holds when comparing the processing times of the correct translations in the most representative 30 NPs. Consequently, the postnominal attribute condition causes more difficulties for Czech readers when reading PL.

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. The full stimuli lists with distance measures and surprisal scores are made available upon request under http://www.coli.uni-saarland.de/%7Etania/incomslav.html.

Thanks

We wish to thank Varvara Obolonchykova for her support in the automatic calculation of orthographic distances and correlations. We owe special thanks to Magda Telus for the consultations on the translations of the NPs and Dietrich Klakow for training the language model on the Czech National Corpus.

References

Broda, B., Piasecki, M., 2011. Parallel, Massive Processing in SuperMatrix – a General Tool for Distributional Semantic Analysis of Corpora, in: *International Journal of Data Mining, Modelling and Management*.
Čermák, F., Rosen, A., 2012. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 13(3): 411–427
Cetnarowska, B., A. Pysz, and H. Trugman. 2011. Accounting for some flexibility in a rigid construction. In P. Bański, B. Łukaszewicz, M. Opalińska and J. Zaleska (eds.), Generative investigations: syntax, morphology and phonology, 24–57. Newcastle upon Tyne: Cambridge Scholars Publishing.
Cetnarowska, B., 2013. The representational approach to adjective placement in Polish. *Linguistica Silesiana* 34: 8–22. ISSN 0208-4228
Crocker, M., Demberg V., Teich, E., 2015. Information Density and Linguistic Encoding (IDEAL). In: *Künstliche Intell* (2016) 30, pp.77-81. doi:10.1007/s13218-015-0391-y
Gooskens, C. 2013. Methods for measuring intelligibility of closely related language varieties. In: Robert Bayley, Richard Cameron, and Ceil Lucas C. (eds.) Handbook of Sociolinguistics.
Jaeger T. F., Tily, H., 2011. On language utility: processing complexity and communicative efficiency. Wiley *Interdiscip Rev Cogn Sci* 2(3): 323–335
Heinz, C. 2008. Semantische Disambiguierung von *false friends* in slavischen L3: die Rolle des Kontexts. *Zeitschrift für Slawistik* 54: 145-166.
Jágrová, K. (forthcoming). The Role of Different Factors for the Intelligibility of Written Polish for Czech Readers. FDSL 12, 2016
Jágrová, K., Stenger, I., Marti, R., Avgustinova, T., 2017. Lexical and Orthographic Distances between Czech, Polish, Russian, and Bulgarian – a Comparative Analysis of the Most Frequent Nouns, in: *Olomouc Modern Language Series* (5): 401–416
Levenshtein, V. 1966. Binary codes capable of correcting deletions, insertions, and reversals. Cybernetics and Control Theory 10. pp. 707–710
Levy, R., 2008. Expectation-Based Syntactic Comprehension. *Cognition* 106 (3), 1126–1177.

Corpora

Czech National Corpus: Srovnávací frekvenční seznamy. 2010. Accessed January 01, 2016. http://ucnk.ff.cuni.cz/srovnani10.php.
Czech National Corpus: InterCorp (version 9). Accessed February 03, 2017. http://treq.korpus.cz/
Lista frekwencyjna. 2016. Grupa Technologii Językowych G4.19 Politechniki Wrocławskiej. http://www.nlp.pwr.wroc.pl/narzedzia-i-zasoby/zasoby/lista-frekwencyjna, Accessed 15/09/2016

Appendix: NP Stimuli

Stimuli were automatically displayed in random order in blocks of 36-37 NPs. Within a block, each NP was displayed only in one of the linearizations, i.e. the informants did not see the same NP twice in different linearizations. Not the CS translation equivalents are given here, but the closest CS translations, i.e. those cognates that orthographic distances and surprisal scores were calculated for. In many cases they can be equal to translation equivalents. Non-cognates are marked with a light grey background, false friends are marked with a dark grey background. Sorted by difference in surprisal between the two conditions, starting with the NP that is most likely to appear in A+N linearization when compared to N+A linearization for Czech readers.

Stimulus NPs PL	**In relation to CS**		**Lex**	**Orth**	**Aggr dist**	**∑ Surp**		**Difficulty**	
						A+N	**N+A**	**A+N**	**N+A**
prywatny pokój	privátní	pokoj	0	0.27	0.27	7.50	11.66	2.02	3.15
następny minister	nastupující	ministr	0	0.34	0.34	7.29	11.15	2.48	3.79
poważna uwaga	vážná	úvaha	0	0.57	0.57	8.65	12.25	4.93	6.98
komunikacyjny węzeł	komunikační	uzel	0	0.38	0.38	7.83	11.39	2.97	4.33
cały dzień	celý	den	0	0.50	0.50	5.36	8.87	2.68	4.43
wewnętrzny głos	vnitřní	hlas	0	0.64	0.64	6.87	10.34	4.40	6.61
ogólna siła	všeobecná	síla	0.5	0.25	0.63	7.76	11.14	4.85	6.96
główne miasto	hlavní	město	0	0.54	0.54	5.15	8.52	2.78	4.60
dodatkowy punkt	dodatkový	punkt	0	0.08	0.08	9.37	12.64	0.75	1.01
wielki pan	velký	pan	0	0.25	0.25	3.41	6.63	0.85	1.66
szczególny wzgląd	neobvyklý	vzhled	0.5	0.50	0.75	8.40	11.60	6.30	8.70
polityczny system	politický	systém	0	0.22	0.22	6.31	9.46	1.39	2.08
chory mężczyzna	chorý	muž	0	0.47	0.47	8.59	11.73	4.04	5.51
zmianowa praca	směnová	práce	0	0.43	0.43	9.79	12.89	4.21	5.54
ciemna noc	temná	noc	0	0.21	0.21	6.98	10.05	1.47	2.11
rządowa akcja	vládní	akce	0.5	0.40	0.70	7.03	10.09	4.92	7.06
określony procent	určité	procento	0.5	0.13	0.57	7.03	10.06	3.97	5.68
biały dom	bílý	dům	0	0.42	0.42	6.25	9.20	2.62	3.86
poszczególne ziemie	jednotlivé	země	0.5	0.42	0.71	6.97	9.79	4.95	6.95
bliski koniec	blízký	konec	0	0.29	0.29	7.87	10.69	2.28	3.10
trudna sytuacja	trudná	situace	0	0.23	0.23	10.29	13.02	2.37	2.99
szybkie pieniądze	rychlé	peníze	0.5	0.39	0.70	7.74	10.42	5.38	7.24
jedyne dziecko	jediné	děcko	0	0.30	0.30	9.12	11.69	2.73	3.51
zielony miesiąc	zelený	měsíc	0	0.39	0.39	8.25	10.80	3.22	4.21
różne cele	různé	cíle	0	0.33	0.33	7.51	9.97	2.48	3.29
nowy rok	nový	rok	0	0.19	0.19	5.76	8.03	1.09	1.53
światowa wojna	světová	vojna	0	0.41	0.41	9.28	11.52	3.80	4.72
droga matka	drahá	matka	0	0.25	0.25	8.42	10.66	2.10	2.66
prosta praca	prostá	práce	0	0.19	0.19	8.47	10.70	1.61	2.03
piękna twarz	pěkná	tvář	0	0.47	0.47	8.15	10.32	3.83	4.85
europejska komisja	evropská	komise	0	0.34	0.34	5.68	7.85	1.93	2.67
polski język	polský	jazyk	0	0.18	0.18	7.86	9.98	1.42	1.80
znany powód	známý	důvod	0.5	0.40	0.70	8.63	10.70	6.04	7.49
pewna ręka	pevná	ruka	0	0.28	0.28	7.26	9.29	2.03	2.60
kolejny raz	další	krát	2		2.00	8.24	10.21	16.49	20.41
silna kobieta	silná	žena	0.5	0.10	0.55	6.78	8.68	3.73	4.77
ostatni okres	poslední	období	2		2.00	7.09	8.87	14.18	17.74
wspólna część	společná	část	0	0.62	0.62	7.54	9.31	4.67	5.77
długi czas	dlouhý	čas	0	0.48	0.48	6.33	8.02	3.04	3.85
otwarty środek	otevřený	střed	0	0.70	0.70	9.28	10.94	6.49	7.66
małe oko	malé	oko	0	0.13	0.13	8.71	10.27	1.13	1.33
potrzebny przepis	potřebný	předpis	0	0.25	0.25	9.41	10.97	2.35	2.74
średni wiek	střední	věk	0	0.49	0.49	6.93	8.45	3.39	4.14
liczne programy	četné	programy	0.5	0.00	0.50	8.97	10.49	4.49	5.24
złota rada	zlatá	rada	0	0.20	0.20	8.62	10.10	1.72	2.02

wschodnia gmina	východní	komuna	0	0.47	0.47	9.05	10.52	4.25	4.94
wojskowa decyzja	vojenská	rozhodnutí	0.5	0.51	0.76	8.55	9.93	6.45	7.50
naturalne środowisko	naturální	prostředí	0	0.48	0.48	9.96	11.34	4.78	5.45
jasna informacja	jasná	informace	0	0.15	0.15	8.89	10.17	1.33	1.53
narodowe państwo	národní	stát	1	0.44	1.44	7.39	8.62	10.64	12.42
zła zasada	zla	zásada	0	0.21	0.21	10.41	11.64	2.19	2.44
gospodarczy związek	hospodářský	svazek	0	0.45	0.45	9.89	11.04	4.45	4.97
dana chwila	daná	chvíle	0	0.27	0.27	9.07	10.14	2.45	2.74
międzynarodowy projekt	mezinárodní	projekt	0	0.25	0.25	6.96	8.02	1.74	2.01
dobry człowiek	dobrý	člověk	0	0.33	0.33	6.49	7.51	2.14	2.48
publiczna droga	veřejná	dráha	0.5	0.40	0.70	10.25	11.27	7.18	7.89
krótkie słowo	krátké	slovo	0	0.33	0.33	8.24	9.20	2.72	3.03
czarne drzwi	černé	dveře	0	0.65	0.65	8.20	9.13	5.33	5.94
miejskie prawo	městské	právo	0	0.42	0.42	8.20	9.11	3.44	3.83
obecny problem	současný	problém	1	0.07	1.07	7.42	8.30	7.95	8.89
młody poseł	mladý	posel	0	0.25	0.25	10.34	11.21	2.58	2.80
stare miejsce	staré	místo	0	0.37	0.37	6.68	7.52	2.47	2.78
krajowa władza	krajová	vláda	0	0.36	0.36	11.49	12.30	4.14	4.43
polska sprawa	polská	záležitost	1	0.08	1.04	9.07	9.84	9.43	10.23
francuski ojciec	francouzský	otec	0	0.39	0.39	8.51	9.19	3.32	3.59
duże zdanie	dlouhé	věta	1	0.58	1.58	10.55	11.22	16.67	17.72
obca rodzina	cizí	rodina	1	0.14	1.14	8.88	9.43	10.12	10.75
czerwona woda	červená	voda	0	0.38	0.38	8.58	9.12	3.26	3.46
istotna prawda	důležitá	pravda	0.5	0.17	0.59	8.76	9.29	5.13	5.43
ciężka głowa	těžká	hlava	0	0.64	0.64	7.86	8.38	5.03	5.36
ogromna firma	ohromná	firma	0	0.11	0.11	9.74	10.25	1.07	1.13
dziwna myśl	divná	mysl	0	0.27	0.27	11.57	12.07	3.12	3.26
podstawowa zmiana	podstatná	změna	0	0.34	0.34	7.16	7.65	2.44	2.60
porządkowe czynności	pořádkové	činnosti	0	0.40	0.40	9.70	10.18	3.88	4.07
wolna szkoła	volná	škola	0	0.36	0.36	9.42	9.87	3.39	3.55
daleki kraj	daleký	kraj	0	0.08	0.08	9.88	10.26	0.79	0.82
poprzednia możliwość	popřední	možnost	0	0.43	0.43	12.61	12.89	5.42	5.54
finansowy wynik	finanční	výsledek	0.5	0.44	0.72	7.92	8.16	5.70	5.87
miesięczne premie	měsíčné	prémie	0	0.29	0.29	13.48	13.72	3.91	3.98
szerokie usta	široké	ústa	0	0.31	0.31	10.51	10.72	3.26	3.32
kwiatowy miód	květový	med	0	0.53	0.53	7.45	7.65	3.95	4.06
właściwa pomoc	vlastní	pomoc	0	0.38	0.38	8.53	8.70	3.24	3.30
konsumpcyjny lód	konzumní	led	0	0.42	0.42	11.16	11.32	4.69	4.76
niewielki rynek	nevelký	rynek	0	0.22	0.22	13.14	13.25	2.89	2.92
głęboka rzecz	hluboká	věc	1	0.43	1.43	10.72	10.78	15.34	15.42
możliwy stan	stav	možný	0	0.41	0.41	10.27	10.27	4.21	4.21
dawny udział	dávný	úděl	0	0.53	0.53	9.29	9.28	4.92	4.92
prawowity rząd	pravoplatný	vláda	0.5	0.50	0.75	12.00	11.92	9.00	8.94
przyszła śmierć	přišla	smrt	0	0.60	0.60	7.66	7.50	4.60	4.50
prawdziwy bóg	pravdivý	bůh	0	0.53	0.53	11.77	11.61	6.24	6.15
niska góra	nízká	hora	0	0.39	0.39	11.14	10.97	4.35	4.28
gotowa poprawka	hotová	oprava	0	0.40	0.40	8.73	8.54	3.49	3.41
specjalny wniosek	speciální	návrh	1	0.33	1.17	9.31	9.11	10.85	10.62
rosyjski numer	ruské	číslo	0.5	0.63	0.82	9.64	9.32	7.86	7.60
naukowy temat	naučné	téma	0	0.44	0.44	11.00	10.64	4.84	4.68
wysoka osoba	vysoká	osoba	0	0.13	0.13	9.54	9.17	1.24	1.19
ciekawe pytanie	zajímavé	ptaní	0.5	0.36	0.68	12.19	11.71	8.29	7.96
północny świat	severní	svět	1	0.70	1.70	9.30	8.71	15.82	14.80
amerykańska grupa	americká	grupa	0	0.25	0.25	12.66	11.77	3.17	2.94
własne życie	vlastní	žití	0	0.65	0.65	9.47	8.36	6.16	5.43
fizyczne ciało	fyzické	tělo	0	0.63	0.63	7.90	6.49	4.98	4.09
społeczna strona	společenská	strana	0	0.33	0.33	10.45	8.95	3.45	2.95
ważna pani	vážná	pani	0	0.31	0.31	12.88	11.13	3.99	3.45

zagraniczny bank	zahraniční	banka	0	0.26	0.26	9.65	7.69	2.51	2.00
zielona mięta	zelená	máta	0	0.38	0.38	10.69	8.57	4.06	3.26
podobny przypadek	podobný	případ	0	0.29	0.29	5.89	8.55	1.71	2.48
państwowa ustawa	státní	stanova	1	0.50	1.50	13.77	10.27	20.65	15.40
pełna godzina	plná	hodina	0	0.34	0.34	9.83	6.09	3.34	2.07
zewnętrzne działanie	zevní	dělání	0	0.66	0.66	13.14	7.07	8.68	4.67
Mean			0.19	0.36	0.52	8.86	9.96	4.65	5.14