

Dissertationen aus der
Philosophischen Fakultät der Universität des Saarlandes

Reading Polish with Czech Eyes: Distance and Surprisal in Quantitative, Qualitative, and Error Analyses of Intelligibility

Klára Jágrová



universaar

Universitätsverlag des Saarlandes
Saarland University Press
Presses Universitaires de la Sarre

 OLMS

The OLMS logo consists of a stylized white 'O' with a horizontal bar through its center, followed by the letters 'OLMS' in a bold, sans-serif font.

Klára Jágrová

Reading Polish with Czech Eyes:
Distance and Surprisal in Quantitative,
Qualitative, and Error Analyses of Intelligibility



universaar

Universitätsverlag des Saarlandes
Saarland University Press
Presses Universitaires de la Sarre

 **OLMS**

© 2022 *universaar*

Universitätsverlag des Saarlandes

Saarland University Press

Presses Universitaires de la Sarre

Postfach 151141, 66041 Saarbrücken

D 291

Dissertation vom 11.07.2019

Dissertation zur Erlangung des akademischen Grades eines Doktors der Philosophie
der Philosophischen Fakultät der Universität des Saarlandes

Dekan: Univ.-Prof. Dr. Heinrich Schlange-Schöningen

Berichterstatter: Prof. Dr. Tania Avgustinova, Prof. Dr. Roland Marti

ISBN: 978-3-86223-322-9 gedruckte Ausgabe

ISBN: 978-3-86223-323-6 Onlineausgabe

Satz: Bernhard Schiestel

Umschlaggestaltung: Julian Wichert

Bibliografische Informationen der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.dnb.de> abrufbar.

Gedruckt auf FSC-zertifiziertem Papier.

Eine Publikation in Zusammenarbeit zwischen *universaar* und dem Georg Olms Verlag.

CONTENTS

CONTENTS	III
TABLES	IX
FIGURES	XIII
LIST OF ABBREVIATIONS	XV
PREFACE	1
Previous Publications	4
Thesis Overview	5
CHAPTER I: BACKGROUND & INTRODUCTION	7
1. The INCOMSLAV Infrastructure	7
1.1. Languages in Focus and INCOMSLAV Publications	9
1.2. Regular Orthographic Correspondences Between PL-CS and BG-RU	9
1.2.1. Hand-crafted correspondences inferred from traditional linguistic assumptions	11
1.2.2. Results of the application of correspondences	11
1.3. Similarity of Linguistic Encoding	14
1.3.1. Lexical distance	16
1.3.2. Orthographic distance	19
1.3.3. Other distance measures in the literature	22
1.4. Asymmetry in Cross-Lingual Intelligibility	23
1.4.1. Conditional entropy	23
1.4.2. Word adaptation surprisal	27
1.5. Surprisal and Context	27
1.6. Context in Intercomprehension	31
2. Thesis Focus: Modelling Linguistic Phenomena of PL for Czech Readers	33

CHAPTER II: COOPERATIVE TRANSLATION EXPERIMENT . . .	35
3. Experimental Setup	35
4. Quantitative Analysis of Written Results and Comparison of Conditions	37
4.1. Hypotheses	37
4.2. Stimuli	39
4.2.1. Linguistic distance of stimuli	41
4.2.2. Surprisal of stimuli	42
4.3. Experimental Conditions: Modification of Stimuli on Different Linguistic Levels	42
4.3.1. Conditions with non-combined modifications	43
4.3.2. Conditions with combined modifications	44
4.4. Results	47
4.4.1. Evaluation of the translations per word	47
4.4.2. Comparison between the conditions	49
4.5. Summary	52
5. Qualitative Analysis	53
5.1. Readers' Strategies	55
5.1.1. Leaving unknown words open and trying to infer them from the context	55
5.1.2. Recognition order as indicator for difficulty	57
5.1.3. Reading again and pronouncing differently	61
5.2. Source of Successful Transfer	61
5.2.1. Inference processes from non-standard CS	62
5.2.2. Inference from languages other than CS	62
5.3. Knowledge of Non-Cognates and Awareness of False Friends	71
5.4. Over-Transfer from Languages Other Than CS	76
5.5. Distrust in Obviously Understandable Words	77
5.6. Revision After Having Already Named the Correct Answer .	85
5.7. Handling Unfamiliar PL Orthography	86
5.7.1. Handling PL diacritics	87
5.7.1.1. Respondents pronounce letters correctly	92
5.7.1.2. Respondents ignore diacritics and pronounce stimulus as if without diacritics	95

5.7.1.3.	Respondents move diacritics to another suitable letter in the word	96
5.7.2.	Handling unfamiliar PL digraphs	97
5.8.	Talking About Grammar	101
5.9.	Problems Caused by Differences in Government Patterns ..	103
5.10.	Problems Caused by Different Prepositions	104
5.11.	Summary	108
CHAPTER III: ON-LINE EXPERIMENTS		111
6.	Hypotheses	111
6.1.	Pronunciation-Based Orthographic Distance	111
6.2.	Surprisal as a Predictor Variable for Context in Intercomprehension	114
7.	Empirical Base	115
7.1.	Online Experiments	115
7.2.	Overview of Experiments and Data Collected	120
7.3.	Participants	121
8.	The Principle of the Closest Possible Translation	122
9.	Measures not Considered	123
10.	Scoring Policy Throughout the Experiments	124
11.	Relevant Statistical Methods in Brief	125
CHAPTER IV: FREE TRANSLATION OF WORDS WITHOUT CONTEXT		127
12.	Cognates with Regular PL-CS Orthographic Correspondences	127
12.1.	Orthographic Distance of the Stimuli	128
12.2.	Results	129
12.3.	Correlations	130
12.4.	Error Analysis	132
12.5.	Summary	135
13.	The 100 Most Frequent PL Nouns	135
13.1.	Results and Correlations	136
13.2.	Error Analysis	138
13.3.	Summary and Outlook	139

CHAPTER V: FREE TRANSLATION OF NPS	141
14. Adjectival Modification in PL	142
14.1. Hypothesis	145
14.2. Method	145
14.3. Distance of the Stimuli	145
14.4. Total Difficulty of the Stimuli	146
14.5. Results	148
14.5.1. Intelligibility	148
14.5.2. Processing time	150
14.5.1. Wrong recognition of POS	151
14.6. Summary	152
14.7. Digression: PL NPs Presented to German Readers	153
14.7.1. Hypothesis	153
14.7.2. Stimuli	153
14.7.3. Orthographic distance	154
14.7.4. Surprisal in context	154
14.7.5. Results	155
14.7.5.1. Relation between intelligibility and orthographic distance ...	156
14.7.5.2. Relation between intelligibility and surprisal	157
14.7.5.3. Mean processing time	157
14.7.5.4. Wrong recognition of POS	158
14.7.5.5. Lexical interferences	160
14.7.6. Summary	161
14.8. Comparison of PL NP Results Between Czech and German Readers	162
CHAPTER VI: TRANSLATION OF TARGET WORDS IN CONTEXT	165
15. Highly Predictable Target Words in Cloze Translation Task ..	165
15.1. Experiment Design	165
15.2. Stimuli	166
15.2.1. Closest translation	167
15.2.2. Surprisal	169
15.2.3. Linguistic distance	170

15.3.	Scoring of Responses	171
15.4.	Results	172
15.4.1.	Comparison: with vs. without context	172
15.4.2.	Different lexical categories of target words	177
15.4.2.1.	Cognates (C)	178
15.4.2.2.	Cognates in other contexts (C-OC)	180
15.4.2.3.	Non-cognates (NC)	181
15.4.2.4.	False friends (FFs)	182
15.4.3.	Analysis of wrong responses	187
15.4.3.1.	Differences in government pattern	188
15.4.3.2.	Ln interferences	188
15.4.3.3.	(Perceived) morphological mismatches	189
15.5.	Correlations and Model	198
15.6.	Summary and Discussion	200
16.	The Impact of Random Context on the Understanding of Particular Words in Sentences from the Cooperative Translation Task	201
16.1.	Method	202
16.2.	Baseline Experiments: Cloze Probabilities in Monolingual Context	205
16.2.1.	Design	205
16.2.2.	Results	205
16.3.	Scoring of Responses	207
16.4.	Results: Target Words at Random Position	208
16.4.1.	Comparison: Types of errors	210
16.4.2.	Comparison: Target words with vs. without context	214
16.5.	Summary	215
CHAPTER VII: CONCLUSION AND OUTLOOK		217
	Conclusion in German Deutsche Zusammenfassung	227

APPENDICES	239
A 1. Alignment Matrices of the PL and CS Alphabets	239
A 1.1 For the Calculation of Trad LD	239
A 1.2 For the Calculation of Pron LD	241
A 2. Questionnaire on Sociodemographic Data	243
A 3. Instruction for the Participants in the Cooperative Translation Experiments	244
A 4. Intelligibility of Stimuli in the Different Experiments	245
A 4.1. Stimuli with Regular PL-CS Correspondences	245
A 4.2. Most Frequent PL Nouns	259
A 4.3. Free Translation of NPs	262
A 4.3.1. PL NPs for CS readers with the most representative data	262
A 4.3.2. PL NP stimuli for German readers	263
A 4.4. Highly Predictable Target Words	264
A 5. Target Words in Highly Predictive Context Categorised as FFs	269
A 5.1. False Friends that Are also Cognates – FF-C	269
A 5.2. False Friends that Are Cognates in Another Context – FF-OC	270
A 5.3. False Friends that Allow for Correct Associations – FF-A ..	271
A 5.4. False Friends – FF	272
A 6. Monolingual Cloze Tests	273
A 6.1. Task in Monolingual Cloze Tests	273
A 6.2. Stimuli	273
A 7. Correlations and Statistical Models	281
A 7.1. Intelligibility of the 100 Most Frequent PL Ns	281
A 7.2. Intelligibility of NPs for Czech Readers – AN Condition ...	282
A 7.3. Intelligibility of NPs for Czech Readers – NA Condition ...	283
A 7.4. Intelligibility of Target Words in Highly Predictive Context ..	284
A 7.5. Model for Intelligibility of Target Words in Highly Predictive Context	286
A 7.6. Model for Intelligibility of the Target Words Without Context	287
REFERENCES	289

TABLES

Table 1:	Experiments conducted with the linguistic levels examined . . .	8
Table 2:	PL-CS word sets used for the extraction of cognate stimuli (cf. Fischer et al., 2015, p. 118).	10
Table 3:	Most frequent PL-CS transformations applied on the different lists (Fischer et al., 2015, pp. 121)	13
Table 4:	Example of lexical asymmetry: non-cognates vs. cognate translations	17
Table 5:	Comparison: the CS and the PL alphabet	19
Table 6:	PL-CS sound correspondences (Heinz & Kuße 2015, pp. 70-72)	19
Table 7:	Example for the calculation of trad LD.	20
Table 8:	Trad LD for PL-CS: internationalisms, Pan-Slavic vocabulary, and Swadesh list	22
Table 9:	Calculation of conditional entropy of a cognate pair	24
Table 10:	Vowel character entropies for the PL-CS language pair	25
Table 11:	Calculation of word adaptation surprisal of a cognate pair	27
Table 12:	PL for Czech readers: comparison of distance and intelligibility in the literature	33
Table 13:	Sentences in cooperative translation experiment and possible translations	40
Table 14:	Data sizes: translated words obtained from informants in each condition	46
Table 15:	Example: recognition order of words within stimuli sentences	58
Table 16:	Expected Ln transfer bases for certain words within the stimuli	63
Table 17:	Ln reading skills indicated by respondents and (partial) non-cognates requiring an Ln transfer base	63
Table 18:	Words containing q , CS cognate translations and applicable correspondences	89
Table 19:	Words containing q , CS cognate translations and applicable correspondences	89
Table 20:	Words containing q and q and their various pronunciations by respondents	91
Table 21:	Words containing z and its various pronunciations by respondents	93

Table 22: Words containing <i>ś</i> and its various pronunciations by respondents	95
Table 23: Recognition of cognates might fail due to wrong division of syllables	98
Table 24: Words containing the digraph <i>cz</i> and its various pronunciations by respondents	100
Table 25: Words containing the digraph <i>rz</i> and its various pronunciations by respondents	100
Table 26: Words containing the digraph <i>sz</i> and its various pronunciations by respondents	100
Table 27: PL-CS alignments that cost 0 for pron LD	112
Table 28: Correspondences that Czech readers are likely to handle through exposure to SK	113
Table 29: Calculation of trad LD of a cognate pair in comparison to pron LD	113
Table 30: Overview of experiments conducted, sorted by topic and section in this thesis	120
Table 31: Overview of main demographic characteristics in the experiments	121
Table 32: Closest translation principle demonstrated on a PL stimulus sentence	123
Table 33: Word length and orthographic distance of cognates with regular PL-CS correspondences	129
Table 34: Intelligibility of cognates with regular PL-CS correspondences	129
Table 35: Correlations: intelligibility of cognates with regular PL-CS correspondences and predictors	130
Table 36: Model for the intelligibility of cognates with regular PL-CS correspondences	131
Table 37: Verbs with PL-CS correspondences and nouns that respondents translated them with	132
Table 38: Correlations: intelligibility of the 100 most frequent PL nouns and predictors	137
Table 39: Model for the intelligibility of the 100 most frequent PL nouns	138

Table 40: Ln interferences among the translations of the most frequent PL nouns	138
Table 41: Example for the calculation of overall difficulty for NP stimuli	147
Table 42: Comparison of linguistic distance and surprisal scores: AN vs. NA	147
Table 43: Intelligibility and mean processing time of NPs: AN vs. NA	148
Table 44: Correlations of predictors with intelligibility of NPs: AN vs. NA	148
Table 45: Regression models for intelligibility of the NPs: AN vs. NA . .	149
Table 46: Correlations: processing time and predictors in AN vs. NA . .	151
Table 47: NP translation experiment with German readers: AN vs. NA . .	156
Table 48: Cases with wrongly recognised POS: AN vs. NA	158
Table 49: Lexical L1/Ln interferences (EN, FR, ES, BCS)	160
Table 50: Comparison: NP translation experiments with Czech vs. German respondents	163
Table 51: Calculation steps for all surprisal-related variables	169
Table 52: Calculation steps for all orthographic distance-related variables of a sentence	170
Table 53: Overview of (sub-)categories of target words included into the statistical model	172
Table 54: Intelligibility of target words with vs. without context in the different categories	177
Table 55: Overview of target non-cognates that offer associations with correct CS translations	182
Table 56: Comparative overview of FF-C with vs. without context . . .	185
Table 57: Comparative overview of FF-OC with vs. without context . .	186
Table 58: Comparative overview of FF-A with vs. without context . . .	187
Table 59: Comparative overview of FF with vs. without context	187
Table 60: PL -ę mistaken for a plural marker or a marker of other grammatical forms ending with -e or -ě in CS	189
Table 61: Target words in instrumental case mistaken for words ending in -a	191
Table 62: Target nouns that differ in grammatical gender between PL and CS	192

Table 63: Target verbs mistaken for nouns with vs. without context . . .	196
Table 64: Stimuli from the cooperative translation task + 10 other sentences presented in the cloze translation experiments with random context	203
Table 65: Results of the cloze translation task with target words at random position	208
Table 66: Comparison: types of wrong responses in different stimuli sets	212
Table A 1: Alignment matrix used for the calculation of trad LD	239
Table A 2: Alignment matrix used for the calculation of pron LD	241
Table A 3: Stimuli with regular PL-CS correspondences, their intelligibility and processing times	245
Table A 4: Intelligibility of the 100 most frequent PL nouns and their predictor variables	259
Table A 5: Intelligibility of the 30 NPs with the most representative data and predictors	262
Table A 6: NP stimuli presented to German readers, their correct DE and closest GER translations	263
Table A 7: Sentences with highly predictable target words in cloze translation experiments	265
Table A 8: Target words classified as FF-C	269
Table A 9: Target words classified as FF-OC	270
Table A 10: Target words classified as FF-A	271
Table A 11: Target words classified as FF-FF	272
Table A 12: Results from the monolingual cloze test (PL)	274
Table A 13: Results from the monolingual cloze test (CS)	277
Table A 14: Regression models: intelligibility of the 100 most frequent PL nouns	281
Table A 15: Regression models: NP translation experiments – AN condition	282
Table A 16: Regression models: NP translation experiments – NA condition	283
Table A 17: Correlation matrix (Pearson's r): intelligibility of target words with and without context and the different predictors . . .	284

Table A 18: Legend containing the abbreviations used in Table A 16 . . . 285

Table A 19: Regression models: Intelligibility of highly predictable target words 286

Table A 20: Regression models: Intelligibility of the target words without context 287

FIGURES

Figure 1: The INCOMSLAV project: overview and workflow 7

Figure 2: Applicability of regular cross-lingual correspondences 12

Figure 3: Lexical distance among the 100 most frequent nouns 18

Figure 4: Orthographic distance of cognate pairs without and with transliterations 21

Figure 5: Expected transformations of unknown characters in a PL stimulus by a Czech reader 26

Figure 6: Transformation of unknown characters observed in cooperative translation experiments 26

Figure 7: Trigrams as they could occur in a PL corpus during training 29

Figure 8: Setup of the cooperative translation experiment in pairs . . . 36

Figure 9: Screen during the cooperative translation experiment 36

Figure 10: Visualisation of a stimulus set in the cooperative translation experiments 46

Figure 11: Results for all conditions in the cooperative translation experiments 49

Figure 12: Correct translations (incl. paraphrases) per condition in relation to total distance 50

Figure 13: Total distance and surprisal among the response categories . . 51

Figure 14: Trad LD vs. pron LD vs. phonetic distance 112

Figure 15: Experimental screen in the free translation experiments . . 117

Figure 16: Experimental screen in the NP translation experiments . . . 117

Figure 17: Experimental screen in the cloze translation experiments . . 118

Figure 18: Brief statistics shown to respondents after a completed experiment 119

Figure 19: Surprisal of the closest CS translation vs. a good CS translation of a PL stimulus 122

Figure 20: Comparison of non-normalised pron LD and pron LD	128
Figure 21: Correlation: intelligibility of cognates with PL-CS correspondences with pron LD	131
Figure 22: Correlation: intelligibility of the most frequent nouns with trad LD vs. pron LD	136
Figure 23: Comparison of typicality of the NP stimuli: AN vs. NA (Jágrová, 2018, p. 130)	142
Figure 24: Typicality of closest CS translations of the NPs: AN vs. NA (Jágrová, 2018, p. 130)	144
Figure 25: Difference in expected processing effort between the two linearisations	147
Figure 26: Correlation: processing time in ms for all correct translations and calculated overall difficulty (Jágrová, 2018, p. 135)	150
Figure 27: Correlation: correct answers and orthographic distance calculated towards DE	156
Figure 28: Correlation: correct answers and orthographic distance calculated towards GER	156
Figure 29: Intelligibility of target words with vs. without context	174
Figure 30: Surprisal graph of a sentence with a low-surprisal target word	175
Figure 31: Surprisal graph of a sentence with a high-surprisal target word	175
Figure 32: Comparison: target cognates (C without C-IB) with vs. without context	179
Figure 33: Comparison: target non-cognates (incl. C-OC, FF-OC, FF-A, and FF) with vs. without context.	181
Figure 34: Comparison: target false friends (FF-C, FF-OC, FF-A, and FF) with vs. without context	184
Figure 35: Comparison: infinitive verb forms with vs. without context	195
Figure 36: Comparison: target words in random context vs. without context	214
Figure A 1: Questionnaire on sociodemographic data on the experiment website (EN version)	243
Figure A 2: Task as displayed to the respondents in the cooperative translation experiment	244
Figure A 3: Instruction as presented to respondents in the monolingual cloze test (EN version)	273

LIST OF ABBREVIATIONS

A	adjective
accu	accusative case
BEL	Belarusian
BG	Bulgarian
BCS	Bosnian, Croatian, Serbian
C	cognate
C-IB	cognate identical in the base form
CNC	Czech National Corpus
C-OC	cognate only in another context
CS	Czech
CSK	Czechoslovak
dat	dative case
DE	German
DK	Danish
EN	English
fem	feminine
FF	false friends
FR	French
gen	genitive case
GER	Germanic
HR	Croatian
impf	imperfective aspect
instr	instrumental case
i.p.m.	instances per million
L1	native language
L2	second language
L3	third language
lex	lexis
Lex dist	lexical distance
LD	Levenshtein distance
LM	language model
Ln	acquired language
loc	locative case

Lx	unknown language
MK	Macedonian
masc	masculine
N	noun
n/a	not applicable
NC	non-cognate
neut	neuter
NL	Dutch
nom	nominative case
NOR	Norwegian
norm	normalised
NP	noun phrase
ns	not significant
orth	orthography
P	pair of respondents
perf	perfective aspect
pers	person
pl	plural
PL	Polish
POS	part of speech
pron LD	pronunciation-based Levenshtein distance
RU	Russian
SD	Standard deviation
SE	Standard error
sg	singular
SK	Slovak
SL	Slovene
SR	Serbian
surp	surprisal
SWE	Swedish
total dist	total distance (pron LD and lex dist in one variable)
trad LD	traditionally calculated Levenshtein distance
UK	Ukrainian
V	verb
voc	vocative case
WAS	word adaptation surprisal

PREFACE

I would like to take the opportunity and express deep gratitude to my supervisors Tania Avgustinova and Roland Marti for giving me the opportunity to commit the last 4.5 years to the topic that has fascinated me since my Russian lessons at school – the ability to understand an unknown but related foreign language – a phenomenon referred to as intercomprehension. Both were always reliable and supportive in professional and interpersonal matters, always willing to provide time and share their absolute expertise both in the field of Slavic linguistics and linguistic theory in general. This possibility was provided to me in the framework of the INCOMSLAV project as part of the collaborative research centre (CRC) 1102: *Information Density and Linguistic Encoding* at Saarland University, funded by the German Research Foundation (DFG, project ID 232722074). Besides the work on the thesis that was closely bound to the project, I was given the opportunity to gather some experience in teaching. In order to experience the subject of my research on myself I took the opportunity to attend language courses and spend time abroad, and to present our research at local and international conferences that resulted in a number of joint publications.

Throughout our work on the project, I always enjoyed working together and exchanging ideas with my colleague and fellow Slavic linguistics PhD student within the project, Irina Stenger, whose dissertation covers the role of orthography in intercomprehension between the Slavic languages written in Cyrillic script. I also owe thanks to Dietrich Klakow and Andrea Fischer who covered the computer scientific part of our project.

I would like to thank all the brilliant colleagues – fellow PhD students, postdocs, and professors – within our CRC for their specialist feedback, be it in the working group on experiment design, the language modelling working group or at presentations at our joint retreats and PhD days. I especially would like to thank Elke Teich for her professionalism, energy, and balance while managing the CRC. I owe special thanks to our coordinators Marie-Ann Kühne, Patricia Borrull Enguix, and Olena Steshenko for their commitment and reliability in all organization-related questions. The same applies to Ekaterina Klüh at the international office who maintains the contact of partner institutions and who has always been a great support in the mobility programs I was able to take part in.

Also, I wish to thank all the student assistants from the field of computational linguistics and computer science that have guided our project throughout these 4.5 years: Aniko Kovač for corpus management, Varvara Obolonchykova for her support in the automatic calculation of orthographic distances, Ali Shah

for the automatic application of cross-lingual correspondences, Muhammad Ahmad for taking care of the experiment website and implementation of experiments, Marius Mosbach for creating the Levenshtein distance and conditional entropy calculation & visualisation tool, Adam Kusmirek for his support and maintenance of the language modelling tool LM GUI, as well as Ayan Majumdar and Hasan Alam for their maintenance, further debugging, and implementation of additional features to the experiment website.

I owe a depth of gratitude to my colleagues at the Slavistics department: Magda Telus, Lucija Šarčević, and Juliana Stoyanova not only for their translations and consultations on the stimuli and the experiment website within the project, but also for welcoming me at their language classes. Special thanks go to Magda Telus for taking the time and checking not only the Polish parts of this thesis, but also going through the typesetting and layout. I also wish to express my gratitude to Bistra Andreeva who has provided me with her expertise in phonetics that was essential for the evaluation of the pronunciations of Czech respondents in the cooperative translation experiments. Apart from the colleagues at the Slavistics department, I wish to thank Olga Petukhova for her advice on the analysis of the protocols from the cooperative translation experiment.

Quite a number of the data presented here was collected during my research stays at our partner institution, the Charles University in Prague. I thank Magda Ševčíková for her advices and critical feedback on the experiments and I thank the colleagues at ÚFAL for sharing their workspace with me during the cooperative translation experiments in Prague. In this context I would like to give regards to Grégoire Labbé who works on a related topic and welcomed me in his course Slovene for students of Croatian at Charles University. I also thank not only the many students at Charles University but also family, friends, and the crowdsourced respondents who participated in the experiments discussed here, but also Iva Poláková-Šolcová and Ivan Rynda at the Faculty of Humanities at Charles University who allowed me to conduct the experiments as a part of their lectures. My friend Miroslav Kiselovský contributed a great part – thank you for your professional transcription of the many hours of audio files from the cooperative translation experiment!

Without Holger Kuße who supervised my state examination thesis on Slavic intercomprehension at TU Dresden and forwarded the posting about the PhD vacancy at Saarland University to me, this all would not have been possible. I was able to acquire solid background knowledge of the Slavic languages in his lectures on historical-comparative linguistics. I value him highly for his engagement and work in many other areas of Slavic linguistics (discourse analysis, cultural studies etc.) and I am thankful that I was welcomed at the Slavic linguistics colloquium at TU Dresden in the last years.

Lastly, the most special of thanks go to my partner Arne for his moral and practical support, patience, dedication, positive distraction, the excellent food, and the music.

As for the topic of this thesis, I chose the intercomprehension scenario in which written Polish is presented to Czech readers for various reasons. Not only was this language-reader combination among those on which the focus of the INCOMSLAV project was laid, but also it turned out to be a scenario with a prominent discrepancy of lexical closeness of the languages on the one hand, but a relatively high orthographic distance on the other. Since PL was shown to be an outlier among the Slavic languages with regard to orthography, it was highly interesting to me how Czech readers would handle this divergent orthography and other distinctive features of the two languages. Last but not least, with CS being my L1 (although not my best language), I can take on the Czech readers' perspective and provide an understanding of the results that was necessary, e.g. in the error-analytical parts of this thesis. At the same time, I had the opportunity to refresh my long ago, incidentally acquired basics of Polish and attend Polish courses both at Saarbrücken and at Warsaw Universities, where I obtained a B2/C1 certificate in summer 2017.

The datasets discussed are made available upon request, allowing interested readers to fully reproduce all results in this thesis or carry out their own analyses. Should you be interested in the data or spot any errors in the data, feel free to contact me at jagrovaklara@gmail.com.

Despite careful verification of the content of all cited links at the time of publication, the author cannot guarantee the continued existence of links used as sources. To view the linked content at the time of defending the work, the use of the Internet Archive's Wayback Machine (<https://archive.org/web/>) is helpful.

Furthermore, no liability is accepted for the content of external links. The operators of linked sites are solely responsible for their content and hosting.

Previous Publications

Section 1.2 summarises the methods and findings for the Polish-Czech language pair published in

Fischer, A., Jágrová, K., Stenger, I., Avgustinova, T., Klakow, D., & Marti, R. (2015). An orthography transformation experiment with Czech-Polish and Bulgarian-Russian. In B. Sharp, W. Lubaszewski & R. Delmonte (Eds.), *Natural Language Processing and Cognitive Science 2015 Proceedings* (pp. 115-126). Venezia: Libreria Editrice Cafoscarina.

Section 1.3 contains methods and findings of the paper

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 J. Edmonds & M. Janebová (Eds.), *Proceedings of the Olomouc Linguistics Colloquium 2016: Olomouc Modern Language Series* (Vol. 5, pp. 401-416). Olomouc: Palacký University. <http://olinco.upol.cz/wp-content/uploads/2017/06/olinco-2016-proceedings.pdf>

Examples and distance measures published in

Stenger, I., Jágrová, K., Fischer, A., Avgustinova, T., Klakow, D., & Marti, R. (2017). Modelling the impact of orthographic coding on Czech-Polish and Bulgarian-Russian reading intercomprehension. *Nordic Journal of Linguistics*, 40(2), 175-199. doi:10.1017/S0332586517000130

are picked up on in the definitions of conditional entropy and word adaptation surprisal (section 1.4.) and in the distance measures of stimuli with applicable cross-lingual correspondences in section 12.

Section 14 in CHAPTER V contains large parts of the paper

Jágrová, K. (2018). processing effort of Polish NPs for Czech readers – A+N vs. N+A. In W. Guz & B. Szymanek (Eds.), *Canonical and non-canonical structures in Polish. Studies in linguistics and methodology* (Vol. 12, pp. 123-143). Lublin: Wydawnictwo KUL.

Section 4 has been submitted for publication and is pending approval as

Jágrová, K. (2016, December). The role of different factors for the intelligibility of written Polish for Czech readers. Paper presented at FDSL 12, Berlin

Note: As of 17 March 2019, section 15 was accepted for publication in a shortened version and from a more results-oriented perspective as

Jágrová, K., & Avgustinova, T. (2019). Intelligibility of highly predictable Polish target words in sentences presented to Czech readers. To appear in *Proceedings of CICLing: International Conference on Intelligent Text Processing and Computational Linguistics*.

and its preprint is made available under http://www.coli.uni-saarland.de/~tania/ta-pub/CICLing_preprint_Jagrova_Avgustinova_2019.pdf with the respective data supplement under https://www.coli.uni-saarland.de/%7Etania//ta-pub/CICLing2019_PL_sentences_resource.xlsx. Parts of section 1.6, 15.2, 15.4.1, and 15.4.3. are identical with parts of the preprint.

Some of the wording might unavoidably overlap between the individual publications listed here and this thesis, or in other sections than mentioned.

Thesis Overview

In CHAPTER I, I first introduce the thesis in the context of the project workflow in section 1. I then summarise the methods and findings from the project publications about the languages in focus. There I also introduce the relevant concepts and terminology viewed in the literature as possible predictors of intercomprehension and processing difficulty. CHAPTER II presents a quantitative (section 4) and a qualitative (section 5) analysis of the results of the cooperative translation experiments. The focus of this thesis – the language pair PL-CS – is explained and the hypotheses are introduced in section 6. The experiment website is introduced in section 7 with an overview over participants, the different experiments conducted and in which section they are discussed. In CHAPTER IV, free translation experiments are discussed in which two different sets of individual word stimuli were presented to Czech readers: (i) Cognates that are transformable with regular PL-CS correspondences (section 12) and (ii) the 100 most frequent PL nouns (section 13). CHAPTER V presents the findings of experiments in which PL NPs in two different linearisation conditions were presented to Czech readers (section 14.1-14.6). A short digression is made when I turn to experiments with PL internationalisms which were presented to German readers (14.7). CHAPTER VI discusses the methods and results of cloze translation experiments with highly predictable target words in sentential context (section 15) and random context with sentences from the cooperative translation experiments (section 16). A final synthesis of the findings, together with an outlook, is provided in CHAPTER VII.

CHAPTER I: BACKGROUND & INTRODUCTION

1. The INCOMSLAV Infrastructure

The work conducted for this thesis is part of the first phase of the INCOMSLAV project – Mutual Intelligibility and Surprisal in Slavic Intercomprehension – at Saarland University. The project itself is part of the DFG-funded collaborative research centre (CRC) 1102: Information density and linguistic encoding and is in its second phase since 11/2018. I will briefly present the project infrastructure in this section.

As one of the projects within the CRC dealing with the phenomena of linguistic variation, the first project phase was settled in the research domain of receptive multilingualism. Its general objective was to information-theoretically and empirically examine the mechanisms by which languages encode and decode information. It is aimed at modelling the performance of Slavic readers in understanding a text in another unknown but closely related language.

Two language pairs for which a relatively high degree of mutual intelligibility is expected were chosen to this end: Polish and Czech – hereafter referred to as PL and CS (both West Slavic, both using the Latin script) – and Bulgarian and Russian – hereafter referred to as BG and RU (South and East Slavic, both using Cyrillic script). This thesis focuses on the PL-CS language pair. The project covered linguistic phenomena on the levels of orthography, morphology, lexis, and syntax which were tested in web-based experiments and correlated to results from language modelling. The project workflow and infrastructure between the three areas of linguistic phenomena, modelling, and experiments is shown in.

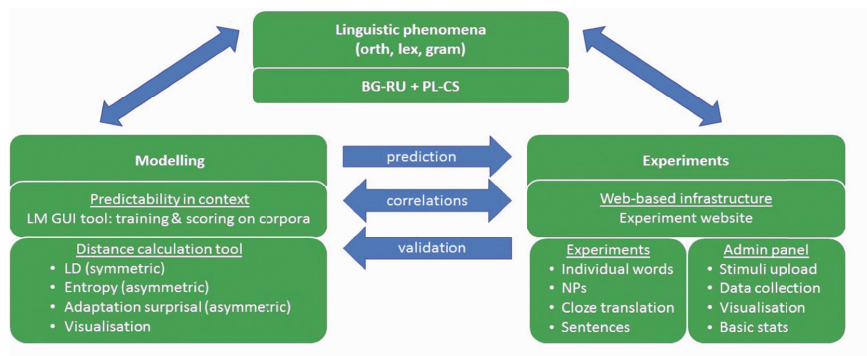


Figure 1: The INCOMSLAV project: overview and workflow.

This thesis represents the entire project workflow, implemented with the focus on one language-reader combination: written PL presented to Czech native speakers. All parts of the project cycle are covered by this thesis and build on each other coherently, since insights about the role of the phenomena on the different linguistic levels were examined in the different kinds of experiments: The context-free translation experiments with individual words cover the topic of orthography in intercomprehension. Besides orthography, morphological and word order features are systematically observed in the experiments with noun phrases (NPs, section 14). All three kinds of factors – orthography, morphology, and syntax – interplay in the sentence stimuli presented in the cooperative translation experiments (CHAPTER II) and in the cloze translation experiments (CHAPTER VI). Table 1 provides an overview of the different experiments and the relevant linguistic levels:

	Orthography	Morphology	Syntax	Word order
Individual words	X			
NPs	X	X		X
Sentences (cloze translation)	X	X	X	X
Sentences (cooperative)	X	X	X	X

Table 1: Experiments conducted with the linguistic levels examined.

The experiment website is one of the three main software resources that have emerged in the project. Details on the website are provided in section 7. Statistical language modelling was implemented with the help of the language modelling tool LM GUI (<https://lm.lsv.uni-saarland.de/>) developed as a resource at the CRC 1102 at Saarland University. It enables researchers to train different pre-defined types of statistical language models on corpora that can be loaded into the tool and saved for later. As of March 2019, the LM GUI is an internal resource. The method of how the tool serves the training of statistical language models (LMs) and how they can be applied to language material is explained in section 1.5 in detail. A tool for calculating orthographic distance and word adaptation surprisal (WAS, explained in section 1.4.2) of parallel word sets was developed in the project. It serves as a visualisation tool at the same time and is planned to be published as a resource in the near future.

Regarding terminology, there have been a number of concepts with slightly different nuances of meaning that the phenomenon of intercomprehension was referred to – *receptive multilingualism*, *semicomcommunication*, *mutual intelligibility*, *receptive bilingualism*. In this thesis, I will use the term intercomprehension to refer to all of them. The phenomenon of intercomprehension reveals a

robust human ability to understand related but unknown languages, without being able to use them actively, i.e. for speaking or writing (cf. Doyé, 2005, p. 7).

Gooskens & Swarte (2017, p. 125) distinguish between acquired and inherent/inherited (both terms appear in the paper) intelligibility, mainly because they investigated mutual intelligibility among the Germanic languages, including EN. While acquired intelligibility is associated with foreign language learning, inherent/inherited intelligibility assumes that the Lx has not been learnt before. In this thesis, only data from those respondents who have not indicated to have learnt PL throughout their lives is considered. In other words, only the inherited/inherent intelligibility is examined. It is practically relevant for all situations in which Czechs encounter the PL language, be it through media or through contact with native speakers of PL at the border area of the two neighbouring countries, in Poland or elsewhere in the world.

The core contribution of this thesis in the research field on intercomprehension are the methods and insights into two topics: first, the systematic analysis of the impact of predictive context in intercomprehension; and second, the assumed pronunciation of the Lx as a reflection of perceived linguistic distance and inner speech during the reading of the unknown but related code, resulting in a pronunciation-based orthographic distance measure (pron LD).

1.1. Languages in Focus and INCOMSLAV Publications

The main focus within the INCOMSLAV project are the four Slavic languages BG, CS, PL, and RU. The project aims at contributing further insights into receptive multilingualism among the selected Slavic languages by using original sources of the languages under focus. There were five fundamental INCOMSLAV publications. The results of three of them are summarised in the following subsections.

1.2. Regular Orthographic Correspondences Between PL-CS and BG-RU

The modern Slavic languages developed from a reconstructed parent language – referred to as Proto-Slavic or Common Slavic – to the modern varieties of BG, CS, PL, and RU (Schenker, 1993). There is a common base in the linguistic systems of the individual modern Slavic languages, which reflects the development from the common ancestor language – Proto-Slavic – in the course of several centuries (Carlton, 1991, p. 9) as a result of both linguistic and sociolinguistic factors.

The first project publication was

Fischer, A., Jágrová, K., Stenger, I., Avgustinova, T., Klakow, D., & Marti, R. (2015). An orthography transformation experiment with Czech-Polish and Bulgarian-Russian. In B. Sharp, W. Lubaszewski & R. Delmonte (Eds.), *Natural Language Processing and Cognitive Science 2015 Proceedings* (pp. 115-126). Venezia: Libreria Editrice Cafoscarina.

which focused on cross-lingual orthographic correspondences in the two Slavic language pairs PL-CS and BG-RU. The objective was to quantitatively validate traditional linguistic assumptions by applying orthographic correspondences on contemporary word material and to obtain suitable stimuli for experiments on the role of orthography in intercomprehension. Since the PL-CS results of this study were used as stimuli in the free translation experiments in section 12, the methods and findings will be summarised in the following.

We tested the automatic applicability of cross-lingual correspondences based on orthographic features of cognates within parallel lists that were available in digital format: Pan-Slavic vocabulary and internationalisms (both adapted from the lists on the EuroComSlav website) as well as Swadesh lists. The analysis was conducted on word lists instead of full texts in order to focus on the orthographic level only and exclude influences that are of morphological nature. All lists were slightly modified: Formal non-cognates (i.e. PL-CS *teraz – ted* ‘now’) were removed and formal cognates were added to the lists where the pairs consisted of non-cognates (i.e. *kobieta* ‘woman’ substituted by *žona* ‘wife’ in PL-CS *žona – žena*) if possible.

Two large word lists were added to obtain a statistically more representative effect: a set of homonyms (false friends) from Szałek & Nečas (1993) and an open-source digital version of a PL-CS dictionary containing more than 80,000 lexemes (Kazojć, 2010). Table 2 gives an overview of the PL-CS lists used for the extraction of transformable cognates together with the number of words per list.

Swadesh list	212
Panslavic list	455
Internationalism list	262
Homonyms	1,553
Dictionary	80,963

Table 2: PL-CS word sets used for the extraction of cognate stimuli (cf. Fischer et al., 2015, p. 118).

The lists contained verbs that were analysed in their infinitive forms in the PL-CS pair, while in the BG-RU lists they were replaced by their third person present tense forms, since there are no infinitive forms in BG.

1.2.1. Hand-crafted correspondences inferred from traditional linguistic assumptions

The orthographic correspondences should act as a substitute for the written representation of units of the Lx in the readers' L1 and reflect the main lines of the sound system evolution, from Common Slavic to the four individual modern Slavic languages in terms of "(i) development of vowels and consonants, (ii) development of specific sound combinations, and (iii) the metathesis of liquids" (Fischer et al., 2015, p. 117).

These correspondences were collected from traditional Slavic comparative literature: Bidwell (1963), Žuravlev (1974-2012), and Vasmer (1973). This resulted in a set of 81 correspondences for PL-CS (e.g., *iq:á, cz:č, ię:ě, gw:hv, hu:lou, dz:z*) and only 48 correspondences for BG-RU (*m:тъ, б:бл, ъ:y, u:ы, я:e, ла:оло etc.*) (Fischer et al., 2015, p. 117). The greater number of correspondences for the PL-CS pair suggests a greater orthographic diversity between PL and CS than between the other two languages. The correspondences were then applied on the parallel word lists with an algorithm and examined for how frequently they apply to the cognates in the word lists.

1.2.2. Results of the application of correspondences

The algorithm automatically classified the words into three categories: (a) *identical*, (b) *correctly transformed* by applying one or more correspondences (correspondences covering strings of characters were prioritised over single character correspondences), and (c) *untransformed* (when the set of correspondences could not cover the necessary transformations) (Fischer et al., 2015, p. 119). The diagrams in Figure 2 (Fischer et al., 2015, p. 120) visualise the different proportions of words in the three categories in both language pairs.

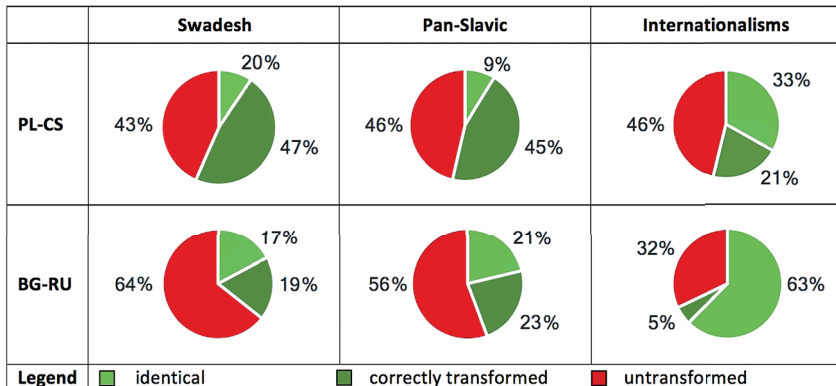


Figure 2: Applicability of regular cross-lingual correspondences.

In total, we obtained 3,404 PL-CS and 1,182 BG-RU word pairs¹ consisting of either identical words or words to which the cross-lingual correspondences apply.² These word pairs were used for further calculations of conditional entropy and word adaptation surprisal in Stenger, Jágrová et al. 2017 as explained in section 1.4. Some of the words from the category *correctly transformed* were used as stimuli in intercomprehension experiments – these are discussed in section 12.

A striking difference in the proportion of orthographically identical words could be observed between the language pairs: The maximum for PL-CS is only about 33% for the internationalisms, while about 63% of the internationalisms are identical in BG-RU. For all word sets, the share of identical words is greater for BG-RU than for PL-CS, which suggests a greater degree of mutual intelligibility for BG-RU than in the other pair. The percentage of identical words is highest for internationalisms in both language pairs, also because this list consists only of nouns, while the other lists contain adjectives and verb forms that would require additional morphological correspondences (these were later extracted in Fischer et al. 2016). Nevertheless, more words of PL-CS can be transformed in the Pan-Slavic (about 45%) and Swadesh list (about 47%) than in the other pair: The results for BG-RU in the Pan-Slavic list amount to only about 23%. The proportion of untransformed cognates remains relatively constant throughout the three lists for PL-CS, while for BG-RU about 64% of the Swadesh list and only 32% of the internationalisms could not be transformed (Fischer et al., 2015, pp. 120).

1 The resource was published under <http://www.coli.uni-saarland.de/~tania/incomslav.html>. An access code can be requested from the authors.

2 Duplicates were removed.

Table 3 displays the five most frequently applicable correspondences on words categorised as *correctly transformed* for each list. The total frequency of application is given next to the correspondences, examples are provided for each correspondence and list.

Swadesh	Pan-Slavic	Internationalisms	Homonyms	Dictionary
ć:t, 24	γ:ý, 45	a:á, 15	w:v, 307	w:v, 991
<i>dać—dát</i>	<i>dynia—dýně</i>	<i>bal—bál</i>	<i>wiec—věc</i>	<i>krowa—kráva</i>
γ:ý, 21	w:v, 42	a:e, 12	γ:ý, 175	ć:t, 663
<i>nowy—nový</i>	<i>woda—voda</i>	<i>linia—linie</i>	<i>wylot—výlet</i>	<i>prac—prát</i>
w:v, 20	ć:t, 37	w:v, 8	ć:t, 163	a:á, 515
<i>dwa—dva</i>	<i>bolec—bolet</i>	<i>kawa—káva</i>	<i>ćma—tma</i>	<i>para—pára</i>
a:á, 10	ł:l, 24	e:í, 5	a:á, 142	a:e, 353
<i>ja—já</i>	<i>zły—zły</i>	<i>talerz—talří</i>	<i>czara—čára</i>	<i>dusza—duše</i>
ł:l, 9	g:h, 20	ł:l, 4	ł:l, 111	γ:ý, 336
<i>cieply—teplý</i>	<i>głowa—hlava</i>	<i>kanal—kanál</i>	<i>łatka—látka</i>	<i>dym—dým</i>

Table 3: Most frequent PL-CS transformations applied on the different lists (Fischer et al., 2015, pp. 121).

Of course, the frequency of applicability in these lists strongly depends on the overall frequency of the characters constituting the correspondences in each word list. A large part of the most frequently applicable correspondences are those with a difference in diacritics: The correspondence $\gamma:\acute{y}$ was originally derived from a historical correspondence in word stems (e.g., *dým* – *dym* ‘smoke’), but it is even more frequent as a typical correspondence in adjective endings. Hence, it is frequent in all lists except in internationalisms (nouns only). Another correspondence that is frequent because of its occurrence in endings is $\acute{c}:t$. This morphological feature of infinitive verb forms is reflected in orthography and is frequently applicable in all lists, again except in internationalisms. This means that some of the cross-lingual orthographic features can be expanded to morphological features, since orthographic correspondences also apply to morphological units in the language pair. The correspondence $h:g$ is only frequent in Pan-Slavic vocabulary. Some of the frequent correspondences describe vowel changes, such as $a:e$ or $e:i$ – both of them apply to noun endings. As a result, there is a frequent applicability of correspondences concerning endings, there are letters that do not exist in the other alphabet, and there is tolerance of diacritical signs (cf. Fischer et al., 2015). This suggests that for

a Czech native speaker reading PL as an Lx, the knowledge of a number of orthographic correspondences might improve reading comprehension to quite some extent already.

The computer code for the implementation of the orthographic transformation rules between language pairs (by Andrea Fischer and Ali Shah) is provided in Fischer et al. (2015). The part for the BG-RU pair (alphabets, correspondences, special characteristics) can be found in the same publication.

1.3. Similarity of Linguistic Encoding

Several linguistic and extra-linguistic factors influence the successful disambiguation of unfamiliar linguistic code. How well reading intercomprehension functions, depends in the first place on the stimulus-decoder combination. In previous research on cross-lingual intelligibility of written text, the role of linguistic distance (lexical, orthographic, morphological, syntactic, phonetic) was investigated as a predictor for human performance in models of intelligibility for different language combinations and in different experimental settings (cf., for instance, Golubović & Goskens, 2015; Golubović, 2016; Gooskens, 2013; Heeringa et al., 2013; Heeringa et al., 2014). Thus, linguistic distance is supposed to reflect the (dis)similarity of two related codes: The smaller the linguistic distance, the more similar and mutually intelligible the two codes are and transfer of knowledge from an L1 (native language) to an Lx (unknown language) is possible.

In the second project publication,

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 J. Edmonds & M. Janebová (Eds.), *Proceedings of the Olomouc Linguistics Colloquium 2016: Olomouc Modern Language Series* (Vol. 5, pp. 401-416). Olomouc: Palacký University. <http://olinco.upol.cz/wp-content/uploads/2017/06/olinco-2016-proceedings.pdf>,

lexical and orthographic distances between all stimulus-reader combinations of BG, CS, PL, and RU with both untransliterated and transliterated cognate pairs were calculated. The study applied existing methods for determining lexical and orthographic distance between related languages as presented by Heeringa et al. (2013) who investigated the Germanic, Romance, and Slavic languages spoken in the EU. It was conducted for a verification of the findings of Heeringa et al. (2013) and for obtaining also distance measures in combination with RU, which Heeringa et al. (2013) did not include, since only official languages of the EU were subject to their study. In general, the methods for measuring linguistic distance throughout this thesis are by and large oriented on these methods.

The translation of words from one language into another is done manually in this thesis, following what I call the *principle of the closest possible translation*: If a cognate translation of a word is possible in at least one context, then this cognate is chosen. The cognate translations can be “pairs of words which have the same meaning in both languages only in some contexts” (Heeringa et al., 2013, p. 103) as well. Cognates are consequently defined as both real cognates and partial cognates. Following this method, I do not distinguish if cognate pairs are etymologically related or if they are loan words as long as they have a common root and share a meaning in at least one possible context (cf. Jágrová, Stenger, Marti & Avgustinova, 2017). This principle represents an intercomprehension situation in which the reader would be able to identify the meaning of a word in a given context. It also holds if the cognate translation chosen is archaic or used in non-standard or literary language only. Cognate translations were preferred over non-cognates even if the meaning of cognates overlapped only in an extremely narrow or obviously infrequent context, such as in

- PL *uwaga* ‘caution’, but also ‘consideration’ and CS *úvaha* ‘consideration’
- PL *ustawa* ‘law’, but also ‘statute’ and CS *ustanovení* ‘designation’, but also ‘statute’ (Jágrová, Stenger, Marti & Avgustinova, 2017).

In a number of cases, ideal translations would be different. The purpose of this analysis is to obtain measures of linguistic distance as predictors of human performance in translating these words. For the same reason, I forego the distinction between *main* translations and rather rare translations. Here, the focus lies merely on the understanding of linguistic code. The question is not how many different signifiers a concept has in the first place, but rather if readers are able to associate the signifier with the signified.

If a PL stimulus word can be translated with a CS cognate, it is assigned a lexical distance value of 0. If there is no suitable cognate translation in at least one possible context, a distance value of 1 is assigned. Translating stimuli and deciding for a translation that is a cognate or not turned out to be complicated in some cases. The decisions were first of all oriented on the results offered by the web application Treq (Vavřín & Rosen, 2015): If one of the words presented by Treq was a cognate, then the stimulus word was considered a cognate and the orthographic distance towards its closest CS form was calculated. Whenever there was more than one possible cognate translation, I choose the orthographically closest option (based on LD), for instance PL *šrodek* has two possible translations into CS: *střed* ‘middle’, ‘centre’ or *prostředek* ‘means’. Since *prostředek* has an LD of only 60% as opposed to *střed* that has an LD

of 71.42%, *prostředek* was chosen for the distance calculation towards *šrodek*. Details on this translation principle and how it is applied on sentence material later in the thesis are provided in section 8.

In Jágrová, Stenger, Marti & Avgustinova (2017), the most frequent BG, CS, PL, and RU nouns were extracted from frequency lists based on the respective national corpora of the individual languages. The most frequent nouns of PL were extracted from a frequency list published by the LT group of the Politechnika Wroclawska. According to Broda & Piasecki (2010), the frequency list was generated from large corpora with an overall size of 1.8 billion tokens, including the IPI PAN corpus, Korpus Rzeczpospolitej, Wikipedia (backup copy from early 2010) and a collection of large internet documents (Lista frekwencyjna. Grupa Technologii Językowych G4.19 Politechniki Wroclawskiej, 2016). I removed country-specific nouns such as *sejm* (lower house of the Polish parliament) from the source list. Obvious errors due to automated processing of the frequency list were corrected, e.g. *proca* ‘sling-shot’ was replaced by *procent* ‘percent’, because the abbreviation of *procent* was apparently mistaken for the genitive plural form of *proca* (*proc*). Since the result of this study was a list of the most frequent PL nouns that was then presented in the free translation experiments discussed in section 13, the methods and findings will be summarised in the following.

1.3.1. Lexical distance

The underlying assumption behind the method introduced by Heeringa et al. (2013) is that the intelligibility of a related L_x is, among other factors, influenced by the common share of cognates and their orthographic transparency. Lexical distance prevails when words in an L_x cannot be correlated to cognates in the reader’s L. The total number of non-cognates is normalised by the number of words in the material: It is determined as the percentage of non-cognates in a language pair and in a certain direction of reading, i.e. it can be asymmetric. Accordingly, the higher the lexical distance of material in a language pair is, the more difficult it should be for readers to understand texts in an L_x. Measurements of lexical distance were mostly applied within sets of words or on short texts (e.g. Heeringa et al., 2013).

The lexical asymmetry between BG, CS, PL, and RU in the lists often emerges not only between two languages, but in some cases, it may persist with the other languages as well. For instance, all languages examined in Jágrová, Stenger, Marti & Avgustinova (2017) share the cognates to PL *grupa* ‘group’:

CS *grupa* and RU *группа*³ (*gruppa*), and BG *група* (*grupa*). However, there is the CS word *skupina* ‘group’ in the list of the most frequent CS nouns that has no cognate translation in any of the other languages. The visualisation of the example in Table 4 (read: first column translated into all other columns) represents a situation in which Czech (and also Bulgarian and Russian) readers should understand PL *grupa* which has a lexical distance of 0, because it is a cognate (green background in Table 4). However, neither of the other readers are likely to understand CS *skupina* (lexical distance is 1), because it is a non-cognate (white background in Table 4).

PL	CS	BG	BG translit	RU	RU translit	ENG
<i>grupa</i>	<i>grupa</i>	<i>група</i>	<i>grupa</i>	<i>группа</i>	<i>gruppa</i>	<i>group</i>
CS	BG	BG translit	PL	RU	RU translit	ENG
<i>skupina</i>	<i>група</i>	<i>grupa</i>	<i>grupa</i>	<i>группа</i>	<i>gruppa</i>	<i>group</i>

Table 4: Example of lexical asymmetry: non-cognates vs. cognate translations.

Jágrová, Stenger, Marti & Avgustinova (2017) found asymmetries on the lexical level for each of the language combinations and decoding direction. The most remarkable lexical asymmetries were observed for CS-RU 20% (CS reader of RU stimulus) vs. RU-CS 26% (RU reader of CS stimulus), as well as for BG-PL 27% (BG reader of PL stimulus) vs. PL-BG 33% (PL reader of BG stimulus). These scores suggest that as far as vocabulary is concerned, CS readers should face less difficulties when reading RU, while RU readers should find it harder to read and understand CS. Accordingly, BG readers are expected to have a slight lexical advantage when reading PL than vice versa. Although both language pairs have similar lexical distances, CS and PL proved to be orthographically more distant from each other than BG and RU (Jágrová, Stenger, Marti & Avgustinova, 2017). Figure 3 shows a matrix of the lexical distances in the twelve language-reader combinations examined (corrected version of Jágrová, Stenger, Marti & Avgustinova, 2017, p. 411).

3 Cyrillic script is transliterated according to ISO 9:1986 (Jágrová, Stenger, Marti & Avgustinova, 2017, p. 407).

		Reader			
		BG	RU	CS	PL
Stimulus	BG		10	27	33
	RU	11		20	23
	CS	29	26		14
	PL	27	20	9	

Figure 3: Lexical distance among the 100 most frequent nouns.

The results display a lexical asymmetry between CS and PL that is larger than in the BG-RU pair, which suggests that PL readers might find it harder to read and understand CS texts because of the higher share of non-cognates. The combination that is least intelligible on the lexical level according to Figure 3 must be BG for a PL reader (33%). BG turns out to have higher distance scores for any reader when compared to the other languages read by other readers, meaning that BG is expected to cause the greatest lexical problems for other Slavic readers. The opposite holds for RU – the scores suggest a maximum distance of only 23% for PL readers, meaning that RU is expected to cause less lexical problems than any of the other languages viewed here. The scores surprisingly also imply that with regard to lexis, it must be slightly more difficult for a PL reader (23% distance) than for a Czech reader (20% distance) to understand RU, even though the fact that Poland is geographically closer to Russia than the Czech Republic might lead to different expectations.

Previous studies on lexical distance (e.g. Heeringa et al., 2013) treated false friends as other non-cognates. This would mean to also assign a distance value of 1 to them. In a regression analysis of experimental results, Jágrová (2018, pp. 127-128) found that predictors calculated with a lexical distance score of 2 for false friends correlate better with processing times of NPs (see also section 14) than if calculated with a distance score of 1. The same observation was made in a study on PL sentences in Jágrová, Avgustinova et al. (2019).

However, the policy with assigning a score of 2 to false friends was changed during the analysis of later experiments: It turned out that some target words can be false friends, i.e. that they are strongly misleading when presented without context, but they still can be cognates in a particular context. This means that words can be both false friends and cognates, which can actually improve their intelligibility in a given context. Therefore, in the analyses in section 13 and in CHAPTER VI, two separate lexical variables are applied to target words – the binary categories cognate/non-cognate (C/NC) and false friend/no false friend (FF/no FF).

1.3.2. Orthographic distance

Even if words in a related Lx are cognates, they can be difficult to identify for readers, for instance if they have a relatively high orthographic distance. Accordingly, readers will be more successful in identifying and understanding cognates when they are spelled more similarly to their L1. The assumption is that the higher the orthographic distance, the more difficult it is to comprehend written cognates of the related Lx (cf. Gooskens, 2007; Vanhove, 2015). In the literature (e.g., Heeringa et al., 2013; Golubović, 2016), orthographic and morphological distances are usually measured as string similarity by means of the Levenshtein algorithm (Levenshtein, 1966) – hereafter referred to as *trad LD* (traditionally calculated Levenshtein distance) – which aligns consonant and vowel letters of cognates separately in slots. Table 5 provides a comparative overview of the CS and the PL alphabet. The characters that both alphabets share are in merged cells, while unique characters are displayed in the respective row. The characters that are displayed in the same column are not supposed to reflect sound correspondences in Table 5, although some of them do. They differ in diacritics and/or pronunciation. The character *ó* is an exception: Even though it carries the same diacritics in the two languages, it still differs in pronunciation.

CS		á	b	č	d	d'	e	é	ě	f	g	h	ch	i	í	j	k	l		m	n
PL	a	ą	b	ć	d	e	e	f	g	h	i	j	k	l	ł	m	n				
CS	ň	ó	q	r	ř	s	š	t	ť	u	ú	ů	v	w	x	y	ý	z	ž		
PL	ń	ó	p	r	s	ś	t	u	u	u	v	w	x	y	z	z	z	z	z	z	z

Table 5: Comparison: the CS and the PL alphabet.

In their coursebook on Slavic comparative linguistics, *Slavischer Sprachvergleich für die Praxis* [Comparison of the Slavic languages in practice], Heinz & Kuße (2015) give an overview of the sound correspondences as they are orthographically represented in six Slavic alphabets. In addition to the comparison of the alphabets in Table 5, Table 6 displays some of the CS-PL sound correspondences as listed in Heinz & Kuße (2015):

CS	Đ đ Dě dě Di di	Dž dž	ě/je	Ň Ň Ně ně Ni ni	Kv kv	š š	šť šť	Ž ž	Ť ť Tě tě Ti ti	Č č	X x
PL	Dź dź Dzi dzi	Dż dż	Je je	Ń ń Ni ni	Kw kw	Sz sz	Szcz szcz	Ż ż Rz rz	Ć ć Ci ci	Cz cz	Ks ks

Table 6: PL-CS sound correspondences (Heinz & Kuße 2015, pp. 70-72).

In order to calculate orthographic distance of cognates, the letters of word pairs are automatically aligned in slots first. The calculations in the INCOMSLAV project were implemented with the help of an algorithm that is fed with letter weight matrices. Such a matrix consists of two alphabets representing the two languages for which distance is measured and numerical values (costs) are assigned for every possible letter alignment of these two alphabets. The PL-CS matrix can be found in Table A 1 and Table A 2 in the appendix. In order to avoid an alignment of vowel to consonant letters, all combinations of vowels and consonants are assigned a cost of 4.5 (most expensive). Combinations of two vowels or two consonants are assigned a cost of 1 and combinations of identical letters in the two alphabets cost 0 (cheapest). If letters differ only in their diacritical signs, they are given a weight of 0.5. The algorithm iterates along a list of word pairs, preferring the cheapest alignment.

Once the algorithm aligned the word pair and determined the length of the alignment, the second step can follow: The actual LD is calculated as demonstrated in Table 7.

# Slots	1	2	3	4	5	6	LD
PL stimulus	s	z	k	o	ł	a	
Closest CS cognate	š		k	o	l	a	
Costs	0.5	1	0	0	0.5	0	2/6 = 33.33%

Table 7: Example for the calculation of *trad* LD.

The cognate pair *szkola* – *škola* ‘school’ in Table 7 requires a deletion of the letter *z* (costs 1), a substitution *s* for *š* (costs 1), and *ł* for *l* (each costs 0.5) from the perspective of a Czech reader. Nevertheless, the perspective of reading is irrelevant in *trad* LD, since the costs are the same for both directions. The different diacritical signs existing in the two alphabets are not distinguished: A difference in diacritics always costs 0.5. The total cost for the transformation of the word pair in Table 7 is 2. This is divided by the number of alignment slots – in this case 6. This results in a normalised orthographic distance of the word pair *skoła* – *škola* ‘school’ of 33.33%. If two words are identical in the way they are spelled, they have an orthographic distance of 0, regardless of their possible semantic differences (cf. Jágrová, Stenger, Marti & Avgustinova, 2017, p. 409).

Figure 4 shows the orthographic distances among BG, CS, PL, and RU, calculated on the lists of the most frequent nouns (Jágrová, Stenger, Marti & Avgustinova, 2017, p. 413). Orthographic distance was calculated both with and without transliteration (upper vs. lower part of the matrix) of the languages using Cyrillic script accordingly. Even though trad LD is a symmetric distance measure, we observe asymmetry in most of the orthographic distances in Figure 4, since they were calculated on different lists – the initial lists are the most frequent nouns of a language and distance is calculated towards the closest cognate translations in the other languages.

		Reader				
		BG	RU	CS	PL	
Stimulus	Untransliterated	BG		13	68	70
		RU	14		70	69
		CS	78	77		35
		PL	77	78	34	
	Transliterated	BG		13	24	31
		RU	14		26	34
		CS	24	24		35
PL		33	34	34		

Figure 4: Orthographic distance of cognate pairs without and with transliterations.

In general, the results reveal that CS and PL display a large discrepancy between lexical closeness (only 9% distance in PL for Czech readers, resp. 14% distance of CS for Polish readers) on the one hand and high orthographic distance (34% PL for Czech readers, resp. 35% CS for Polish readers) on the other hand. The orthographic distance of PL-CS is the greatest among all combinations viewed here. When comparing the languages sharing the same script, there is a remarkably lower orthographic distance in the pair with Cyrillic script (RU reader of BG stimulus 13% vs. BG reader of RU stimulus 14%) than in the pair with Latin script. This suggests that CS and PL are less orthographically intelligible to each other than BG and RU are. For the transliterated distances, the highest orthographic distances can be observed in all combinations with PL, both in the en- and decoding direction: The distances predict not only the greatest difficulties when the other readers try to understand PL, but also when Polish readers would try to understand CS, BG, and RU (the latter two in transliteration). This confirms the findings of Heeringa et al. (2013) that PL is an outlier among the Slavic languages in terms of orthography. It also confirms previous orthographic

distance calculations on the lists of Pan-Slavic vocabulary, internationalisms and cognates from the Swadesh list discussed in 1.2 – these are indicated in Table 8 (Stenger, Jágrová et al., 2020, p. 487).

Word list	CS-PL orthographic distance by means of trad LD
Internationalisms	17%
Pan-Slavic	39%
Swadesh	42%

Table 8: Trad LD for PL-CS: internationalisms, Pan-Slavic vocabulary, and Swadesh list.

The reason for these high orthographic distance values for PL in all combinations is probably that other languages use single letters where PL uses digraphs. The digraphs *cz*, *rz*, and *sz* require insertion of additional letters in cognate pairs that contain the letters *č*, *ř*, or *š* in CS, which leads to greater costs in the Levenshtein alignment and a higher orthographic distance.

The translated lists and the word alignment matrices for the Levenshtein distance (LD) calculations in Jágrová, Stenger, Marti & Avgustinova (2017) were made available online under <http://www.coli.uni-saarland.de/~tania/incomslav.html/> (CC-NC-SA). An access code can be requested from the authors.

1.3.3. Other distance measures in the literature

The role of morphology and syntax has not been investigated as thoroughly as the other linguistic distance measures (Hilton et al., 2013). While syntactic and morphological distances have been included into statistical models of mutual intelligibility before and weighed against the influence of other predictors (Gooskens & Swarte, 2017), the topic of divergent morphology is approached with a systematic morphological modification of stimuli in section 4 in this thesis.

Gooskens & Swarte (2017) used a broad phonetic transcription of stimuli in a study of mutual intelligibility between the Germanic languages, where spoken audio recordings were played to respondents in translation experiments. They found that besides lexical and orthographic distance, also phonetic/phonological distance was one of the most important linguistic predictors of intelligibility between the five Germanic languages Danish, Dutch, English, German, and Swedish. Phonetic or phonological distance is not considered here, because no audio recordings but only written stimuli were presented to the respondents. Instead, the aspect of phonetic representations of the written

stimuli will be discussed as a matter of perceived distance reflected in respondents' utterances in the cooperative translation experiments and in the calculation of a pronunciation-based Levenshtein distance (section 6.1).

This thesis does not account for the role of syntactic distance as examined, for instance, by Golubović (2016) who measured syntactic distance of parallel texts as the correlation between the POS trigram frequencies of related languages. Obolonchikova (2017) compared the number of crossings and clusters between words of aligned sentences from parallel corpora between BG, CS, PL, RU, and UK as a measure for cross-lingual similarities in word order. The difficulty caused by divergent word order is expected to be better reflected by statistical language models – hereafter referred to as LMs – which inform about the (un)predictability of particular words (and not only their POS) in context. Details about this method are elucidated in section 1.5.

1.4. Asymmetry in Cross-Lingual Intelligibility

This section picks up on the methods and results published in

Stenger, I., Jágrová, K., Fischer, A., Avgustinova, T., Klakow, D., & Marti, R. (2017). Modelling the impact of orthographic coding on Czech-Polish and Bulgarian-Russian reading intercomprehension. *Nordic Journal of Linguistics*, 40(2), 175-199. doi:10.1017/S0332586517000130

1.4.1. Conditional entropy

In addition to linguistic distance, not only the similarity of two languages, but also their cross-lingual regularity was estimated with two other information-theoretic measures – conditional entropy and surprisal (Shannon 1948). Conditional entropy assigns lower values to cross-lingual correspondences with greater regularity and turned out to be a good predictor when it comes to the comparison of cognate sets of identical size between language pairs (Stenger, Avgustinova & Marti, 2017). As this thesis does not deal with the comparison of distances between several language pairs, but focusses only on the PL-CS pair, conditional entropy and the related measure – word adaptation surprisal – are touched upon only briefly here.

Stenger, Avgustinova & Marti (2017) applied the measures conditional character adaptation entropy and word adaptation surprisal in order to account for the asymmetries in the mapping of one orthographic system on another in language pairs. They found that word-length normalised adaptation surprisal was a better predictor for mutual intelligibility than aggregate Levenshtein distance when the same stimuli sets in different Slavic language pairs with Cyrillic script were compared.

Despite the higher correlations of these measures as predictors of mutual intelligibility between language pairs and the consideration of asymmetry which LD cannot account for⁴, these measures have a disadvantage when it comes to analysing the processes involved in intercomprehension within one language pair: The entropy values always depend on the word list that they have been calculated on. The longer the cognate list, the more reliable the values should be. Also, they cannot (or at least should not) be applied to material containing non-cognates, which confines its applicability to cognate lists. If one is interested to estimate how difficult only one sentence, phrase or one word in a related Lx would be for a reader to whom only this one stimulus is presented, one could only calculate the regular distribution of cross-lingual correspondences as indicated in the example with a cognate pair from Stenger, Jągrová et al. (2017) in Table 9:

	1	2	3	4	5	6	7
CS	<i>m</i>	<i>l</i>	<i>a</i>	<i>d</i>	<i>o</i>	<i>s</i>	<i>t</i>
PL	<i>m</i>	<i>ł</i>	<i>o</i>	<i>d</i>	<i>o</i>	<i>ś</i>	<i>ć</i>
CS reader	1:1	1:1	1:2	1:1	1:2	1:1	1:1
PL reader	1:1	1:1	1:1	1:1	1:1	1:1	1:1

Table 9: Calculation of conditional entropy of a cognate pair.

Table 9 demonstrates the calculation of conditional entropy on the PL-CS cognate pair *młodość* – *mładost* ‘youth’ (Stenger, Jągrová et al., 2017, p. 183), character by character. The alignment rules are the same as those that apply for the alignment in the Levenshtein algorithm (cf. section 1.3.2) – the vowel and consonant characters are aligned separately. The basic idea in this example is that Polish readers should have an advantage in understanding CS *mładost* ‘youth’ over Czech readers attempting to understand PL *młodość* ‘youth’. There is 0 entropy for the correspondences from a Polish reader’s perspective, whereas from a Czech reader’s perspective, the two characters *o* in *młodość* can either transform into a CS *o* or *a* with equal probabilities (50% each). “ $p(o|o)$ and $p(a|o)$ is 0.5, the entropy of *o* is $(1/2(-\log_2(0.5)) + 1/2(-\log_2(0.5)))/2 = 1$, and the overall entropy for this direction is $(2 * 1 + 5 * 0)/7 \approx 0.29$ ” (Stenger, Jągrová et al., 2017, p. 187) which is higher than 0 for the Polish reader. If the CS reader is aware of these probabilities and the correct solutions, then the model can be a suitable predictor even for this individual cognate pair.

4 Or only when calculated on different word sets.

However, how can a reader know these correspondences without knowing the correct translation of a cognate? Moberg et al. (2006) state that this measure of complexity reflect[s] the difficulties with which a reader is confronted in guessing the correct correspondence.

CS entropy for PL readers	Characters	PL entropy for CS readers
0.08	<i>a</i>	1.16
	<i>q</i>	1.14
0.20	<i>á</i>	
0.91	<i>e</i>	0.87
	<i>ę</i>	0.72
0.07	<i>é</i>	
0.74	<i>ě</i>	
0.09	<i>i</i>	0.74
1.69	<i>í</i>	
0.14	<i>o</i>	0.21
0	<i>ó</i>	1.69
0.64	<i>u</i>	0.02
0.92	<i>ú</i>	
0	<i>ů</i>	
0.09	<i>y</i>	1.63
0	<i>ý</i>	

Table 10: Vowel character entropies for the PL-CS language pair.

Table 10 summarises the conditional entropy values of CS and PL vowel characters calculated on 1,182 word pairs (rounded values as of Stenger, Jągrová et al., 2017, p. 188). The entropy, for instance, of the CS vowel character *o* for Polish readers is lower (0.14) than that of PL *o* for Czech readers (0.21): “More precisely, the PL *o* can map into 6 CS characters (*o*, *e*, *a*, *á*, *ů*, or *i*) and the CS *o* can map only into 2 PL characters (*o* and *ó*) or to nothing” (Stenger, Jągrová et al., 2017, p. 188). Of course, again, in an intercomprehension scenario, neither a Czech nor a Polish reader can be expected to know these mappings or their probability distributions. The results from the cooperative translation experiments (section 5) reveal that entropy-based predictions do not always agree with human performance.

Let us consider the PL word *ręka* ‘hand [instr]’ as it was presented to Czech respondents in the sentence

Nie widziałam, że jego żona pokazuje ręką, żebyśmy poszli do rektora.
 ‘I have not seen that his wife is showing with her hand that we should go to the rector.’

In the cooperative translation experiments, respondents were asked to first read out the stimuli aloud and then try to translate them. It turned out that already when reading the PL stimuli aloud, readers ignore, replace, or shift diacritics (section 5.7.1). Figure 5 visualises the regular PL-CS correspondences extracted from large cognate sets (Fischer et al., 2015, section 1.2.1) as they theoretically should apply for the vowels in the cognate forms *ręka*:*rukou* according to traditional linguistic assumptions.

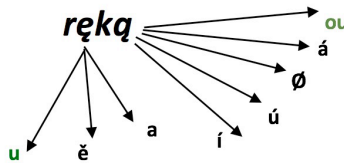


Figure 5: Expected transformations of unknown characters in a PL stimulus by a Czech reader.

Instead, respondents turned *ręka* into *řeka* ‘river’ (shift of diacritics from *ę* to *ř*), *reka* (gen of *rek* ‘hero’, ignoring diacritics) or *řiká* ‘she says’ (replacing diacritics in *ę* for *á*), but not into the correct translation *rukou* (instr of *ruka* ‘hand’). For a comparison, Figure 6 visualises the actual processes observed when Czech respondents read and translated this stimulus word.

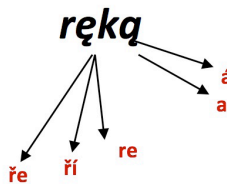


Figure 6: Transformation of unknown characters observed in cooperative translation experiments

In other cognate pairs, Czech readers were able to apply regular PL-CS correspondences, but the application is not always consistent. For instance, respondents successfully applied the correspondence *rz*:*ř* in *porządkowe*:*pořádkové* (100% of all read-out instances) in sentence 11 of the cooperative translation

experiments, but in 75% of all read-out instances of the stimulus word *brzozy* ‘birch [gen]’, respondents made a syllabic division between *r* and *z* pronouncing it /br-zo-za/ which led to wrong translations in some cases.

1.4.2. Word adaptation surprisal

The calculation of word adaptation surprisal follows the same principles as that of conditional entropy, with the difference that the surprisal for the individual character correspondences is counted on a large list of cognate pairs and not on only one word pair. The underlying calculation for the same cognate pair as in Table 11 would be: “*m:m* (surprisal: 0.001), *l:l* (surprisal: 0.0), *o:a* (surprisal: 6.724), *d:d* (surprisal: 0.0), *o:o* (surprisal: 0.036), *s:s* (surprisal: 0.0), *ć:t* (surprisal: 0.002)” (Stenger, Jágrová et al., 2017, p. 190). These values are summed up per word and then divided by the number of alignment slots for the normalised measure (norm WAS) as demonstrated in Table 11: $6.78/7 = 0.97$ for a Czech reader vs. $13.9/7 = 1.99$ for a Polish reader (Stenger, Jágrová et al., 2017, p. 190).

	1	2	3	4	5	6	7	normalised WAS
CS	<i>m</i>	<i>l</i>	<i>a</i>	<i>d</i>	<i>o</i>	<i>s</i>	<i>t</i>	
PL	<i>m</i>	<i>ł</i>	<i>o</i>	<i>d</i>	<i>o</i>	<i>ś</i>	<i>ć</i>	
CS reader	0.001	0	6.724	0	0.036	0	0.002	6.78/7 → 0.97
PL reader	0	2.229	6.984	0	0.026	2.968	1.700	13.9/7 → 1.99

Table 11: Calculation of word adaptation surprisal of a cognate pair.

Again, the smaller the WAS value, the easier it should be to guess the correct cognate with the applicable orthographic correspondence. For instance, the *l:l* correspondence has an adaptation surprisal value of 0 for Czech readers, because PL *ł* always corresponds to CS *l* in the cognate list that the adaptation surprisal was calculated on.

1.5. Surprisal and Context

Parts of this section pick up on the project publication

Jágrová, K., Avgustinova, T., Stenger, I., & Fischer, A. (2019). Language models, surprisal and fantasy in Slavic intercomprehension. *Computer Speech and Language* 53. 242-275. doi:10.1016/j.csl.2018.04.005

In psycholinguistic research, processing effort in monolingual reading situations has been measured in terms of event-related potentials (ERPs, e.g. Block

& Baldwin, 2010), fixation duration and eye movements in eye tracking experiments (e.g., Demberg & Keller, 2008; Rayner & Well, 1996), by self-paced reading time of stimuli (Smith & Levy, 2013) or cloze probabilities (Bloom & Fischler, 1980; Block & Baldwin, 2010). These measures correlate strongly with predictability scores from statistical LMs. Levy (2008) showed that trigram LMs performed well at predicting the processing difficulty measured by the reading times of texts of various difficulties. The measure employed is called surprisal.

Surprisal is widely used in information-theoretic modelling of human language and captures frequency and predictability effects. It reflects the information content conveyed by a linguistic unit, the unpredictability of units in context and the cognitive effort that is required to process this information (Crocker et al., 2015). In contrast to the mere frequency data of independent words that can also be obtained from large corpora, surprisal measures probability of a word w_1 depending on its preceding words w_2 , w_3 etc. For a unit, surprisal is defined as the negative log-likelihood of encountering this word in its preceding context. It is defined as:

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

The lower the surprisal, the more predictable a word is in a sentence, given its preceding words. Whenever there is a drop in surprisal after a word, this indicates that the word with the lower surprisal is highly predictable after the preceding word.

Surprisal theory in principle includes three areas that are fundamental in a communicative situation as described in Shannon's noisy channel model (Shannon, 1948): (a) Coding of a message (related to language production), (b) channel constraints, and (c) noise. According to the UID (uniform information density) hypothesis, speakers tend to distribute information as close to constant as possible over the duration of an utterance, avoiding peaks and troughs in surprisal (Jaeger, 2010). The theory concerns the production-related features of a message and is not of primary relevance in the present intercomprehension setting, since there is no communicative partner in the translation task, and therefore the UID hypothesis is not a topic here.

The aspects relevant for the present intercomprehension setting are the channel constraints and the noise. The channel constraint part concerns language perception and comprehension and aims to explain why, for instance, some sentences are more difficult to read than others. The noise aspect within the theory concerns the uncertainty that noise injects into the raw input. Noise can be plain acoustic noise from the environment, a coffee stain on a letter

or, in the present setting, an imperfect linguistic signal – that of a related, but unknown foreign language. Specifically, this might be an unexpected orthographic unit in a still understandable cognate word, such as the character *w* in the PL word *woda* ‘water’ when a Czech reader would expect a *v* as in the CS translation equivalent *voda* ‘water’. In the PL-CS setting, there can be characters with a known base, but noise in the shape of unknown diacritics, e.g. in the characters *ą*, *ł* or *ź* which do not exist in CS. On the sentence level, Czech readers can encounter noise in the form of non-cognates within an otherwise understandable sentence, e.g. the word *rowerze* ‘bike [loc]’ in the sentence

PL: *Dobrym sposobem zachowania dobrej kondycji jest jazda na rowerze.*

CS: *Dobrym způsobem zachování dobré kondice je jízda na kole.*

EN: ‘A good way to maintain a good condition is to ride a **bike**.’⁵

which, except for the word *rowerze*, consists only of cognates and should otherwise be understandable for Czech readers. Of course, there can also be combinations of different sources of noise in one message.

In order to use LMs for any kind of linguistic application, the LMs have to be trained on a corpus. The corpus is usually pre-processed according to the needs of the user (the researcher). The corpus language that the LMs are trained on then represents a monolingual reader of this language.

Figure 7: Trigrams as they could occur in a PL corpus during training.

Figure 7 visualises the training of a trigram LM on a PL corpus – the algorithm counts all combinations of words in a window of three words, whereby punctuation signs are counted as words. In this example, the trigram *to jest nasz* ‘this is our [masc]’ occurs twice and would hence be the most frequent here.

5 Original version of the sentence as of Block & Baldwin (2010): “A good way to exercise is to ride a **bike**.”

In comparison, *to jest nasze* ‘this is our [fem]’, as all other trigrams, occurs only once. A bigram LM would be with a window of two words, a four-gram would be with four words accordingly etc.

After an LM has been trained, it can be used to score language material that should be in the same language as the training corpus and to which the same pre-processing steps were applied. Such an n-gram LM can predict words only in a limited context, i.e. in the context of the window of n words that it was trained on. For the sample corpus in Figure 7, this would mean that after *to jest* ‘this is’, *nasze* ‘our [fem]’ would be assigned a higher surprisal than *nasz* ‘our [masc]’.

In this thesis, only n-gram LMs with Kneser-Ney smoothing (Kneser & Ney, 1995) are used for modelling the difficulty or unpredictability of words in context. “The Kneser-Ney smoothing technique leverages available information from overlapping, smaller n-grams to ensure that surprisal scores computed for unseen word combinations do not turn out extremely high” (Jágrová, Avgustinova et al., 2019, p. 251). The PL stimuli (NPs and sentences) in this thesis were scored by a trigram LM trained on the PL part of SCD InterCorp (size: 118,651,918 words, Čermák & Rosen, 2012) and the CS literal translations (as close as possible translations) of these were scored by a trigram LM trained on the Czech National Corpus (CNC – SYN version 5, released in 2015, size: 4,599,643,984 words, Křen et al., 2015). The LMs provide surprisal values in the unit hartley (symbol Hart). Hartley measures information or entropy and is the pendant of the bit. While hartley uses the common logarithmic base 10, the unit bit uses the binary logarithm to the base 2.

Surprisal scores can also be interpreted as a measure for the typicality of certain constructions. As for phrases and sentences, surprisal can not only estimate which word order is more typical. In NPs, for instance, it can estimate how likely particular nouns are after particular adjectives, respectively how likely particular adjectives are to appear after particular nouns. NPs are subject of the analysis in section 14. The lower the surprisal score of an NP, the more expectable it should be for a reader. 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 the PL LM which assigns a high probability – and hence low surprisal score (1.12 Hart) – to *dom* after *biały*, while assigning a low probability – and hence a high surprisal score (7.02 Hart) – to the word *sześciokąt* after *biały*. If both words in the NPs are scored accordingly, one can obtain a total surprisal score for both words of the NP: 3.05 Hart for *biały dom* ‘white house’ (1.93 Hart + 1.12 Hart) and 11.20 Hart for *biały sześciokąt* ‘white hexagon’ (4.18 Hart + 7.02 Hart).

Thus, if a noun is highly unexpected after a certain adjective, it will lead to a high total surprisal score for the NP. The same should apply for the typicality of word order in sentences.

Among the work correlating measures obtained from LMs with context, Bernardy, Lappin & Lau (2018) have investigated the influence of document context on human acceptability judgements for machine-translated EN sentences. They presented stimuli sentences once in an experimental setting without any other text and in another experimental setting with their original document contexts. They assessed the accuracy of two different types of LMs (those that incorporate context during training and those that do not) as predictors of human judgements. They found that human acceptability increased for ill-formed sentences when presented with context, but also decreased for well-formed sentences from a certain threshold level of the ratings (Bernardy, Lappin & Lau, 2018, p. 460). A possible explanation for that could be that humans, when presented with context, focus more on semantic and pragmatic coherence than on grammaticality, which could also be of relevance in an intercomprehension setting. They also found that agreement between human ratings increases when context is introduced and that the LM incorporating context performed better at modelling this human performance.

1.6. Context in Intercomprehension

The role of context for the understanding of a particular Lx has been subject to relatively few studies on intercomprehension, although it is crucial for the cognitive processes involved in the human comprehension system. Jágrová (2018) examined surprisal as a predictor for the added difficulty in NA (noun + adjective) word order in NPs for Czech readers. Jágrová, Avgustinova et al. (2019) qualitatively examined surprisal and intelligibility on three PL sentences translated by Czech respondents. They found that “linguistic distance (encoding similarity) and in-context surprisal (predictability in context) appear to be complementary, with neither factor outweighing the other, and that our distinguishing of these two measurable dimensions is helpful in understanding certain unexpected effects in human behaviour.” (Jágrová, Avgustinova et al., 2019, p. 242). With regard to intercomprehension, it is still not entirely clear to what extent predictability in context interplays with other linguistic factors in understanding a related but unknown language. A systematic examination of surprisal on larger data sets in order to capture the role of sentential context as a measurable variable is still missing.

In a study on the disambiguation of cross-Slavic false friends in divergent sentential contexts, Heinz (2009) 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 (Heinz, 2009), meaning that the target word might be misinterpreted due to misleading or misinterpreted context.

Muikku-Werner (2014) 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. 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” (Muikku-Werner, 2014, p. 105). She defines intercomprehension as a holistic process in which “perceived similarity leads to different comprehension results in single items and in texts.” (Muikku-Werner, 2014, p. 102). Muikku-Werner refers to Sinclair’s definition of collocations: “The occurrence of two or more words within a short space of each other in a text” (Sinclair, 1991, p. 170). She distinguishes six different semantic links between words in collocations (Muikku-Werner, 2014, p. 108):

- a) same semantic field
- b) hyponymy
- c) schematic implication
- d) two or more co-ordinated co-hyponyms of some semantic category
- e) antonymy
- f) cause-consequence

Another concept that is therefore likely to play a role in the intercomprehension of sentences is that of semantic priming (cf. Harley, 2007). Gulan & Valerjev (2010) 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 (Gulan & Valerjev, 2010, p. 54). During associative priming, a word causes associations of 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” (Gulan & Valerjev, 2010, p. 54), for instance to play – to have fun. Semantic priming in intercomprehension, of course, can only work if the prime in the Lx is correctly recognised as such.

2. Thesis Focus: Modelling Linguistic Phenomena of PL for Czech Readers

PL and CS both belong to the West Slavic language group, together with the official languages Slovak (SK), and Sorbian (Lower and Upper Sorbian). As mentioned in section 1.3.2., Heeringa et al. (2013) and found that PL is an outlier in terms of orthography among the other Slavic languages spoken in the EU (Heeringa et al., 2013, p. 119). Golubović measured the linguistic distances between the Slavic languages spoken in the European Union and confirmed that PL is an outlier in terms of orthography, having the greatest orthographic distance to the other five Slavic EU languages (Golubović, 2016, p. 49). Jágrová, Stenger, Marti & Avgustinova (2017) found that in relation to the small lexical distance (9%)⁶ between PL and CS, their orthographic distance (34%) is extraordinarily high when compared to BG and RU that have similar levels of both orthographic (13.5%) and lexical distance (10.5%). As for the linguistic distance and intelligibility of PL sentence material for Czech readers, findings from the literature are summarised in Table 12 (cf. Jágrová & Avgustinova, 2017).

Distance	Heeringa et al., 2013	Golubović ⁷	Jágrová, Stenger, Marti & Avgustinova, 2017	Jágrová, Avgustinova et al., 2019	Stenger, Jágrová et al., 2017
Lexical	23%	17%	10%	12%	-
Orthographic	31%	31%	34%	38%	32%
Morphological	-	31%	-	-	-
Intelligibility	64% ⁸	41%	71% ⁹	-	67% ⁹

Table 12: PL for Czech readers: comparison of distance and intelligibility in the literature.

6 Corrected value, differs from the value in Jágrová, Stenger, Marti & Avgustinova (2017)

7 Data for the written cloze test published in Golubović (2016)

8 Data for the written translation task of the most frequent nouns from the British National Corpus as published in Golubović (2016, p. 77) on the material of Heeringa et al. (2013)

9 Published in this thesis, section 12.2 and 13.1.

According to Golubović (2016, pp. 47-49), PL has an orthographic distance of about 32% and a lexical distance of nearly 18% if read by Czech readers. This again suggests that divergent orthography alone might crucially impair the intelligibility of PL for Czech readers, since the two languages are lexically relatively close. Also, these two languages can be expected to be phonetically closer than orthographically in many cognate pairs – some cognates are pronounced almost identically, but written differently, e.g. PL *woda* and CS *voda* (both ‘water’).

Taken together, all these results suggest that although Czech readers can profit from the large percentage of common cognate vocabulary, the intelligibility of PL might be unsuccessful because of the different orthography. The same might apply for Polish native speakers trying to read CS. For all these reasons that make the PL-CS pair so interesting to compare and for reasons of personal interest, I chose the scenario *written PL presented to Czech native speakers* to be the subject of an in-depth research effort with different methods and materials that constitutes the core of this thesis.

CHAPTER II: COOPERATIVE TRANSLATION EXPERIMENT

3. Experimental Setup

The sentence translation experiments were designed as a cooperative task to be solved by a pair of informants while they were audio-recorded. Section 4 discusses only the quantitative part of the evaluation (responses per stimulus word), while the qualitative analysis of the actual audio recordings is in section 5. The stimuli sentences were also tested in web-based cloze translation experiments in order to obtain a more representative sample – their results are discussed in section 16.

The experiments were conducted at Charles University, Prague in late 2016. The objectives of the experiments were a) to compare the performance of Czech native speakers reading PL sentences in original vs. how they read and understand sentences with systematic modifications and b) to learn about the processes that take place during this reading intercomprehension scenario. The idea of the experiment design is to gain a meaningful and measurable insight into the respondents' minds while they are solving the translation task. The experiment was designed for pairs of informants, because if an individual would have been only prompted to say what s/he is thinking, the informant might actually not pronounce all her or his ideas. Whereas in a cooperative task, informants have to communicate with each other.

16 pairs of informants (32 persons) of which 14 were females and 18 males, aged 16 to 30 (mean age: 22.2), Czech native speakers and students or graduates of the Charles University who did not study linguistics as a subject and who never learned PL participated in the experiment. They were equipped with headsets and communicated with each other over Skype throughout the experiment.

Before the actual experiment, the informants filled out an informed consent form and a questionnaire with the standard empirical data, their L1(s), Ln, exposure to languages (see Figure A 1 in the appendix), followed by a self-assessment of skills for all languages they had indicated. The self-assessment scale was designed as a drag-and-drop bar with a continuous 7-point scale, ranging from 0 to C2, oriented on the Common European Framework of Reference for Languages (CEFR). For each indicated language, the skills for speaking, hearing, reading, and writing were enquired separately.

After the informants confirmed that they had read the instructions (see Figure A 2 in the appendix) and were ready to start the experiment, individual sentences, each in one of the conditions (explained under 4.3) were presented

to the persons on two separate screens simultaneously. By presenting every condition only once to the informants, learning effects (e.g. about regularly occurring PL-CS correspondences) should have been avoided.

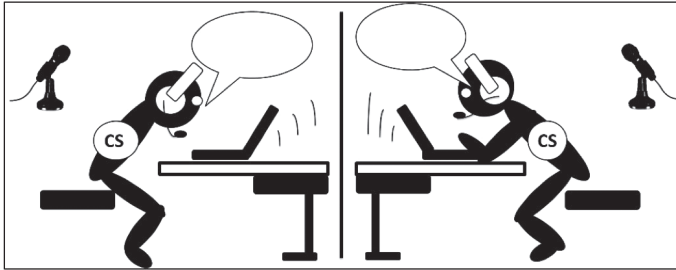


Figure 8: Setup of the cooperative translation experiment in pairs.

The respondents were audio-recorded while trying to cooperatively translate the PL stimuli, each of them working on a separate screen, as visualised in Figure 8. They were placed in separate rooms in order to avoid cross-talk on the audio recordings. Only one person was able to enter the written response at a time (with changing turns after each stimulus). The informants' task was to read the whole sentence aloud first (one of them, again with changing turns) and then to try and translate it into CS cooperatively. They were explicitly asked to discuss with their partner what they think they do or do not understand. They were also asked to try to translate the entire stimulus and even if they would not know a certain word, they should guess it from the context.

4:59

Ne widziała jsem, že jeho žena ukazuje rękou, aby chom szli k rektorovi.

Neviděla jsem, že ...

souhlasím

Figure 9: Screen during the cooperative translation experiment.

The example in Figure 9 shows the experimental screen with sentence 8 in the all no orth condition. Respondents entered their joint solution in the field underneath the stimulus. The other person was able to see what the partner was writing. The time limitation for each stimulus sentence was set to 5 mins. When both informants clicked on the *souhlasim* 'I agree' button before the 5 minutes expired, their translation was stored and the next stimulus was presented. The informants did not get any feedback on the correctness of their translations during the experiment.

The experiment output consists of two parts: About 10 hours of audio recordings of the 16 participant pairs trying to decode the stimuli and the written translations they have entered during the experiment. The complete transcripts and can be made available upon request.

4. Quantitative Analysis of Written Results and Comparison of Conditions

This section examines the written answers collected in the cooperative translation experiment from a quantitative approach. These stimuli were systematically modified on the different linguistic levels in order to control for the role of the cross-lingual phenomena on the individual levels (orthography, morphology, closed class words, lexis, and word order) and in order to assess to which degree the differences on these levels might influence reading intercomprehension. The overall aim is to understand the processes that take place when Czech native speakers read PL. As there are certain regularities on the different levels of the linguistic hierarchy that influence mutual intelligibility of related languages, the assumption is that a reader's knowledge about such regularities or about lexis can promote intelligibility. For instance, most Czech native speakers might be aware of the fact that the PL digraph *cz* regularly corresponds to the CS *č*, because they are exposed to it in the way *Czech* is spelled in EN. Basically, a reader's knowledge of regularities can be imitated by a modification of the foreign text towards the reader's L1 and thus the linguistic distance of the text can be minimised and controlled for systematically.

4.1. Hypotheses

The huge discrepancy between the small lexical and the high orthographic distance between CS and PL, as discovered in Jágrová, Stenger, Marti & Avgustinova (2017), suggests that the potential for an improvement of intelligibility by an orthographic modification between CS and PL must be relatively high. A hypothesis resulting from this is that the mutual intelligibility of CS

and PL could be greater if both languages used the same orthographic coding. Accordingly, it is expected that if readers overcome the differences in orthography (e.g. by the knowledge of corresponding units such as *cz:č*), mutual intelligibility between the two languages would be higher. This potential for an increase in intelligibility through modifications might not be that prominent on other linguistic levels or in other language combinations. In the case of BG and RU, for instance, the more serious differences are situated on other linguistic levels, such as in morpho-syntax (e.g. missing grammatical case and infinitive in BG). In particular, the potential to modify for instance BG towards RU only by orthographic means can be expected to be lower than for PL towards CS.

Regarding the impact of word order in intercomprehension, Hilton et al. (2013) found that within the Scandinavian languages, non-native word order has a greater impact on intercomprehension than morphological differences. However, compared to phonological differences between the languages, non-native morpho-syntax was found to have a rather negligible effect (Hilton et al., 2013). The word order in some of the PL stimuli is different from a correct CS word order. Hence, it can be expected that this divergent word order might cause additional difficulty and, consequently, a modification of the word order in a PL sentence towards correct CS might increase its intelligibility for Czech readers.

In this section, these hypotheses are tested for Czech native speakers reading PL with and without different features of CS. To this end, the PL stimuli sentences were modified orthographically. Furthermore, modifications of the same stimuli sentences with CS morphology, lexis, closed class words, and word order were added in order to test how much intelligibility would increase if these were adapted to CS. Of course, closed class words are also part of the lexis. However, they constitute a limited set of words that could be relatively easily learned by hypothetical readers/learners. It also has to be mentioned that there is an interplay of the different levels in a sense that, e.g., morphological units or closed class words contain orthographic features. The method applied here aims to represent a reading situation in which readers have overcome the difficulties on one of the individual levels in order to be able to estimate the relative importance of the difference on the individual levels.

In section 4.3, the modifications applied to the stimuli sentences are explained in detail. Section 4.4 presents the results with regard to the modifications and the written responses from the experiments.

4.2. Stimuli

The stimuli sentences were selected or constructed along different criteria. First of all, each of the stimuli sentences contains at least one PL morpho-syntactic construction or word order feature that is either non-typical or ungrammatical in CS. The cognates within the stimuli were chosen in a way that each of the cross-lingual orthographic correspondences that were gathered in a previous study (Fischer et al., 2015, see section 1.2) were represented at least once throughout the experimental set. Finally, lexically difficult items in the form of non-cognates and known false friends (or such that were expected to be false friends) were added to the existing sentences. The 12 stimuli in their original (unmodified) condition are listed in Table 13: false friends are marked **bold and red**, non-cognates are marked **bold**, differences in word order and morpho-syntax (including the spelling of compound vs. separate words and morphemes) are underlined. The EN translations provided in Table 13 should assist the comprehension of the differences between the PL stimuli and the possible CS translations and are therefore not entirely identical with the EN sentences listed in Table 64 in CHAPTER VI. The modified versions of the stimuli are explained in section 4.3. They can be provided with the written responses to interested parties upon request.

PL stimulus — ORIG condition	Possible good CS translation	EN translation
1 Głoby nie było książek, czytałbym Ci z oczu.	Kdyby nebyło knížek, žetl bych Ti z očí.	If there were no books, I would read from your eyes.
2 W 2000 roku wzrósł do ponad 900 mln. marek obrót towarami, w procesie produkcji których nie używano substancji zagrażających środowisku naturalnemu wilka.	V roce 2000 narostl obrát zboží, u kterého při procesu výroby není užíváno látek ohrožujících životní prostředí vlka, na více než 900 mil. marek.	In the year 2000, the turnover of goods in the production of which no substances that are harmful for the natural habitat of the wolf are used, rose above 900 million German mark.
3 Kolegium dało mi pozwolenie, aby zrealizować ten projekt nad jeziorą.	Kolegium mi dalo povolení, abych zrealizoval ten projekt u jezera.	The council gave me the permission to realize the project at the lake.
4 Praga to ważny węzeł komunikacyjny.	Praha je významný komunikační uzel.	Prague is an important traffic hub.
5 Czy pani będzie głosowała? Czy chciałbyście, aby stały się one gwiazdami?	Paní, budete hlasovat? Chtěli byste, aby omý se staly hvězdami?	Madam, are you going to vote? Do you want [plural] that they [group of females] become stars?
6 Kupiliśmy nie tylko czerstwy chleb, ale jeszcze gorzej — też stary żółty samochód.	Koupili jsme nejen tvrdý chléb, ale ještě hůř — také staré žluté auto.	Not only did we buy stale bread, but even worse — also an old yellow car.
7 Teraz rosna również możliwości odbycia interesujących praktyk w kraju.	Nyní rovněž rostou možnosti absolvování zajímavých praxí v zemi.	Right now, also the possibilities of undergoing interesting internships at home are growing.
8 Nie widziałam, że jego żona pokazuje ręką, zebymy poszli do rektora.	Neviděla jsem, že jeho žena ukazuje rukou, abychom šli za rektorem.	I have not seen that his wife is showing with her hand that we should go to the rector.
9 Skąd jesteś przekonana, że za pięćdziesiąt lat ludzie nie będą już latali samolotem?	Proč jsi přesvědčená, že za padesát let lidé již nebudou létat letadlem.	Why are you convinced that in fifty years people will no longer fly aeroplanes?
10 OCZEKIWANIA: doświadczenia w pracy przy produkcji mięsa; pełna dyspozycyjność od poniedziałku do piątku. OFERTA: realne możliwości awansu w firmie; 12,00 brutto/godzinę + premie miesięczne	POŻADAVKY: pracovní zkušenosti při zpracování masa, ochota pracovat (být plně k dispozici) od pondělka do pátku. NABÍZÍME: realní možnosti postupu ve firmě, 12,00 hrubého/hodinu + měsíční prémie	EXPECTATIONS: work experience in the meat production; full availability from Monday to Friday. OFFER: realistic promotion opportunities within the company, 12.00 gross/hour + monthly bonuses
11 OBSŁUGA SKLEPU — ZAKRES OBOWIĄZKÓW: znajomość języka polskiego; ekspozycja towarów; gotowość do pracy zmianowej; czynności porządkowe	OBSŁUHA OBCHODU — ROZSAH POVINNOSTÍ: znalost polského jazyka, vystavování zboží, ochota práce na směny, udržování pořádku	SALESPERSON IN A RETAIL STORE — SCOPE OF DUTIES: knowledge of Polish; display of goods; readiness for shift work; cleaning activities
12 NAPÓJ Z MIĘTY I MIODU: mięta zielona suszona: 25 g; miód kwiatowy: 50 g; cytryna: 1 szt.; lód konsumpcyjny: 5 kostek; sok z brzozy: 100 ml; jarzębiny: 50g.	NÁPOJ Z MÁTY A MEDU: sušená zelená máta: 25g; květový med: 50g; citron: 1 kus; konzumní led: 5 kostek; šťáva z brzozy: 100ml; jeřabiny: 50g.	MINT AND HONEY DRINK: dried green mint: 25 g; blossom honey: 50 g; lemon: 1 piece; consumable ice: 5 cubes; birch sap: 100 ml; rowan berries: 50 g.

Table 13: Sentences in cooperative translation experiment and possible translations.

4.2.1. Linguistic distance of stimuli

The methods for measuring linguistic distance of the stimuli in the cooperative translation experiments are based on those described in section 1.3 through the principle of the closest possible translation described in section 8. A **total linguistic distance** measure was applied as a predictor variable in this section, unifying both lexical and orthographic distance and relying largely on an overall processing difficulty. Every word that does not have a cognate translation equivalent in the other language is counted as a non-cognate and is assigned a distance score of 1. If this non-cognate is also a false friend, such as *przekonana* in sentence 9 of Table 13 (it could easily be mistaken for *překonaná* which means ‘overwhelmed’ in CS), it is counted with a distance score of 2, assuming that readers are less likely to translate such a word correctly than if it was a random non-cognate. All other words are cognates for which LD is calculated and hence they can have distance scores of ≤ 1 . The distances are calculated for every word within the stimuli in all (modified) conditions.

It was not always trivial to categorise the individual words into the categories cognate, non-cognate or false friend. Here are some examples for unclear cases and how they were treated:

- A translation of *awans* ‘promotion’ occurs only once in InterCorp as the verb *avancirovat* ‘to be promoted’, ‘to advance’. Because of its low frequency and the difference in POS, *awans* was treated as a non-cognate.
- *skąd* ‘where from’ is counted as a cognate of CS *odkud* ‘where from’, because these prepositions have the same stem and only differ in their prefixes. Another argument for considering *skąd* a cognate to *odkud* is that the archaic form *skud* existed in CS (occurrences documented in SyD/CNC until 1875 (Cvrček & Vondříčka, 2011a). The orthographic modification also results in *skud*. Later, from the recordings it became apparent that the informants were not familiar with this archaic interrogative pronoun, but still most of them were able to figure out its syntactic function.
- *teraz* ‘now’ is considered a non-cognate, but should be discussed in a separate analysis on the role of the informants’ multilingual lexicon, because CS native speakers know it by their exposure to SK (Nábělková, 2007), which is also documented in the recordings.
- *interesujący* ‘interesting’ – the closest translation variant *interesující* can be found in the CNC and therefore this stimulus word is counted as a cognate.

- *towar* ‘good’, ‘product’ – although Treq does not offer the CS translation *towar* in the query direction PL to CS, it does so when querying the PL translation of CS *towar*. Although *towar* occurs only 3 times in InterCorp, I consider it a cognate.
- *od poniedziałku* ‘from Monday’ would actually be translated *od pondělka*, but the variant *pondělek* resp. *pondělku* is documented in the CNC (cf. also Šimandl, 2011). Therefore, the LD of it was calculated towards *od* and *pondělku*.
- *jesteś* ‘you are’ would be translated with *jsi* ‘you are’ (LD: 0.67) in standard CS, but the Common CS equivalent is *jseš* (LD: 0.42). Therefore, LD is calculated towards the latter variant.
- Some modal expressions such as *žebyśmy* ‘that we would’ can be translated with two separate words in CS. Although the standard CS translation would be *že bychom* or *abychom* ‘that we would’, the Common CS variant *bysme* ‘we would’ exists in the CNC, its LD is calculated towards *že* ‘that’ and *bysme* ‘we would’.
- Instances in which different prepositions are used, e.g. *nad* ‘on top of’, ‘above’, ‘at’ in *nad jeziorem* ‘at the lake’. For PL *nad*, Treq offers *nad* ‘on top of’, *o* ‘about’, *na* ‘on’, or *u* ‘at’ as CS translation equivalents. Therefore, *nad* is considered having a LD of 0. It is not treated as a false friend, although the correct CS translation of the prepositional phrase would be *u jezera*. Due to the fact that the phrase could also mean ‘above the lake’ in PL, the translation *nad jezerem* ‘above the lake’ is counted as correct.

4.2.2. Surprisal of stimuli

The surprisal of the stimuli sentences is determined in the same way as described later in section 15.2.3. The results concerning the role of linguistic distance and surprisal as predictors are displayed in Figure 13 in subsection 4.4.

4.3. Experimental Conditions: Modification of Stimuli on Different Linguistic Levels

This section explains the modification variants that were applied to the 12 stimuli sentences in the 12 different conditions. It presents (parts of) the original PL stimuli sentences that were tested in the experiment (see Table 13), together with examples for each of the modification variants applied to it. The modifications on the different levels were carried out systematically by substituting certain units from the stimuli with units from the reader’s L1: (i) orthographic

correspondences (*ORTH*), (ii) morphological units (*MORPH*), (iii) closed class words (*CLOSED*), (iv) lexis (*LEX*), and (v) word order (*ORDER*) (section 4.3.1) plus combinations of (i-v) with each excluding one of (i-v) (section 4.3.2).

4.3.1. Conditions with non-combined modifications

Five basic modification variants were applied to the original PL stimuli. The substituted units are green and underlined in the following examples. Please also consider the abbreviations for each modification method given in italics. These are later used for reference in the analysis and in Figure 10-Figure 12 and Table 14.

Substitution of orthographic correspondences – *ORTH*

Orthography can be viewed as a first interface in reading. The regular PL-CS orthographic correspondences gathered in the study by Fischer et al. (2015) and explained in section 1.2. were applied here. The substitution was implemented as visualised in this example:

ORIG: *NAPÓJ Z MIĘTY I MIODU: mięta zielona suszona*

ORTH: **NÁPOJ Z MÁTY I MEDU: máta zelená sušená*

Substitution of morphological correlates – *MORPH*

All inflectional and derivational affixes in the PL sentences are replaced by their CS equivalents.

ORIG: *ekspozycja towarów: gotowość do pracy zmianowej*

MORPH: **ekspozice towarů: gotowost do praci smianové*

Substitution of closed class words – *CLOSED*

In this modification variant, all POS from the PL sentences that belong to closed classes (prepositions, determiners, conjunctions, pronouns, auxiliary verbs, numerals, interjections) are replaced by their CS counterparts.

ORIG: *Gdyby nie było książek, czytałbym Ci z oczu.*

CLOSED: **Kdyby ne bylo książek, czytał bych Ti z oczu.*

Substitution of non-cognates for cognates – LEX

Non-cognates from the original PL stimuli are replaced by pseudo-CS cognates that are spelled according to PL orthographic rules.

ORIG: *Skąd jesteś przekonana, że ludzie nie będą już latali samolotem?*



LEX: **Skąd jesteś przeświadczona, że ludzie nie będą już latali latadłem?*

Optimisation of word order – ORDER

The original PL word order was optimised by re-ordering the original PL words according to an appropriate CS word order. This modification concerns e.g. the positions of clitics or the post-modification vs. pre-modification inside NPs. This means that there is no change in linguistic distance for the CS reader, but only in linearisation.

ORIG: *mięta zielona suszona: 25 g; miód kwiatowy: 50 g*

ORDER: **suszona zielona mięta: 25 g; kwiatowy miód: 50 g*

4.3.2. Conditions with combined modifications

There were 6 conditions of stimuli with combined modifications. In the following examples, those units that are marked **red** are not substituted.

All modifications except orthography – ALL NO ORTH

All modifications were applied to the stimuli with the exception of the orthographic correspondences. This concerns only stems, as morphological units or closed class words are substituted anyway.

ORIG: *Nie widziałam, że jego żona pokazuje ręką, żebyśmy poszli do rektora.*

ALL NO ORTH: **Ne **widziała** jsem, že jeho **žona** ukazuje **rękou**, abychom **szli** k rektorovi.*

All modifications except morphology – ALL NO MORPH

Here, all modifications were applied with the exception that derivational or inflectional affixes were not exchanged. However, if orthographic correspondences (from Fischer et al., 2015) could be applied on the affixes, e.g. in *-ości*, then *ś* becomes *s* and *ci* becomes *ti*. If morphological correlates would be applied to this example, then *w kraju* (PL ‘in the country’) would have been replaced by *v kraji* (CS ‘in the region’).

ORIG: *Teraz rosną również możliwości odbycia interesujących praktyk w kraju.*

ALL NO MORPH: **Nyní rosnou rovněž možnosti zajímavých praxí v kraji.*

All modifications except closed class words – ALL NO CLOSED

All modifications are applied except the substitution of closed class words. For instance, *to* ‘this’ is not replaced by the copula *je* ‘is’ here.

ORIG: *Praga to ważny węzeł komunikacyjny.*

ALL NO CLOSED: **Praha to významný komunikační uzel.*

All modifications except lexis – ALL NO LEX

The CS counterparts replace all units except the NCs. Still, orthographic correspondences were applied to the NCs.

ORIG: *Kupiliśmy nie tylko **czerstwy** chleb, ale jeszcze gorzej – też stary **żółty samochód**.*

ALL NO LEX: **Koupili jsme ne jen **čerstvý** chléb, ale ještě hůř – též starý **žlutý** **samochod**.*

All modifications except word order – ALL NO ORDER

Here, all modification variants are applied, but the original PL word order remains. This means that for the CS reader there is only a change in linguistic distance, but not in linearisation. This modification variant should inform about the difficulty caused solely by the different linearisation.

ORIG: *W 2000 roku wzrósł do ponad 900 mln. marek obrót towarami, procesie produkcji których nie używano substancji zagrażających środowisku naturalnemu wilka.*

ALL NO ORDER: **V 2000 roce narostl na více než 900 mil. marek obrat tovarů, v procesu produkce kterých ne užíváno substancí ohrožujících prostředí přírodní vlka.*

All modifications at once – ALL

A combination of all modification variants leads to an acceptable CS translation of the originally PL sentences. Therefore, the remaining average total distance in this modification variant is zero in most of the stimuli sentences.

ORIG: *Kolegium dało mi pozwolenie, aby zrealizować ten projekt nad jeziorem.*

ALL: **Kolegium mi dalo povolení, abych zrealizoval ten projekt u jezera.*

Sentences:	1	2	3	4	5	6	7	8	9	10	11	12
ORIG	x	x	x	x	x	x	x	x	x	x	x	x
ORTH	x	x	x	x	x	x	x	x	x	x	x	x
MORPH	x	x	x	x	x	x	x	x	x	x	x	x
CLOSED	x	x	x	x	x	x	x	x	x	x	x	x
LEX	x	x	x	x	x	x	x	x	x	x	x	x
ORDER	x	x	x	x	x	x	x	x	x	x	x	x
ALL NO ORDER	x	x	x	x	x	x	x	x	x	x	x	x
ALL NO ORTH	x	x	x	x	x	x	x	x	x	x	x	x
ALL NO MORPH	x	x	x	x	x	x	x	x	x	x	x	x
ALL NO CLOSED	x	x	x	x	x	x	x	x	x	x	x	x
ALL NO LEX	x	x	x	x	x	x	x	x	x	x	x	x
ALL	x	x	x	x	x	x	x	x	x	x	x	x

Figure 10: Visualisation of a stimulus set in the cooperative translation experiments.

Figure 10 visualises a possible set of stimuli (marked yellow) as it was presented to a pair of informants. One stimulus set consisted of 12 PL stimuli with a total of 170 words and 1169 signs (in the original PL condition).¹⁰ The experiment was originally designed for 12 respondent pairs so that each condition of each sentence is tested once. Due to a technical failure during one of the experiments (one stimulus sentence in the ALL NO ORTH condition could not be displayed), one of the stimuli sets was presented again to another respondent pair. For this reason, the stimuli were tested 13 times in each of the respective modified sentences, except ALL NO ORTH that was tested only 12 times. Additionally, 3 informant pairs were presented with the complete stimulus set only in the ORIG condition. This is the reason why the data size for the ORIG condition is the largest.

	ORIG	ORTH	MORPH	CLOSED	LEX	ORDER	ALL NO ORDER	ALL NO ORTH	ALL NO MORPH	ALL NO CLOSED	ALL NO LEX	ALL
Words	633	160	164	169	172	183	185	156	180	185	170	166

Table 14: Data sizes: translated words obtained from informants in each condition.

10 Among the stimuli, there were 9 sentences and 3 fragments: 1 recipe and 2 job advertisements.

4.4. Results

4.4.1. Evaluation of the translations per word

Intelligibility is expressed as the percentage of correctly translated words, whereby every word counts as 1 unit for evaluation. The written translations entered by the informants during the experiment were exported from the software and were then evaluated manually. They were analysed word by word and categorised as *correct*, *paraphrase*, *partly wrong*, *wrong* or *nothing* accordingly. As an objective basis for what can be considered a **correct** translation of a word, the web application Treq, which facilitates querying translation equivalents based on InterCorp (Vavřín & Rosen, 2015), is used as a reference. In unclear cases, the PL stimulus was queried (case insensitive) and if a translation given by informants could be found among the translations suggested by Treq, it was classified as clearly correct.

If the translations given are different from those offered by Treq, but are still reasonable, they are categorised as **paraphrase** in the evaluation. This is especially the case for certain noun and prepositional phrases. For instance, the phrase *realne možnosti awansu w firmie* ‘realistic promotion opportunities within the company’ would be best translated by *reální možnosti postupu ve firmě* in CS. When *reální* was not written down by the informants, but only *možnosti postupu*, the translation of *realne* is counted as paraphrase; when *ve firmě* ‘in the company’ was not translated, then these words were also counted as a paraphrase, as the recordings prove that people consider this a redundant information. If *odbycia* ‘undergoing [gen]’ in the phrase *możliwości odbycia interesujących praktyk* ‘possibilities of undergoing interesting internships’ has not been explicitly translated by the informants, but as *možnosti zajímavých praxí* ‘possibilities of interesting internships’, this is counted as a paraphrase. For the word *sok* ‘juice’, translations such as *extrakt* ‘extract’ or *voda* ‘water’ were counted as paraphrases. Treq suggests the translations *džus* ‘juice’, *šťáva* ‘juice’, *sirup* ‘syrup’, and *výtažek* ‘extract’, which are categorised as clearly correct. The NP *lód komsumpcyjny* ‘consumable ice’ was translated in 6 of 16 cases with only *led* ‘ice’, which is a more appropriate CS translation than *konzumní led* ‘consumable ice’ would be. Therefore, the omission of the translation of *komsumpcyjny* was counted as a paraphrase. The clearly correct translation of *miód kwiatowy* ‘blossom honey’ would be *květový med*, but also *luční med* ‘meadow honey’ is counted as a paraphrase. In the phrase *doświadczenia w pracy przy produkcji mięsa* ‘work experience in the meat production’ a translation without the explicit translation of *produkcji*, such as *pracovní zkušenosti s masem* ‘work experience with meat’ was counted as a paraphrase for the word *produkcji* and correct for the rest of the phrase.

Those words that were translated with the wrong voice, wrong person, wrong number, wrong tense, wrong mood, or a wrong derivational affix are considered **partly wrong** in the evaluation. Some examples are

- *chcielibyście* ‘would you want’, referring to a group of persons, translated as *chtěla byste* ‘would you want’, referring to a female;
- *aby stały się* ‘that they would become’ translated as *aby se stala* ‘that she would become’;
- *interesujących* ‘interesting [gen pl]’ translated as *zájmových* ‘interest-related [gen pl]’ instead of *zajímavých*;
- *czytałbym* ‘I would read’ translated as *četli by jsme* ‘we would read’ or *četl by* ‘he would read’.

If a translation given could neither be classified as correct, paraphrase or partly wrong, it is simply **wrong**. If *nie* ‘no’ is not translated, it is also evaluated as wrong and not as part of the category nothing. Those words that were not translated by the informants are categorised as **nothing**, assuming that the informants did not understand the respective stimulus word and therefore could not come up with any translation equivalent.

4.4.2. Comparison between the conditions

Figure 11 displays a comparison of the results for the different conditions with an evaluation per word and shows that the modifications led to different results.

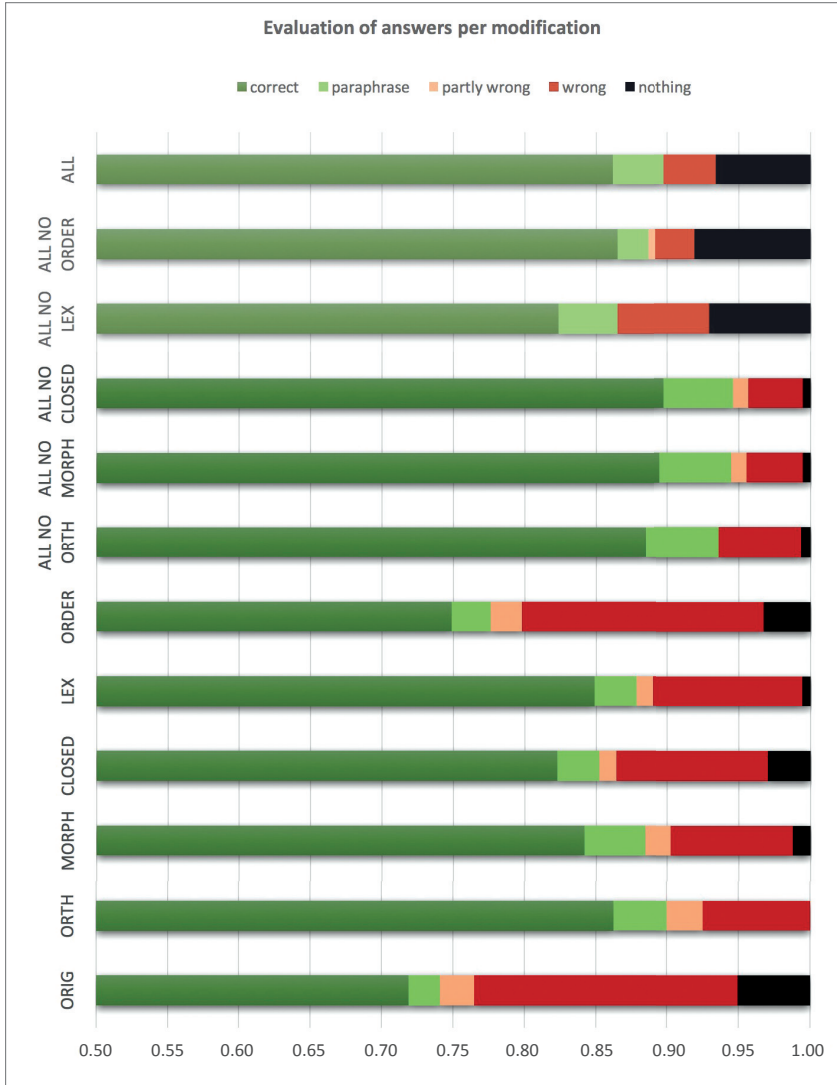


Figure 11: Results for all conditions in the cooperative translation experiments.

The results reveal that informants performed worst in the original condition, as expected. When viewing only the non-combined conditions, then the substitution of orthographic correlates (*ORTH*) led to the greatest rate of correctly¹¹ translated words (90%), followed by the substitution of morphological affixes (88.41%), lexis (87.79%), and closed class words (85.21%). An optimisation of only word order led to the lowest increase in the share of correctly translated words (77.6%) compared to 74.09% in the original condition.

When viewing the combined modifications, the best results occurred for the condition *ALL NO CLOSED* with 94.59% of correctly translated words, followed by *ALL NO MORPH* (94.44%) and *ALL NO ORTH* (93.59%). It is somewhat remarkable that the condition in which all modifications were applied at once (*ALL*) did not lead to the best results, but resulted in even slightly less correct translations (89.76%) than only the orthographic modification (*ORTH*) alone. This observation is basically open for interpretation – the reason for this unexpected result might be in the experimental setting and that informants do not expect a sentence that is declared to be “Polish” to be that similar to CS and re-interpret the sentence according to what they think “makes sense”. For the combined conditions, *ALL NO LEX* (86.47%) and *ALL NO ORDER* (88.65%) led to the lowest intelligibility results. This, on the one hand, suggests that word order might be an important influencing factor in the setting, but, on the other hand, only word order alone does not lead to any significant improvement when compared to the original condition.

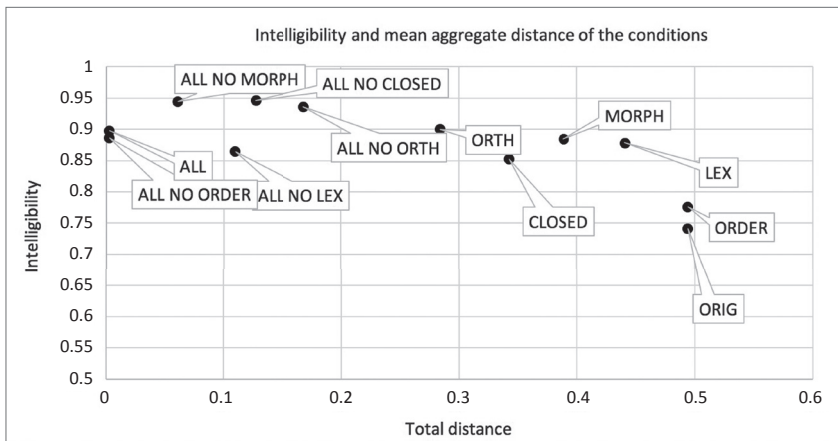


Figure 12: Correct translations (incl. paraphrases) per condition in relation to total distance.

11 Correctly in this subsection means correct answers plus paraphrases (green + dark green in Figure 11).

Figure 12 shows an alternative visualisation of the intelligibility scores in the different conditions (in addition to Figure 11) together with the mean total distances of the individual conditions. This gives an overview over the extent to which the individual modifications influenced the total linguistic distance of the stimuli and how the stimuli in the conditions differ from each other.

At first glance, a decrease in total distance is visible for every condition compared to the *ORIG* condition, except the *ORDER* condition. The total distance decreased strongest for the orthographic modification *ORTH* and also led to the highest intelligibility among the non-combined conditions. Although the *CLOSED* condition resulted in a lower total distance than *LEX*, its intelligibility was not as high as that of *LEX*. Although the distances of *ALL NO ORTH* and *ALL NO CLOSED* were highest among the combined conditions, they still led to the best intelligibility scores – together with *ALL NO MORPH* that was among the lowest in total distance. A statistical correlation of the intelligibility of the stimuli in the individual conditions and total distance is not calculated here, since it is not relevant for the present hypothesis.

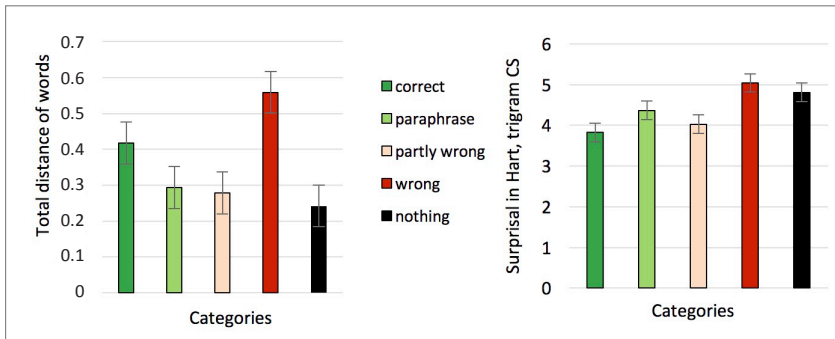


Figure 13: Total distance and surprisal among the response categories.

Figure 13 shows histograms of the mean total distances and mean surprisal values of the stimuli for the different categories of responses. It is clearly visible that the mean total distance and the surprisal values of the words that were translated wrong is highest. The words for which no response was entered (category *nothing*), however, have a relatively low total distance on the average. In contrast to that, the surprisal values of these words are relatively high compared to the other categories. Nevertheless, the differences in surprisal values between the categories are not that prominent as the differences in total distance.

4.5. Summary

An intercomprehension experiment was conducted in which Czech readers in pairs were supposed to translate different PL sentences cooperatively into CS. This study was designed in order to find out to which extent the individual linguistic levels can impair or perpetuate reading intercomprehension of PL sentences for Czech readers. In order to control for the different factors that could play a role, the stimuli sentences were modified on the different linguistic levels so that they were more similar to CS. The stimuli were presented in twelve different conditions so that each of the twelve pairs of participants translated one stimulus sentence from every condition. There were five basic (“non-combined”) modifications that were applied to the originally PL sentences: orthographic, morphological, lexical, closed class words, and word order. Additionally, combinations of these modifications were applied to the same sentences, each excluding one modification variant. There was one more condition in which all modifications were applied at once.

In a previous study, it was found that CS and PL do not differ in lexis to such an extent as they do in orthography (Jágrová, Stenger, Marti & Avgustinova, 2017). The hypothesis that the intelligibility of PL for Czech readers can be improved by modifying a PL sentence with certain CS units was tested. The hypothesis was found true for modifications on all linguistic levels, but to different degrees. When viewing only the conditions with the non-combined modification variants, then the substitution of orthographic correlates led to the greatest rate of correctly translated words, followed by the substitution of morphological affixes, lexis and closed class words. This suggests that if Czech readers were aware of the regular orthographic correspondences and knew how to apply them to PL cognates, they could understand more (about 90% in total) than without knowing these correlates. An optimization of only word order led to the lowest increase in the share of correctly translated words, compared to the condition without modification. This is in line with other findings about a limited, but existing effect of morpho-syntactic differences on mutual intelligibility of closely related languages, e.g. between DK and NOR (Hilton et al., 2013). As for the combined modification conditions, the condition in which everything but lexis was substituted (ALL NO LEX) resulted in the lowest intelligibility (86.47%), suggesting that the divergent lexis alone (closed class words excluded) accounts for about 13.5% of problems in PL-CS reading intercomprehension. This result is very close to the lexical distance of PL for Czech readers (10%) as determined in Jágrová, Stenger, Marti & Avgustinova (2017, p. 411) – see section 1.3.

This section furthermore proposes an aggregate linguistic distance measure for parallel sentence material – referred to as *total distance*. The means of correctly translated words per condition were calculated and related to the total distance of the stimuli. As opposed to findings from previous experiments and related research, it is for the present experimental setting not necessarily true that the higher the measurable distance, the lower the share of correctly translated words. In contrast to this, when viewing only the combined modifications that are very close to acceptable CS already, their intelligibility is not significantly higher than that of those conditions with a modification on only one level. This result is open for interpretation. From the audio recordings during the experiments, it becomes apparent that informants do not trust sentential contexts that are considered too “unusual” or “do not make sense”. Even if unexpected words in the sentences are perfectly transparent or even identical to CS (see section 5.5), respondents tend to dismiss correct interpretations of these words in favour of other words that “make more sense” in the context. This is especially the case in sentences that express complex situations with more than one event or agent. Here, dominant concepts in the stimuli sentences and associations with them might play an important role and the predictive power of such distance measures as orthographic and lexical distance reaches its limitations.

5. Qualitative Analysis

The modified versions of the original PL stimuli that were part of the quantitative analysis in section 4 are not discussed here. In this section, the aim is not to analyse the influence of the modifications, since the hypothesis formulated in 4.1. is not subject to this analysis. This section focusses on the unmodified stimuli and results for the modified conditions are mentioned only in especially interesting cases.

Selected passages from the recorded and transcribed experiment protocols are cited in a manner that, for instance, P8/6 means respondent pair 8, sentence number 6 (see list of sentences in Table 13). The citations are given both in CS and as translations in EN below. The CS passages are cited in exactly the same manner that the Czech native transcriber wrote them down, meaning that, for instance, non-words are represented in CS orthography as written down by the transcriber and not in phonetic transcription. The transcriber’s work was compared to the actual recordings and corrected if necessary and relevant. In some cases, a broad phonemic transcription is added in the quoted passages if the utterances cannot be translated into EN or if it is relevant in terms of how respondents pronounced a particular stimulus. The relevant sequences are marked bold.

The system of categories for this analysis was established in a mixed explorative – deductive and inductive – qualitative data analysis method (Mayring, 2010) with the help of the MAXQDA software. Some of the categories were formulated before the analysis of the results in a hypothesis-driven approach. Another part of the categories was observed during the analysis of the recordings and added to the previous categories accordingly.

The related method of think-aloud protocols in research on intercomprehension was previously used by Berthele (2011) and Möller & Zeevaert (2015), however, with individual respondents and not in a pairwise and cooperative setting. Berthele (2011) conducted a study with 163 young Swiss German native speakers who were asked to disambiguate Danish and Swedish verbs without context in written and aural condition. Verbal protocols of a subsample of the participants were recorded during the task, face to face with a field worker who wrote down the responses. The written responses were evaluated statistically with the aim to identify characteristics of an “ideal interlingual inferer” (Berthele, 2011, p. 199) and linguistic features of the stimuli relevant for inferability. For the characteristics of the ideal inferer, Berthele identified the factors “1. age (the older, the better); 2. vocabulary learning ability; 3. English proficiency; # of languages in the repertoire” statistically meaningful, accounting for 62% of the variance in the data. In a comment subsequent to the publication of the study, Berthele points out the fact that the correlation coefficient for the number of languages in the participants’ repertoire is negative, meaning that multilingualism here actually correlates “with a smaller amount of correct inferences” (p. 199). With regard to linguistic features, Berthele found a significant correlation ($r(28) = -0.416$, $p < 0.05$ for the written condition and $r(25) = -0.349$, $p < 0.05$ for the aural condition) of Levenshtein distance of the stimuli with their EN cognates (2011, p. 202), which speaks for the fact that multilinguals activate not only their L1, but rather a repertoire of acquired languages during intercomprehension. He facilitates the insights from the think-aloud protocols for support of a number of conclusions in the discussion section of the study.

Möller & Zeevaert (2015) conducted a think-aloud study with 17 German students trying to recognise cognates and text segments from other GER Lx. The participants commented on how they were proceeding during the task. They evaluate the protocols according to four different categories: (i) comments on similarity, (ii) conscious and unconscious semantic associations, (iii) words in text context, and (iv) respective roles of semantic and phonetic associations. It was found that when disambiguating words without context, the participants do not only “attribute the same importance to other associations as they do to phonetic ones – even in the recognition of isolated words semantic

connections in the mental lexical are involved” (p. 313). They point out that in text context, participants combine phonetic similarity and inference, whereby “the aspect of semantic probability manifestly overrides intuitions about phonetic similarity” (p. 313).

5.1. Readers’ Strategies

As also pointed out by Berthele (2011) and Möller & Zeevaert (2015), respondents use different strategies, be it consciously or unconsciously, to comprehend the related but unknown foreign language. Two of the most prominent strategies in the present study proved to be i) the use of placeholder words for incomprehensible words within otherwise understandable sentences and ii) repeated reading of certain difficult words with different pronunciation variants. The two strategies will be discussed in the subsections 5.1.1 and 5.1.3.

5.1.1. Leaving unknown words open and trying to infer them from the context

As already observed by Muikku-Werner (2014, p. 103) in experiments with Finnish readers translating Estonian sentences, omission of hard to comprehend words was a frequent strategy if the understanding of the greatest part of the remaining sentence was not disturbed badly. From the present recordings, this finding can be confirmed. Placeholder words were used in order to overcome difficulties. In many cases, indefinite pronouns such as *něco* ‘something’, *někam* ‘somewhere’ or *někdo* ‘somebody’ were mentioned explicitly. Some examples are shown in the following:

Respondent 1B had read the word in question – *przekonana* ‘convinced’ – which is a false friend to CS *překonaná* ‘overwhelmed’. As *overwhelmed* does not fit the remaining context of the sentence semantically, the respondent replaces it with *něco* ‘something’ and reads the whole translated sentence with this pronoun instead of the word in question:

P1/9: B: Překonaná. Tak, proč jsi **něco**, že za padesát let lidé už nebudou létat letadlem.

‘B: [reading *przekonana*]. Well, why are you **something** that in fifty years people will no longer fly aeroplanes.’

Respondent 3A proceeds in the same manner with the unknown stimulus word *brutto* ‘brutto’, which actually also exists in CS, but is not known to everybody, as the more frequent CS expression is *hrubý/hrubého* ‘gross’:

P3/10: A: Bruto nevím, ale tak prostě **něco** na hodinu.

‘A: [reading brutto] I don’t know, but simply **something** per hour.’

Sometimes this placeholder can substitute whole phrases, such as *do rektora* ‘to the rector’ in sentence 8:

P5/8: A: Tak... že bysme šli **někam. Do rektora**, no.

B: No, takže, ale prostředě... prostředku je jeho žena, na **něco** poukazuje.

‘A: So ... that we should go **somewhere. To the rector**, well.

B: Well, so, but middle... the middle is his wife, pointing at **something**.’

The presence of such placeholder words can be an indicator for difficult words or such words that can be inferred from the context of the remaining sentence. The correct use of a placeholder in a grammatically congruent form can be an indication for the correct recognition of its POS, i.e. its grammatical function, without understanding the word entirely. In the following examples, respondents used the pronoun *něco* ‘something’ correctly in genitive case – *něčeho* – in order to replace the word *towarów* ‘goods [gen]’ in sentence 2 (respondent 6B), respectively the word *książek* ‘books [gen]’ in sentence 1 (respondents 14A, 5B, and 11A):

P6/2: B: Vzrostl obrat, ne?

A: Mhm, jo, jo, jo...

B: **Něčeho**, u kterých... v procesu...

‘B: Turnover increased, or not?’

A: Hm, yeah, yeah, yeah ...

B: **Of something**, with which ... in the process ...’

This strategy was used more often for certain sentences, for instance in sentence 1:

P14/1: A: Kdyby **něčeho** nebylo, tak by to schytal.

‘A: If there was no **something**, he would get his fair share.’

P5/1: B: Tak... Kdyby ně bylo ksiasek, čítal-bym či z oču. Tak kdyby nebylo **něčeho**...

A: Knížek, knížek.

‘B: So ... [reading ...]. So, if there were no **something** ...

A: Books, books.’

P11/1: A: [...] Kdyby nebylo **něčeho**, četl bych ti z očí.

‘A: [...] If there was no **something**, I would read from your eyes.’

In the following case, the respondent is first using a placeholder word, then mentions what the actual word *jeziorem* ‘lake [instr]’ from the stimulus “looks a lot like”, but replaces it with the placeholder again, which might be a sign of insecurity about the correctness of the translation:

P8/3: B: Dobře, dobře, projekt u **něčeho**. U če- u čeho by to bylo? U něco místo a vypadá to hodně jako **jezera**. [...] Zrealizoval ten projekt u, u **něčeho**, já nevím.

‘B: Good, good, project at **something**. At wh- at what could it be? At something place and it looks a lot like **lake**. [...] Implement that project at, at **something**, I don’t know.’

5.1.2. Recognition order as indicator for difficulty

This method for analysing the identification order of words within sentence stimuli is oriented on the method used by Heinz (2009) who presented audio recordings of sentences to respondents and let them note down all words identified during each turn of listening. In Table 15, this method is adapted to reading. Some words have been correctly recognised by the respondents right away during the first reading, while other words were recognised only after several attempts of reading the whole sentence, if they were recognised at all. This informs us about the difficulty of certain words. The more difficult words are recognised last (or not recognised at all). Table 15 visualises the recognition order in sentence 7: **Green cells** are correct, **light green cells** are paraphrases, **blue cells** are placeholders and **red cells** are wrong translations.

P	Teraz	rosną	również	możliwości	odbycia	interesujących	praktyk	w	kraju.
			rovněž						
	teď	něco							
							něco	v	krajích
					nějakých	zajímavých			
						zajímavých	událostí?		
						zajímavé	události	v	kraji
					s bicyklem?				
14		něco	rovněž						
			rovněž	můžete?					
				můžete	být svědky	zajímavých	událostí?		
		rosna?							
		tohoto jara?							
		letošního nebo nynějšího							
		růže, kytka, rosa?							
		květen?							
		jaro							
		Letošní jaro	rovněž	můžete	být svědky	zajímavých	událostí	v	kraji.
16								v	kraji
	teď								
			rovněž						
						zajímavých	praktik		
							zvyků?		kraj?
									země?
									země
				myslivost					
				se může					
			zároveň						
			možnost						
	teď	máme	zároveň	možnosti	poznávat	zajímavé	zvyky	v	zemi

P	Teraz	rosną	również	możliwości	odbycia	interesujących	praktyk	w	kraju.
3	ted'								
		nevím	rovněž						
				možnosti					
					nevím				
						zajímavých			
								v	oblasti
									země
					ukončení				
					udělat				
					provádění				
							věcí		
					vykonávání				
					provádění				
					zabývání se				
		roste							
Teď	roste	rovněž	nabídka		zajímavých	činností	v	oblasti.	

Table 15: Example: recognition order of words within stimuli sentences.

The verb form *rosną* was definitely regularly the last word that was disambiguated in sentence 7. However, the overall meaning of the sentence can be captured without understanding the exact meaning of the word and with just replacing it by a form of *být* 'to be', which many respondents did. Respondent pairs who pronounced it /rosna/ by simply ignoring or omitting the diacritics (97.3%, see Table 15 and 5.7.1), provided a number of different (intermediate) responses, such as *rovná* 'straight [fem]', *různá* 'various', *zrovna* 'right now', *letošní* 'this year's', *nynější* 'present', *růže* 'rose', *kytka* 'flower', *rosa* 'dew', *květen* 'May', *jaro* 'spring', but also correct *rostou* 'they grow' – this huge variance on the one hand reflects a great entropy about the POS and the meaning of this verb form. On the other hand, it displays the associations that respondents have with *rosna* which can be considered a non-word in CS.

The respondents in pair 3 put two placeholders in the positions of the two words they did not understand – *rosną* and *odbycia*. They then continue their discussion with speculations and try to find suitable synonyms for the noun *odbycia* until they finally translate *rosną* as *roste* 'it grows' (see Table 15).

P3/7: A: No, to by mohlo být. Ono to totiž strašně zní jako *rosna* a je to zavádějící, ale...

‘A: Well, that could be. This totally sounds like /rosna/ and it’s misleading, but ...’

One of the words that was most difficult to pronounce was *pięćdziesiąt* ‘fifty’ – an indicator for this might be the many variants in which it was pronounced by the different respondents. Nevertheless, its meaning was relatively easy to guess – many respondent pairs replaced it by its CS equivalent *padesát* right after the first reading attempt, before even reading through the rest of the sentence – maybe because it is so difficult to pronounce:

P16/9: B: [...] Tak jo. Skad jesteš překonaná, že za pięcdzjesat’ – za padesát let lidé ...

‘B: [...] Alright. [reading PL till *pięćdziesiąt*] – in fifty years people [reading already in CS] ...’

The ease of understanding *pięćdziesiąt* might, of course, be due to the low neighbourhood density of the word, that is a low number of possible other words with minimal differences.

P6/9: A: Jesteš překonana...

B: Ježiš, co to je?

A: ...že za pieč- piedziesianc lat ludzie ně bjendza już latali samulotem? Ou, to je nějaký složitý.

B: Ou.

A: Pješdziešanč je padesát. Padesát let je pješčdžišanc lat. Lidi... ne...

B: Jo, lidi, ne...

A: Bjendza, ne- ne- ty jo, nevím.

B: Tědka, co je *běda*?

‘A: [reading]

B: Jesus, what is this?

A: [reading on]? Oh, that is somehow difficult.

B: Oh.

A: [reading *pięćdziesiąt*] is fifty. Fifty years is *pięćdziesiąt* lat. People ... don’t ...

B: Yes, people, don’t ...

A: [reading *będa*] not, not, man, I don’t know.

B: Now, what is [reading *będa* as *běda* ‘woe’]?’

5.1.3. Reading again and pronouncing differently

A strategy of flipping and trying different modifications of vowels was previously observed and presented in an example with Swiss German native speakers trying to disambiguate the SWE verb *skulle* ‘should’ in a study by Berthele (2011). Evidence for this strategy can be found, for instance, with pair 5 and pair 2 trying different ways of how to pronounce *zagražajících* ‘harmful’ and *węzeł* ‘knot’:

P5/2: A: Za... zagražaj... za... počkej, jak to přečíst, **zagrázaja...**
zajucích... **zagrážajúcích.**

B: Látek za- **zagražajoucích**, za- **zagražajacích...**

A: Zahraža... To je jako, mně se to zdá jako *zabraňujících* nebo něco takovýho...

‘A: [reading] ... wait, how to read that, [reading in different variants].

B: Substances [reading with different variants].

A: [reading] ... That’s like, it seems to me like preventing or something like that ...’

P2/4: B: Praga to vazni komunikacyjny **vezel. Vezejl...** To měkký L neumim říct.

‘B: [reading sentence 4, reading *węzeł* again differently] ... I cannot pronounce this soft L.’

Flipping characters and/or sounds, such as here with pair 5, can be a good strategy when encountering cases of metatheses, e.g. in *zólty* vs. *žlutý* ‘yellow’ in which the order of the corresponding sounds *ól* vs. *lu* is divergent due to the historical metathesis of liquids:

P5/6: A: Též starý **žlotý...** **žoltý.**

‘A: Also an old /zloti:/ ... /zolti:/.’

Interestingly, respondent 5A pronounces the word first with an order of sounds that is more similar to the CS translation – *lo* and only then *ol* –, which suggests that the respondent has understood the word already during reading.

5.2. Source of Successful Transfer

Among the cases of successful inference processes, three sources of transfer could be identified and will be distinguished in the following subsections: the L1 (CS), non-standard CS, and acquired languages (Ln) in the respondents’ repertoire.

5.2.1. Inference processes from non-standard CS

One of the respondents mentioned to speak the dialect typical for the Ostrava region – the Moravian dialect. East Moravia is an area close to the Polish border and the Moravian dialect shares some common lexical and morphosyntactic features with SK and PL (cf., for instance, Karlík et al., 2002).

For instance, in sentence 8, the NP *do rektora* ‘to the rector’ caused confusion for some of the respondents. In standard CS, the preposition *do* carries the meaning of ‘into’. In the Moravian dialect, however, the preposition *do* is occasionally used to express a movement to a destination and even to a person, for instance *idu do doktora* ‘I am going to the doctor’ (Kosek, 2014, p. 96) – and so it does in PL. Respondent 8A attempted to explain this dialectal phenomenon, referring to the NP *do rektora* which would be correctly translated *k rektorovi* in standard CS, but is understandable through similar constructions with the preposition *do* in the Moravian dialect. Nevertheless, instead of choosing an example in which a movement towards a person is expressed, the respondent chose an example that is considered correct in standard CS, too. Still, the explanation and the inference process are interesting:

P8/8: A: Oni říkají, to říkáme i my v Ostravě, že jdeš **do** prostě... **do** bazená nebo tak, prostě, **to znamená jako kam.**

‘A: They say, we also say that in Ostrava, that you simply go **into** ... **into** the swimming pool or so, simply, **that means like where.**’

The other respondent pairs who have not mentioned to have an Eastern dialectal background handled this NP in two different ways. Either the divergence in the preposition did not pose any problem – 11 of the 16 respondent pairs decided to translate *do rektora* with a phrase containing a form of *rektor*: *za rektorem* or *k rektorovi* – both meaning ‘to the rector’ (correct). In the other cases, the preposition was dominant in a sense that respondents expected the noun to be an institution or a building that can be entered and thus modified the original *rektor* to *doprava* ‘to the right’ (n = 2), to the more frequent *řediteli* ‘to the headmaster’ (n = 2) or *za učitelem* ‘to the teacher’ (n = 1).

5.2.2. Inference from languages other than CS

In order to overcome lexical difficulties, it is expected that the words within the stimuli sentences listed in Table 16 require the knowledge of an Ln transfer base:

Ln	Words within stimuli	CS
SK	<i>kraj</i> 'country'	<i>země</i> , partial cognate of CS <i>kraj</i> 'region'
	<i>teraz</i> 'now'	<i>ted'</i>
RU	<i>samolot</i> through <i>samolet</i> 'plane'	<i>letadlo</i>
	<i>sok</i> through <i>sok</i> 'juice'	<i>šťáva, džus, sirup</i>
DE	<i>szt.</i> (abbreviation for <i>sztuka</i> 'piece') through <i>štuka</i> 'piece'	<i>ks.</i> (abbreviation for <i>kus</i>)
EN	<i>szt.</i> (abbreviation for <i>sztuka</i> 'piece') through <i>Stück</i> 'piece'	<i>ks.</i> (abbreviation for <i>kus</i>)
EN	<i>awans</i> 'advance'	<i>povišení</i>

Table 16: Expected Ln transfer bases for certain words within the stimuli.

Vanhove & Berthele refer to such Ln transfer bases as suggested in Table 16 as “supplier languages” (2015, p. 2). According to their results, the LD “between an Lx stimulus and a known cognate in German or English [...] is the most important item-related predictor of cognate guessing accuracy” (2015, p. 20), suggesting that respondents do not only rely on their L1 but also on the supplier languages. In other words, if respondents have indicated the knowledge of one of the possible supplier languages, they are likely to provide the correct translation of the words within the stimuli. Table 17 gives an overview of the Ln skills as indicated by the respondents in the sociodemographic survey before the experiment and the words within the stimuli that require Ln transfer bases. Pair 8 is not included, because all of the critical words were substituted for cognates in their stimulus set.

P	Ln skills indicated			Words within stimuli requiring Ln transfer base				
	RU	DE	EN	<i>awans</i>	<i>samolot</i>	<i>sok</i>	<i>szt.</i>	<i>teraz</i>
1A	0	42	100	✓	✓	✓	✓	n/a
1B	0	68	0	✓	✓	✓	✓	n/a
2A	0	86	100	✓	✓	X	X	✓
2B	0	4	98	✓	✓	X	X	✓
3A	8	0	68	✓	n/a	✓	X	✓
3B	38	0	78	✓	n/a	✓	X	✓
4A	0	54	11	✓	✓	n/a	n/a	n/a
4B	0	32	90	✓	✓	n/a	n/a	n/a
5A	0	72	76	0	n/a	n/a	n/a	✓
5B	0	60	70	0	n/a	n/a	n/a	✓
6A	0	68	68	n/a	✓	n/a	✓	✓
6B	0	50	66	n/a	✓	n/a	✓	✓
7A	0	0	50	✓	✓	n/a	n/a	✓
7B	0	0	54	✓	✓	n/a	n/a	✓
9A	0	0	66	✓	✓	n/a	n/a	n/a
9B	0	0	94	✓	✓	n/a	n/a	n/a
10A	0	56	84	n/a	n/a	✓	X	n/a
10B	0	0	70	n/a	n/a	✓	X	n/a
11A	50	0	100	n/a	✓	✓	✓	n/a
11B	0	72	76	n/a	✓	✓	✓	n/a
12A	0	32	94	n/a	n/a	✓	X	n/a
12B	16	80	78	n/a	n/a	✓	X	n/a
13A	0	0	42	n/a	n/a	n/a	n/a	✓
13B	0	8	64	n/a	n/a	n/a	n/a	✓

P	Ln skills indicated			Words within stimuli requiring Ln transfer base				
	RU	DE	EN	<i>awans</i>	<i>samolot</i>	<i>sok</i>	<i>szt.</i>	<i>teraz</i>
14A	0	0	100	✓	✓	✓	✓	✓
14B	0	4	62					
15A	0	38	74	X	✓	✓	✓	✓
15B	50	0	70					
16A	0	0	100	✓	✓	✓	X	✓
16B	0	0	100					
Correct in free translation experiment (%)				n/a	21.9	n/a	5.9	90.6

Table 17: Ln reading skills indicated by respondents and (partial) non-cognates requiring an Ln transfer base.

A ✓ sign in Table 17 means that the word was translated correctly, an X means that the response was wrong, and 0 means that no response was given for this word. Not applicable (n/a) indicates that the stimulus was presented in one of the modified conditions (section 4.3) and therefore cannot be compared here. The colour code in the background of the cells indicates whether the prediction matches the correctness of the response: a green background indicates a correct prediction while a red background indicates a wrong prediction. For instance, it was correctly predicted that pair 1 would translate *szt.* ‘piece’ correctly, because at least one of them indicated to have some knowledge of DE. In contrast, it was wrongly predicted that pair 1 would not be able to translate *sok* ‘sap’ without any knowledge of RU. In total, 21 of the predictions in Table 17 were correct and 7 were wrong. This, however, is not an objective measure for an analysis, since, among other factors, the similarity of the Lx stimulus to an Ln and also the respondents’ level of Ln command play a crucial role here.

Intelligibility scores of these words from the free translation experiments are given in the last line of Table 17 for a comparison. In the following, the successful inferences drawn from SK, RU, DE, and EN are listed and explained.

- **Inferences from SK:**

Some of the respondents were able to draw lexical inferences from SK when encountering the PL words *kraj* ‘country’ and *teraz* ‘now’. The noun *kraj* carries the meaning of ‘region, area’ in standard CS – the Czech Republic is divided into 16 administrative units – *kraje*. Therefore, Czech readers are likely to associate the concept of region or area with the word *kraj* in the sense of region and not with a whole state or country. Nevertheless, according to the Dictionary of Standard Czech, *kraj* also retained its meaning synonymous to *země* ‘country’ (Havránek, 1964, as cited in Šmerk et al., 2009). However, The Internet Language Reference Book which also contains data from this Dictionary draws attention to the fact that the Dictionary “was published in

the 1960s” and “the information given in it, which in some cases is perceived as obsolete, complies with the form existing at that time” (Šmerk et al., 2009). According to the online corpus tool Treq, PL *kraj* is translated most frequently as CS *země* ‘country’ (84%) and only relatively rarely (0.4%) as CS *kraj* in the sense of ‘country’ (Škrabal & Vavřín, 2017) – for instance, *w drodze do obcych krajów* ‘way to foreign lands’ is translated as *cesta do cizích krajů* in Ajtmatov’s *Scaffold* (as cited in Škrabal & Vavřín, 2017). Pair 3 discusses the possibility that *kraj* means the whole country through the knowledge of the word *krajina* ‘country’ in SK:

P3/7: A: No, no, no, no. Anebo, viš co, **slovensky je jako země krajina**.
No, ale to zas asi ne.

B: No dobře. A definitivně **to nebude kraj**, protože oni maj vojvodstva.

A: Jakože **jestli by to nemohla být celá země**, ale, ale to nevím.

‘A: Yeah, yeah, yeah, yeah. Or, you know what, **country is krajina in Slovak**. Well, but maybe it’s not that.

B: Alright. And this is definitely **not kraj**, because they have voivodeships.

A: Well, **this could be the whole country**, but, but I don’t know that.’

The adverb *teraz* ‘now’ can be considered as commonly known among the Czech population by exposure to SK through the media and popular culture (Nábělková, 2007). As expected, 90.6% of the respondents translated *teraz* ‘now’ correctly (CS *ted’*), most probably through their exposure to SK *teraz* which is identical to PL *teraz*. The pairs 2 and 16 discussed whether *teraz* could mean the same in PL as it does in SK:

P2/7: A: A nebo... hm, **teraz bude snad ted’**, prostě snad, nevím, se **slovenštinou**.

B: Myslíš s **poľštinou**, jo?

A: Rovněž by mohlo bejt rovněž. Počkej ted’, rosna, no ne **teraz** je v, v tom, ne? **Ve slovenštině. Tak to by mohlo bejt stejný jako v poľštině**.

‘A: Or ... hm, [reading *teraz*] **is probably now**, simply like, I don’t know, with **Slovak**.

B: You think with **Polish**, yeah?

A: [reading *również*] could be also. Wait now, [reading *rosną*], well now **teraz** is in, in that, isn’t it? **In Slovak. Now that could be the same as in Polish.**’

P16/7: B: Tak tam rozumím, v kraji, na konci, ty poslední dvě slova. **Teraz, tak teraz bude teď. To je to stejné ve slovenštině, ne?**

‘B: Well, there I understand in the region, at the final position, the last two words. [reading *Teraz*], well **Teraz is now. That’s the same as in Slovak, isn’t it?**’

The following example demonstrates a discussion about the divergent use of the negation particle *nie* ‘no, not’ in PL, respectively *ne-* in CS and its use in SK. While the negation particle is attached as a prefix to the verb form in CS, its position in PL is separate from and in front of the verb form. The respondents in pair 16 are quoting *ja som neni* ‘I am isn’t [literally]’ which is non-standard SK for ‘I am not’ (standard SK would be *ja nie som*). They manage to comprehend the negation function of the particle, probably by knowing that there is some variation in the construction of negations in the closely related languages:

P16/6: B: Protože **ve slovenštině... ve slovenštině se říká, ja som neni, ne ja nejsem**. Takže jestli **kupili sme ně**, tak možná to bude *nekoupili jsme*.

A: Jo...

B: Terazky som majorom, no.

A: Terazky som majorom, přesně...

‘B: Because **in Slovak ... in Slovak they say I am isn’t, not I am not**. So, if [reading *kupilišmy nie*], then maybe this is *we didn’t buy*.

A: Yeah ...

B: [quoting from a popular film ‘Now I am a mayor’], right.

A: [repeating the quote], exactly.’

The divergent SK form that respondent 16A most probably had in mind could be the third person singular form *nie je* ‘is not’ which would be *není* in CS – the negation of the verb *být* ‘to be’ is irregular in CS.

Pair 8 discussed the negation particle in the relative clause of sentence 2. There is no verb form in the relative clause and *nie užívano* ‘is not used’ would be CS *není užíváno* with the irregular negated third person singular form of *být* ‘to be’. Pair 8 transforms the PL negation particle into the SK equivalent that is orthographically less distant to the PL particle than *není*.

P8/2: A: Jo, jo. **To je jak slovenština, že?** [...] V procesy produkce, ktorých **ňuje** užíváno substancí [...]

‘A: Yeah, yeah. **That’s like Slovak, isn’t it?** [...] in the process of production, of which substances [saying SK *isn’t*] used.’

- **Inferences from RU:**

Lexical inferences were made for the stimulus word *sok* ‘juice’, which is also *сок* (*sok*) in RU, but can be translated as *šťáva* ‘sap’, *džus* ‘juice’ or *sirup* ‘sirup’ in CS (tolerating minor semantic differences in the characteristics of the different beverages).

P3/12: A: [...] Pět kostek, **sok** z břízy, nevím, co je **sok** z břízy.

B: **Sok... rusky, tuším, sok je, je džus**, ale nejsem si jistý. Ale to může bejt šťáva, šťáva z břízy. Neslyšel jsem, že by to někdo pil nikdy.

‘A: [...] Five cubes, [reading **sok**] from a birch, I don’t know what [reading **sok**] from a birch is.

B: [reading **sok**] ... **Russian, I guess, sok is, is juice**, but I’m not sure. But it can be sap, birch sap. I’ve never heard that someone would ever drink that.’

The same pair of respondents was aware of the false friend *czerstwy* ‘stale’ through the knowledge of RU *čerstvyj* ‘stale’. Both are explicitly mentioning that they know that it means the opposite of CS *čerstvý* ‘fresh’. The interesting thing here is that they were actually presented the lexically modified version of sentence 6 in which *czerstwy* was substituted by *twardy* ‘hard’ – a cognate to CS *tvrdý* ‘hard’. Still they were discussing what PL *czerstwy* means:

P3/6: A: **Čerstvý**, to znamená polsky jakože opak, jakože **starý**. [...]

B: Já jsem to tušil, protože **je to úplně stejně v ruštině**. Já jsem, já jsem si říkal, ano, protože zaprvé, kdo by kupoval tvrdý chléb, kdo by kupoval staré auto. I když to staré, staré třeba to mi tam mate i nepřijde. Takže, já jsem si taky říkal, že... já úplně teďka jako nevím, jak je to v polštině, ale vím, že je to tam... jo, **že čerstvý znamená naopak starý**.

‘A: **Čerstvý**, that means like the opposite in Polish, like **old**. [...]

B: I thought so, because **this is absolutely the same in Russian**. I was, I thought to myself, yes, because first, who would buy stale bread, who would buy an old car? Although this old, old maybe that even seems weird to me, mate. So, I was also thinking that Right now, I don’t really know how it is in Polish, but I know that it’s there ... yeah, **that czerstwy on the opposite means old**.’

Regardless of this intermediate discussion, the pair came up with the correct translation for the word in question.

Apart from lexis, differences in prefixes were successfully overcome by respondent pair 1 when translating *pokazuje* ‘[she] is showing’ in stimulus sentence 8 which would be *ukazuje* in CS. From an orthographic or morphological perspective, words that are less distant to the PL stimulus in a reader’s Ln than in CS are expected to be inferred more easily if respondents can resort to this Ln. The prefix *po-* instead of *u-* in combination with the stem *kaz* is apparently previously known from RU, although RU is not explicitly mentioned here:

P1/8: B: [...] **Pokazaj**, není to něco jako **ukaž**? Aby jeho... **ukázala**.

‘B: [...] [saying RU **pokazaj**], isn’t that something like **show** [imperative sg]? That his ... **should show**.’

RU *покажи* (*pokaži*) ‘show’ is the imperative singular of *показать* (*pokazat’*) ‘to show’ – the CS translations being *ukaž* and *ukázat*. It is especially interesting here that the verb form in question is in another tense, person and mood, but the respondents are still able to infer its correct meaning by the knowledge of the corresponding prefixes in a known RU verb form, without mentioning it explicitly. However, this pair did not enter the correct translation, but decided to enter *Nepřeji si, aby jeho žena navrhovala [...]* ‘I don’t wish that his wife suggests [...]’ as their written response.

From the many cognates that are spelled differently in PL, the word *żółty* ‘yellow’ was correctly recognised as *żlutý* ‘yellow’ through RU *žoltyj*, even though respondent 3B pronounces it in a wrong way (wrong order of the sounds /l/ and /o/), explicitly mentioning the inference from RU:

P3/6: B: Jako to, to **żłoté... Rusky je to taky żłoty**. Hm. Co ty na to?

A: [...] Jakože **żlutý**? Hm.

‘B: Like this, this [reading **żółty**] ... is also [reading **żółty**] in Russian. Hm. What do you think of that?’

A: [...] Like **yellow**? Hm.’

This observation supports the finding that orthographic distance of another closer cognate from an Ln is a better predictor than a more distant cognate or non-cognate from L1. It also indicates that readers are able to draw inferences through cross-lingual correspondences. Here, the metathesis of liquids rule *ól:ol:lu* (PL:RU:CS) can be applied through another Ln.

Syntactic inferences could be made when translating the sentence *Praga to ważny węzeł komunikacyjny*. ‘Prague is an important traffic hub.’ Instead of a verb form, there is only the demonstrative pronoun *to* ‘this’. The pronoun *to* also exists in CS, but cannot replace a finite verb in a sentence as in the example here. Acceptable CS translations would in this case be *Praha je ...* ‘Prague is a ...’ or *Praha, to je ...* ‘Prague, that is a ...’. The absence of a finite verb

form of *to be* in the sentence might be known from RU. Although none of the two respondents indicated any knowledge of RU in the self-assessment, pair 1 mentions to be aware of this syntactic phenomenon in RU:

P1/4: B: Hm, dobrý. **To je myslím v ruštině, že tam můžou vypouštět [...] slovesa.**

‘B: Hm, good. **I think it’s in Russian where they can omit [...] verbs.**’

This absence of the finite verb form and how respondents handled this is further discussed under 5.8 *Talking About Grammar*.

- **Inferences from Non-Slavic languages:**

An example for a successful inference of a lexically different stimulus word from non-Slavic languages occurred with the respondent pairs 8 and 11 when they tried to decipher the abbreviation for *sztuka* ‘piece’ – *szt.* The correct translation of *sztuka* would be the CS non-cognate *kus*, which is abbreviated as *ks*. PL *sztuka* is a loanword from DE *Stück* in which the original umlaut *ü* is represented by a *u* in PL. There is a very infrequent Germanism (0.02 i.p.m. according to the CNC) in CS – *štyk* – in which the originally German umlaut is represented by a *y*. Therefore, it might not be so transparent to respondents without any knowledge of DE. CS *štyk* also occurs in the compound loanwords *kunstštyk* ‘piece of art’ (from DE *Kunststück*) and in the more frequent *majstrštyk* ‘masterpiece’ (0.37 i.p.m. according to the CNC) from DE *Meisterstück*. Pair 8 inferred the correct translation of the PL abbreviation *szt.* through both DE *Stück* and a decomposition of CS *majstrštyk*:

P8/12: B: Kvjatový mjod padesát gramů, cytryna je- jeden š...

A: Jak je to? **Jak je německy kus? Protože oni berou hodně z...**

B: **Štuk.**

A: **Štyk** nebo? **Štyk** myslím, nevím teďka. No, to je jedno.

B: *Štyk, štyk, štyk* jo, to je, to je možný. **Jako majstrštyk – mistrovskéj kousek.** Takže *štyk* by to mělo být.

‘B: [reading *kwiatowy miód*] fifty grams, [tsitřina] o- one [ʃ] ...

A: How is it? **What’s piece in German? Because they take a lot from ...**

B: [reading [ʃtuk]]

A: [ʃtuk] or? [ʃtuk] **I think**, I don’t know now. Well, doesn’t matter.

B: [ʃtik], [ʃtik], [ʃtuk] yeah, that’s, that’s possible. **Like majstrštyk – master piece.** So that should be *štyk*.’

It would have been possible for respondent pair 11 to infer *sztuka* from DE *Stück* ‘piece’, too, because one of the two respondents indicated reading skills of 72/100 in DE in the self-assessment questionnaire part of the survey. Instead, they infer the correct meaning from the EN abbreviation for *piece* – *pc*:

P11/12: B: No a citrón jeden *st*. Set?

A: Ne.

B: Když to přečteš, tak je to *s t*, že jo.

A: No jo, ale to je **zkratka**, že jo, nezapomeň, že tam je tečka za tím. [...] To právě nikdo neví, že jo. Já si myslím, že to fakt bude jako jeden, jeden kus prostě. [...] **Že to bude jako *pc* v angličtině...**

‘B: Well and lemon one [reading *szt.*]. Set?’

A: No.

B: If you read that, then it’s /s - t/, right.

A: Well yeah, but it’s an **abbreviation**, right, don’t forget that there is a full stop after it. [...] Nobody knows this, right. I think that this is really one, simply one piece. [...] **That’s like *pc* in English.**’

Both respondents had indicated better skills in EN (76 and 100) than in DE, which might explain the dominant role of EN here.

In the following case, the respondents were also able to infer the correct meaning of PL *sok* ‘juice’ through EN *sap*. Although the words linguistically are not considered to be cognates, they have word length and initial letter in common and therefore might evoke some association. It is also well possible that respondent 16B who is a native speaker of CS, but lives in Great Britain, was exposed to the word *sok* ‘sap, juice’ for instance through PL labels of packages of juice that can be bought where she lives.

P16/12: B: **Sok** z břozem... **Sok nebude jako... *sap* jako sirup** z bezu? Jako bezový...

A: Jo, ty seš, ty seš dobrá, ty jo, jasně, no. To úplně ted’ka jak to řekneš, tak to úplně dává smysl.

B: Hm...

A: Sirup z...

B: **Ale zajímavý teda, že ted’ka, jak jsem to odvodila spíš z angličtiny než z češtiny. Jako *sap* jako *sap* v angličtině.**

A: Jo jasně, no, jasně, jasně.

‘B: [reading *sok z brzozy*] ... **Sok is not like ... *sap* like sirup** from elderflower? Like elderflower ...

A: Yeah, you're, you're good, wow, sure, yeah. Totally just as you're saying, that totally makes sense.

B: Hm ...

A: Sirup from ...

B: But it's interesting actually that now that I have inferred it more from English than from Czech. Like *sap*, like *sap* in English.

A: Yeah sure, well, sure, sure.'

5.3. Knowledge of Non-Cognates and Awareness of False Friends

In previous research, non-cognates (profile words) were included in free translation experiments to test whether the respondent indeed has no substantial prior knowledge of the experiment language and had not lied during the self-assessment of language skills. Vanhove (2015) ascribes the few correct translations of profile words "to a small degree of incidental learning, e.g. during holidays or due to popular culture" (p. 68). The fact that some respondents knew about false friends can be an argument for not removing false friends from stimuli sets.

There also seems to be some awareness about cross-lingual lexical differences:

P15: A: No jako jo, ale tak Slováci maj taky plno slov, který vůbec příbuzný češtině nejsou. Jakože většina jo, ale.

'A: Well, like, the Slovaks also got loads of words that aren't related to Czech at all. But most of them are, yeah.'

P16: A: ...v té polštině ty slova jsou podobný, ale jako znamenají diametrálně jiný věci [...].

'A: [...] words are similar in Polish, but they often mean vastly other things [...].'

There were cases in which respondents knew about some of the PL-CS false friends and thus they managed to comprehend these in the stimuli **successfully**. The most frequently known of these false friends seem to be *czerstwy* 'stale' and *sklep* 'basement':

- *czerstwy* ‘stale’ is a false friend of CS *čerstvý* ‘fresh’:

P6/6: A: Tak tady je zrada. *Čerstvý chléb* není vůbec *čerstvý* chléb, ale měl by to být *zkažený* chléb. **To je jedno z těch slov, který právě, jak jsem se bavila s tou kamarádkou z Krakova. A ona mi říkala, že to... je vtipný slovo, že jako *čerstvý* u nich znamená...**

‘A: So this is treason. *Čerstvý chléb* [reading] is not fresh bread at all, but this is supposed to be rotten bread. **This is exactly one of these words that I talked about with my friend from Cracow. And she said that ... it’s a funny word, that *czerstwy* for them means ...**’

P8/6: A: Hej, tohle slovo zrovna vím, kámo, protože my jsme se jednou bavili to, s jedním Polákem a jakože jaké slova máme různé a on přímo říkal, že *čerstvý* znamená u nich... že *čerstvý* znamená u nich prostě opak.

‘A: Hey, I know this word, mate, because we once talked that, with a Polish guy about which words are different and he directly said that *czerstwy* means ... that *czerstwy* simply means the opposite for them.’

- *sklep* ‘shop’ is a false friend of CS *sklep* ‘basement’:

P3/11: A: [...] Mhm. Obsluha **sklepu**.

B: O tom jsem slyšel, to je **obchod**.

A: [...] Hele já vím, co je *sklep*, tak si zkus tipnout ty, protože nevím, ty to možná nevíš.

‘A: [...] Mhm. [reading *obsługa sklepu*].

B: I heard about that, that’s a **shop**.

A: [...] Look, I know what *sklep* is, try and guess, because I don’t know, maybe you don’t know it.’

P12/11: B: [...] Tak to je obsluha obchodu, ne? **Sklep je obchod, to nám říkali na občance.**

A: Fakt? Ty jseš dobrá, to bude ono.

‘B: [...] So that’s service in a shop, isn’t it? **Sklep is a shop, so they told us at civic education.**

A: Really? You’re good, that’s going to be it.’

P14/11: B: **Sklep je obchod, ten sklep je obchod**, no. To je jediné, co si pamatuju, teda.

A: Aha, no jasně.

B: **To vím, že *sklep* je obchod.**

A: Jasně, to jsi mi vlastně říkala.

‘B: ***Sklep is a shop, that sklep is a shop***, yeah. That’s the only thing I remember, right.

A: Aha, well, sure.

B: **I know that *sklep* is a shop.**

A: Sure, you said that to me actually.’

The following pair actually did not mention to have known the word *sklep* as a false friend before, but it seems as if they had successfully inferred it from the context:

P13/11: B: Počkej, tak obs... obsluha sklepu? Já myslím, že **ten *sklep* bude znamenat něco jinýho než sklep.**

A: Já si taky myslím. Obzluga a jako obsluha jako, myslíš, že to je obsluha?

B: No, to nevím, to bysme potřebovaly vědět, co znamená ten *sklep*. Obsluha na...

A: Sk... obsluha, ale to by, jakoby, obsluha, sklep? Obsluha sklepu... Sklep? Vinný sklep?

B: No a budou tam vystavovat nějaký výrobky ve vinnym sklepu.

A: Ale třeba by to mohlo být, když jako vinný sklep, tak tam může jakoby vino... **to je obchod. Co když je ten *sklep* obchod? Obsluha obchodu...** to by dávalo, to by dávalo smysl, ale obsluha sklepu...

‘B: Wait, so, [reading *obsluga sklepu*]? **I think that *sklep* means something else than basement.**

A: I think so, too. [reading] and like service, like, do you think this is service?

B: Well, I don’t know, we’d need to know what that *sklep* means. Service for ...

A: Sk ... service, but that, like, service, basement? Basement service ... Basement? Wine cellar?

B: Yeah and there they will display some products in that wine cellar.

A: But that might be, if that’s like a wine cellar, there could be like wine ... **that’s a shop. What if that *sklep* is a shop?** Service in a shop ... that’d make sense, that’d make sense, but service in a basement ... [...]

It is, of course, possible that they had heard about the fact that *sklep* means something else in PL. However, they both seem to change their mind, probably due to having misunderstood the sentence onset *zakres obowiązków* ‘job description’ as *zakaz obouváků* ‘ban on shoe horns’:

P13/11: A: **Zákaz...**

B: Bot, *zázaz*, to je jako **obouváky**.

A: Ale proč by do sklepu byl **zakaz obouváku**? Znajomoc polského jazyka. Expozicja. [...]

B: No a proč, proč tam nesměj v botech?

A: Obsluha obchodu...

B: No, ale nedává smysl, proč tam nesměj v botech.

A: Nebo jako obsluha zaměsc... za- zaměstnance obchodu. *Zákres...*

B: Tak tam napíšem, co si myslíme, že to je.

A: Tak tam napiš jakoby zam- zaměstnanci obchodu...

B: **A jseš si jistá, že to bude obchod?**

A: Jó, *sklep*, to bude. Protože jakoby...

B: A nebude to jakoby zaměstnanci, jenom nějak jako zaměstnanci budovy? To bysme ale mohli napsat. **Budovy...**

A: Jo. *Sklep*, já nevím, já...

B: *Zá-* *zakaz* vstupu v botech, jo?

‘A: **Ban ...**

B: On shoes, /za:zas/, that’s like **shoe horns**.

A: But why should there be a **ban on shoe horns** in the basement? [reading *znajomość*] of the Polish language. [reading *expoziycja*].

B: Yeah, and why, why are they not allowed to enter in shoes?

A: Shop assistant ...

B: Yeah, but it doesn’t make any sense why they shouldn’t be allowed to enter in shoes.

A: Or like service emplo... em- employee of the shop. Plot ...

B: Then we’re going to write what we think it is.

A: Then write like emp- employees of the shop ...

B: **And are you sure that this is a shop?**

A: Yeah, *sklep*, that’s it. Because like ...

B: And maybe it's not like employees, just like employees of the building? But we could write that down. **Of the building ...**

A: Yeah. *Sklep*, I don't know, I ...

B: Ba- ban on entrance in shoes, yeah?

In the end, pair 13 decided to enter *budovy* 'building [gen]' as a translation at the position of *sklep*. Besides the above mentioned non-cognate *szuka* 'piece', there were other non-cognates that some of the respondents might have "heard before", but were not always able to understand:

- **samochód** 'car' which is a non-cognate to CS auto, but which as a compound consists of the cognate units *samo* 'self' (in both PL and CS) and *chód/chod* (PL/CS) 'walk(ing), motion':

P5/6: A: Ten *samochod*, to jsem někde slyšel, ale nevím, co to je.

'A: That *samochód*, I've heard that before, but I don't know what it is.'

P6/6: A: Hele ten *samochód*, to se mi zdá, že by mohlo být fakt, fakt auto, protože na auto se to hrozně často používalo, když jsme stopovali přes Polsko.

B: Určitě to tak bude, *samochód*... to je hrozně vtipný slovo.

'A: Look, that *samochód*, seems to me, that it really could be, really a car, because this was used terribly often for car the time we hitchhiked through Poland.

B: Surely it's like that, *samochód* ... that's a terribly funny word.'

Pair 6 explicitly mentioned this incidental learning of the word *samochód* as one of them "hitchhiked through Poland". This pair was also aware of the existence of false friends and discussed whether *kraj* 'region' (see section 5.2.2) was one of these words:

P6/7: A: Mhm. **A jestli kraj je kraj, že jo.** Vždycky takový ty **nejpodobnější slova vždycky znamenají něco úplně jiného**, totiž. Takže jsem, takže nevím...

B: Tak zajímavý zvyky vlastně v kraji... **Tak co by mohlo být kraj, kraj, nebude kraj, země?**

'A: Hm. And now whether [reading *kraj*] is region, right. It's always **those most similar words that always mean something completely different**, I tell you. Well, I'm, well I don't know ...

B: So, interesting habits actually in the region ... **Well, what could kraj be, kraj, couldn't it be country?**

5.4. Over-Transfer from Languages Other Than CS

While numerous successful L1 or Ln inference processes took place, also cases of over-transfer from languages other than CS occurred.

EN:

- *będzie* ‘will [3rd pers.]’ – over-transfer through EN *bad*:

P6/5: B: **Bedzie** a nemůže to být, jako že to je jakože **něco špatnýho**? Ne, ne, ne, ne.

‘B: [**bedzje**] and can’t that be like that’s like **something bad**? No, no, no, no.’

SK:

- *tylko* ‘only’ – over-transfer through what respondents think that SK *něskoro* ‘late’ means:

P1/6: B: Nebo třeba... nje tylko, jestli to neni jako *něskoro* slovensky, že to je jako málo chleba. Koupili jsme **málo, málo** čerstvýho chleba. Já nevím.

‘B: Or maybe ... [nje tilko], might be something like *něskoro* in Slovak, like little bread. We bought **little, little** fresh bread. I don’t know.’

HR:

- *gotowość* ‘readiness’ – over-transfer through HR *gotovina* ‘cash’:

Respondent 11B referred to HR and most likely meant *gotovina* ‘cash’ and interestingly did not attempt for a transfer through the possible CS cognate transfer bases *pohotovost* ‘availability’ or *hotovo* ‘ready’:

P1/11: B: To je podobný chorvatsky, nějak podobně. Takže... **hotovostní, hotovostní** operace. A činnosti pořádkové, to by mohlo být třeba jako úklidové práce.

‘B: That’s similar in Croatian, kinda similar. So ... **cash, cash** operation. And order activities, that might be something like housekeeping.’

DE:

- *do rektora* ‘to the rector’ with over-transfer through DE *Recht* ‘right’ or *rechts* ‘to the right’ instead of the identical, but infrequent CS *rektor* ‘rector’ (see also 5.5):

P5/8: B: Žě bychom mě- měli jít. **Rekt... recht** z **němčiny**, by bylo.
A: Jo, jo, jo.

B: Abychom měli jít **doprava**. To by mohlo být, mhm, něco takovýho. To zní dobře.

‘B: That we should go. /*rekt*/ ... *recht* from **German**, would be.

A: Yeah, yeah, yeah.

B: That we should go **to the right**. That could be, mhm, something like that. That sounds good.’

- *glosovala* ‘voted [fem]’ with over-transfer through DE *Glas* ‘glass’ or EN *glass*:

P3/5: B: [...] To bude jako nějaké pani bude **gla- glasovala**...

A: Paní buďeš hlasovala...

B: Jo, jako by měla být **posklená** nebo něco. [...] No, paní bude ze **skla**, no to je, to je ještě horší věta. [...] Právě jsem taky uvažoval nad tím zpíváním, nevím proč.

A: **Glas**...

‘B: [...] That will be like the lady will /**gla**/ ... /**glasovala**/

A: Madam, will you vote ...

B: Yeah, like she should be covered in **glass** or something. [...] Well, the lady will be made of **glass**, well that is, that is an even worse sentence. [...] I was just thinking about that singing, I don’t know why.

A: /**glas**/ ...’

In general, cases of over-transfer from languages other than CS have occurred less frequently than the cases of correct Ln inferences explained in section 5.2.2.

5.5. Distrust in Obviously Understandable Words

There seemed to be a specific distrust in some of the internationalisms in this experiment. This does not comply with the observations from a previous study by Jágrová, Stenger & Avgustinova (2017) where it was found that in a free translation experiment with context-free Polish internationalisms and Indo-European cognates presented to German readers, the internationalisms were translated three times more often correctly than Indo-European cognates with the same orthographic distance.

This distrust might have different reasons. One possible explanation is that respondents might not be sure of the actual meaning of a foreign word or loanword for which another, possibly more frequent, CS synonym exists. In many cases, the respondents make these words briefly a subject of discussion in order to make sure they both have a similar understanding of the foreign word. In the following overview, the critical internationalisms are listed with their most frequent translations extracted with the parallel corpus tool Treq (Vavřín & Rosen, 2015), their CS corpus frequencies from the CNC – SYN2015 (Křen et al., 2015), and examples from the discussions. If available, the respective

intelligibility scores from subsequent web-based context-free translation experiments with the individual words (section 7 and 16) are listed for a comparison.

- **brutto** ‘gross’ – The most frequent translation (89.6%) into CS is *hrubý* ‘gross’ which has a corpus frequency of 42.79 i.p.m. The term *brutto* ‘gross’ also exists in CS, however, with a very low corpus frequency of 0.27 i.p.m. and, according to Treq, only 2.8% of PL *brutto* are translated as CS *brutto*.

P3/10: A: **Brutto** nevím, ale tak prostě něco na hodinu.

‘A: I don’t know about **brutto**, but it’s just something per hour.’

Pair 3 did not explicitly indicate that they know the word, but in the end decided for the correct translation *hrubého* ‘gross’ in their written response.

Pair 13 assumed that brutto is the currency of Poland:

P13/10: A: Dvanáct tisíc za hodinu. Dvanáct korun.

B: Dvanáct... Dvanáct nějakých těch polských, ne? Co já vím, čím se platí.

A: **Brut.**

B: Hm, asi.

A: **Dvanáct brutů za hodinu.**

‘A: Twelve thousand per hour. Twelve crowns.

B: Twelve ... Twelve some of these Polish, right? Who knows with what they pay there.

A: **Brutts.**

B: Hm, maybe.

A: Twelve brutts per hour.’

Pair 2 finds the right solution, mentioning what the word “sounds” like:

P2/10: B: Já bych řekl hrubého, *brutto* zní prostě **hrubě**.

‘B: I would say gross, *brutto* just sounds **gross**.’

- **cytryna** ‘lemon’ – The correct translation would be CS *citron* or *citrón* (both variants are acceptable) with corpus frequencies of 9.34 and 0.93. Besides the difference in spelling, PL *y* vs. CS *i* in the stem which is one of the regular PL-CS correspondences found in internationalisms, *cytryna* and *citron* differ in their grammatical gender. The word *cytryna* was subsequently also tested in the context-free translation experiments where only 38.23% of the online respondents translated it correctly.

P14/12 B: **Citrón** anebo **kyselina citronová**?

A: *Cytryna*... No, citrón dám, jo?

‘B: **Lemon** or **citric acid**?’

A: *Cytryna* ... Well, I’ll put lemon there, ok?’

P1/12: A: [...] *Cytryna*, citron.

B: Asi, ale možná by to byl **citron**, **teda jako citrón by byl polsky citron**, jestli to náhodou není, ta, **limetka**.

A: Tak jo, no. **Limetka**.

‘A: *Cytryna*, lemon.

B: Probably, but that might be **citron**, **like lemon would be citron in Polish**, so I wonder if this is not a ... **lime**.

A: Alright then, yeah. **Lime**.’

- **ekspozycja** ‘exposition’ – According to Treq, the most frequent translation into CS is *expoziice* ‘exposition’ (66.1%) with a corpus frequency of 66.46 i.p.m. After successfully recognising the PL-CS suffix correspondence *cja:ce*, pair 12 initially mentions *dispozice* ‘disposition’ (93.17 i.p.m.) as a possible translation before they consider *expoziice*, varying the prefix:

P12/11: B: Tak **expoziice** by mohla bejt **dispozice**, ne? [...]

A: Znalost polského jazyka je určitě dobře. Expoziice tovarův, ty kráso... **Expozice**, jako, že něco ukazuješ, takže, jakoby, že bys tam prováděl?

B: A proč bysi prováděl po obchodě, kterej budeš uklízet? [...] Čas, co časová dispozice? [...] Protože ten čas by se hodil tam i k těm předchozím továrnám možná trochu. [...] A říká se časová dispozice, není to nějak, jak se to říká česky?

‘B: So, [reading **ekspozycja**] could be **disposition**, or not? [...]

A: Knowledge of the Polish language is surely correct. Exposition of [reading *towarów*], man ... **Exposition**, like, you’re showing something, so, like, you’re guiding people there?

B: And why should you guide people through a shop that you’re going to clean? [...] Time, what about time availability? [...] Because time would fit there with these previous factories maybe a bit. [...] And do you say time availability, isn’t it somehow, how do you say that in Czech?’

- **kolegium** ‘council’ – According to Treq, the most frequent translation into CS is the identical *kolegium* (40.6%) which has a relatively low frequency of 3.55 i.p.m.:

P3/3: A: *Kolegium*... co je *kolegium*?

B: Jakože, pokud si dobře vzpomínám, jak to říkat, jak někdo říkal, možná už si... možná si to s něčím pletu, ale že *kolegium*, to je prostě množina kolegů. Jestli mi rozumíš.

‘A: [reading *kolegium*] ... What’s *kolegium*?’

B: Like, if I remember that correctly, how to say that, like someone said, maybe I’m already ... maybe I’m confusing it, but *kolegium*, that’s just like a number of colleagues. You know what I mean.’

P5/3: A: Tak *komise* nebo něco takového podobného? *Rada*?

‘A: So, it’s like a commission or something similar? A council?’

For a comparison, the mentioned alternatives have higher corpus frequencies: *komise* ‘commission’ (0.6% of all translations according to Treq) has a frequency of 73.13 i.p.m. and *rada* ‘council’ 154.79 i.p.m. The synonym suggested by pair 14, *sněm* ‘assembly, parliament’, has a corpus frequency of only 6.53 i.p.m, which nevertheless is higher than that of *kolegium*. Respondent 14A then realises that *kolegium* and *sněm* might be synonyms and expresses her favour to leave *kolegium* in the translation:

P14/3: B: *Kolegium* dalo mi pozvolenie, aby zrealizovač ten projekt nad jezzerom. Hm, mi dalo povolení, aby... zrealizovat projekt nad jezzerem. [...] *Kolegium*, třeba *kolegium*.

A: To je i česky. *Sněm*...

B: *Sněm*?

A: No já nevím, to už hledáme synonyma. Klidně napiš *kolegium*.

‘B: [reading sentence]. Hm, gave me the permission to ... realise the project at the lake. [...] *Kolegium*, maybe *kolegium*.

A: That’s also Czech. *Assembly* ...

B: *Assembly*?

A: Well, I don’t know, we’re already looking for synonyms. Just write down *kolegium*.’

P11/3: B: *Kolegium* bude *kolegium*, ne?

A: To je, no, *kolegium* asi. Povolení [...] Abych zrealizoval ten projekt nad jezzerem. I když slovo zrealizovat teda není úplně česky, ale...

B: Jó, to je.[...] Tak *kolegium* taky není úplně nejlepší.

A: No, tak, *kolegium* se používá i v češtině. Třeba *kolegium* děkana, ty jo, to jsou ty lidi, co mu raděj.

B: A není to jako, jako nějaký latinský [...] převzatý? Řekli bychom nějaký **shromáždění** nebo ...?

‘B: [reading *kolegium*] is *kolegium*, isn’t it?’

A: That’s right, yeah, *kolegium* probably. Permission [...] to realise the project over the lake. Although the word *zrealizovat* is also not really Czech, but ...

B: Yeah, it is [...] Well, *kolegium* isn’t the best either.

A: Well, yeah, *kolegium* is used in Czech, too. Like *kolegium* of the dean, man, that are people that give him advices.

B: And isn’t it like, like some Latin [...] loanword? If we said some **assembly** or ...?’

P9/3: A: **Společenství**, no... Ale hej, nevíš zas, co je *kolegium* v polštině, že jo?

‘A: **Community**, yeah ... But hey, you never know what *kolegium* is in Polish, right?’

P12/3: B: A přemejšlej ještě nad tím *kolegium* teda, to se mi zdá, zní divně. Není to nějakej **spolek**?

‘B: And think again about that *kolegium*, that seems to me like, sounds weird. Isn’t it some **association**?’

Again, for a comparison, the alternatives for *kolegium* that were considered by the respondents have the following corpus frequencies: *shromáždění* ‘assembly’ – 17.86 i.p.m., *společenství* ‘community’ – 25.56 i.p.m. and *spolek* ‘association’ – 32.49 i.p.m. The fact that all suggested synonyms or alternative translations for PL *kolegium* have higher corpus frequencies than CS *kolegium* might indicate that it is those very infrequent internationalisms in the readers’ L1 that cause distrust.

- ***konsumpcyjny*** ‘consumable’ – According to Treq, the most frequent translation into CS is *spotřebitelský* ‘consume [A]’ (37.1%) which has a corpus frequency of 6 i.p.m. The closest CS cognate translation *konzumní* ‘consumable’ is slightly less frequent with 3.41 i.p.m., but neither of the NPs *spotřebitelský led* nor *konzumní led* is found in the corpus. Pair 9 expresses its doubts about the adequacy of the translation of *konsumpcyjny*:

P9/12: A: No není, more, když budeš mít nějaký technický led na chlazení nějakých...

B: Okej, [...] tak **konzumní, konzumní**. To jsem ještě neslyšela, led **konzumní**.

A: Já taky ne, ale možná v Polsku to mají rozlišené. [...]

‘A: This isn’t like that, dude, if you have some technical ice for chilling of some ...

B: OK, [...] then **consumable, consumable**. I’ve never heard of that, **consumable** ice.

A: Me neither, but maybe they distinguish that in Poland. [...]

The results of the subsequent web-based context-free translation experiments confirm the problems respondents had with this word: Only 50.0% translated it correctly, whereas e.g. the internationalism *komunikacyjny* ‘communication [A]’, although sharing not only the feature of the suffix, was translated correctly by 79.9%.

- **oferta** ‘offer’ – According to Treq, the most frequent translation into CS is *nabídka* ‘offer’ (80.6%) which has a frequency of 145.77 i.p.m. The identical CS translation *oferta* is a rarely used term with a frequency of 0.01 i.p.m. and Treq provides only 7 hits where PL *oferta* is translated as CS *oferta* (> 0.1% of all translations of PL *oferta*). Pair 15 assumes that it means *odpověď* ‘answer’ (133.61 i.p.m.) or *poptávka* ‘request, demand’ (28.74 i.p.m.), although they entered the correct translation in the end:

P15/10: A: **Oferta** bude **odpověď**. Nebo to bude pop... jako že se ptá. **Pop-távka**, ne... Jakože, víš co, někdo píše do nějaký produkce masa a na něco se ptá a voni mu potom odpoví.

‘A: [reading *oferta*] is **answer**. Or it’s req... like asking. **Request**, or not? Like, you know, someone is writing to some meat producer, asking for something and they give him an answer.’

Respondents might therefore rather be able to infer the meaning of *oferta* from EN *offer*, although none of the respondent pairs mentions it explicitly. Four out of nine respondent pairs who saw the noun *oferta* in the original condition transformed it into the verb form *nabízíme* ‘we offer’ (inflected form: 5.73 i.p.m.; infinitive form *nabízet* ‘to offer’: 198.35 i.p.m.), such as pair 1:

P1/10: A: [...] Tak, oferta.

B: **Nabízíme**.

A: **Nabízíme** – no, no, no, přesně, to je ono.

‘A: [...] Ok, *oferta*.

B: **We offer**.

A: **We offer** – yeah, yeah, yeah, exactly, this is it.’

• **rektor** ‘rector’ – According to Treq, the most frequent translation (61.6%) into CS is the identical *rektor* ‘rector’ with a corpus frequency of 8.84 i.p.m.

The word *rektor* ‘rector’ aroused a similar situation of distrust in identical words in stimulus sentence 8. This sentence is expected to be lexically transparent to Czech readers, except for the difference in the preposition *do* ‘to’ which also exists in CS but carries the meaning ‘into’. Given that both prepositions express a direction, a correct understanding of *do* by the Czech readers can be expected.

On the other hand, respondents might question a certain internationalism because they consider it does not fit the context. Viewing the surprisal levels within the CS translation of the sentence, we observe the highest levels for the sentence onset *neviděla* ‘I didn’t see’ as well as for the end of the sentence *rektorovi* ‘rector [dat]’ (Jágrová, Avgustinova et al., 2019). While some respondents only raised the question to ensure the partner agrees with the assumption that PL *rektor* is CS *rektor*, such as:

P15/8: A: Jako *rektor* bude asi rektor, ne?

‘A: Like, *rektor* is probably rector, or not?’

other respondents doubted that PL *rektor* is CS *rektor*:

P5/8: A: To asi nebude rektor jako takovej.

‘A: That’s probably not a rector as such.’

P1/8: A: Poslat pro... **ten rektor je divnej ale.**

B: A nebude to třeba ředitel?

A: Jo, ale to je pravda. Měli bychom poslat pro...

B: **Ředitele...**

A: **Ředitele...**

‘A: Send for ... **but that rektor is weird.**

B: And isn’t that a **headmaster** maybe?

A: Yeah, that’s true. We should send for ...

B: **The headmaster.**

A: **The headmaster.’**

P16/8: A: Žebysmi posli do rektora. Že bysme šli, ale teďka co je *rektor*, že jo. [...]

A: Že to asi nebude jako **rektor na univerzitě**, podle mě.

B: To je nějaký jako, no... Jako **učitel**? Třeba teďka zase z... nevím, třeba jako **mentor** je učitel, tak rektor by taky mohl...

‘A: [reading *Żebyśmy poszli do rektora*]. That we should go, but now what is *rektor*, right?’

B: That’s probably not going to be a **university rector**, as for me.

A: That’s kind of like, well ... Like a **teacher**? Maybe now again like with ... I don’t know, maybe like a **mentor** is a teacher, a rector might also ...’

Pair 16 considers the more frequent translation *učitel* ‘teacher’ with 96.82 i.p.m., but also the less frequent *mentor* ‘mentor’ with 1.99 i.p.m. The response *učitel* was also among the translations given in the cloze translation design of the experiment conducted subsequently to this pairwise cooperative experiment (section 16).

The corpus frequency data of the responses suggest that, with the exception of *mentor* ‘mentor’ and *kyselina citronová* ‘citric acid’, respondents tend to dismiss the closest CS translations of seemingly understandable internationalisms in favour of more frequent words that are often wrong translations.

The reason why respondents considered *kyselina citronová* as a translation of PL *cytryna* might be due to the different grammatical gender of PL *cytryna* and its CS translation *citron*. As observed in later experiments, respondents tend to maintain the grammatical gender when translating target words in context (Jágrová & Avgustinova, 2019; see section 15.3) – the case of *kyselina citronová* confirms this finding.

A possible conclusion that can be drawn from this subsection is that even though internationalisms, foreign words or loanwords can be found in a corpus or a dictionary, it does not mean that they are part of a native speaker’s transfer base or even lexikon. In fact, a baseline for evaluating the reading or interpretation ability of the respondents in their own language could be addressed in future experiments. This should help identify potential biases and extreme cases – not only with regard to internationalisms, but also e.g. archaic Panslavic vocabulary.

When looking for linguistic predictors for the intelligibility of individual words, the procedure usually is to see if there is a cognate in the readers’ language and, if it existed, to calculate orthographic distance. Most probably, when dealing with internationalisms in context, attention has to be paid to both the frequency and the contextual factor – both outweigh linguistic distance as a predictor in this scenario, whereby frequency seems to have greater impact on intelligibility of certain internationalisms than the contextual factor.

5.6. Revision After Having Already Named the Correct Answer

Revision and discard of items for which the correct translation was already mentioned does not only happen with internationalisms or identical words as shown in section 5.5., but also with more distant cognates as well as non-cognates.

- *sztu*ka ‘piece’, in other contexts also ‘art’:

Pair 16 discarded the correct translation of *sztu*ka ‘piece’ – *kus* – in favour of *lži*čka ‘spoon’:

P16/12: B: Aha a zkratka [...] *sz*t.?

A: To je určitě jeden, jeden **kus**... jako určitě, ale jako nevím, co je to *sz*t.

B: Myslíš? Jeden... nebo on taky, že... no...

A: Jo, počkej. Jeden... jo počkej, jeden...

B: No, pak mě napadlo třeba lžička, víš? Ono se pak třeba jako table spoon, teda... teda v angličtině.

A: Jo, to je blbost, aby to byl jeden kus, to je fakt, no. [...] Tak jedna stol...

B: Tak jedna stolní lžička nebo kávová lžička.

A: Tak dáme jedna **lži**čka.

B: No.

A: A do závorky kávová, ne?

B: No, no.

‘B: Aha and the abbreviation [reading *sz*t.]?’

A: That’s surely one, one **piece** ... for sure, but I don’t really know, what *sz*t. is.

B: Do you think so? One ... or he’s also ... well ...

A: Yeah, wait. One ... yeah, wait, one ...

B: Well, then I had the idea of a **spoon**, you know? That might be something like table spoon, like ... I mean in English.

A: That’s nonsense, that one piece, that’s a fact, yeah. [...] So, one table ...

B: So, one table spoon or coffee spoon.

A: Let’s put one spoon there.

B: Yeah.

A: And coffee in brackets, right?’

Since *kus* and *sztuka* are non-cognates (except if the respondents had knowledge of DE *Stück* ‘piece’ or RU *штuka* ‘piece’), there is no option for comparing more similar words, but rather relying on context information.

Pair 15 discarded the correct translation *knížek* ‘books [gen pl]’ of the stimulus word *książek* in favour of *slov* ‘words [gen pl]’:

P15/1: A: Kdyby nebylo **knížek** [...]

Kdyby nebylo *ksizek* [...]. Kdyby nebylo *ksi... ksiazek*, hm. [...] No, **knížky, to asi nebude ono**. *Ks... žek*. [...] Kdyby nebylo čeho? [...] Kdyby nebylo... ty jo, co to znamená? [...] *Ksiazek...* hm. Kdyby nebylo... no, ty jo. Četl by z očí. [...] To fakt nevím. Hm, tak třeba to bude, vid'. [...] No, tak přijdeme na to, co je *ksizek*? Asi ne, no.

‘A: If there were no **books** [...]

If there were no [reading *książek*] ... If there were no [reading *książek*], hm. [...] Well, **books, that’s probably not it**. [reading *książek*]. If there were no what? [...] If there were no ... man, what does that mean? [... reading *książek*] ... hm. If there were no ... well, man. He would read from the eyes. [...] I really don’t know. Hm, maybe that’s it, right. [...] Well, are we going to find out what is [reading *książek*]? Probably not, hm.’

The examples demonstrate that discarding of correct translations can happen due to contextual reasons. However, the discarding of intermediate wrong translations is much more frequent in the protocols than the discarding of correct translations.

5.7. Handling Unfamiliar PL Orthography

This section attempts to systematise the phenomena observed when Czech respondents encounter PL orthography with special attention to unknown diacritics and the PL digraphs. The categories in the subsections of 5.7.1. should not be interpreted as mutually exclusive. Respondents are not consistent in how they encounter words with unknown diacritics or unusual character sequences and it is mostly the case that respondents vary in their strategies, i.e. ignoring diacritics when pronouncing a word once and another time pronouncing it differently. The recordings reveal that the most problematic letters to pronounce were *ł*, the nasal vowel letters *ę* and *ą* and the digraphs/diphthongs *rz*, *sz*, and *cz* and combinations of these.

Since the task was first to try and read the stimulus aloud, the respondents produced utterances of what is likely to be the manifestation of their inner speech when reading PL. The Czech respondents are not expected to know

the correct pronunciation of the stimuli, since they have not learnt PL before. Some of the mistakes in pronunciation interestingly reveal the causes for wrong translations.

5.7.1. Handling PL diacritics

The PL letters *a*, *e*, *l* and *ž* have diacritics that do not occur in the CS alphabet. Also, there are no acutes (*čárky*) with consonant letters as basic glyphs in CS, but in PL there are *ć*, *ń*, *ś* and *ź*. CS consonants are palatalised by a háček *ˇ*, respectively by a similar sign if the letter does orthotactically not allow for this sign as in the case of *l'* and *d'*.

Pair 2 referred to the PL letter *l* as “měkký L” ‘soft L’, most probably because the function of the stroke was assumed to be a sign of palatalisation, similar to the CS háček.

P2/4: B: Praga to vazni komunikacjny vezel. Vezejl... To **měkký L** ne-umim říct.

‘B: [reading sentence 4] ... I cannot pronounce this **soft L**.’

Another explanation could be that the respondents knew the “soft L” through their exposure to SK (Nábělková, 2007). In fact, SK has two different L-characters with diacritics: the syllabic *ḷ* (*dlhé el* ‘long L’) and *l'* (*mäkké el* ‘soft L’) and none of them is identical to the PL letter *l*. Pair 15 referred to the unknown character as “crossed-out L”, assuming to know “these signs” from the Slovak alphabet:

P15: A: Tahle latinka... Akorát maj nějaký **takový ty znaky** jako **přeškrtnutý L** a ty maj i Slováci. [...]

‘A: This Latin script ... They just got some of **these signs** like the crossed-out L and the Slovaks got that, too. [...]

Pair 8 seems to be aware of the existence of nasals in PL, but wrongly interprets the letter *l* as a nasal:

P8/6: B: Tež starý **žoltý** samochód.

A: *Žon-žontý*, myslím. Ne, *žontý*.

B: *Žo-žo-žontý, žontý*, okej.

A: To... nebo počkat, ne. *Žontý*? Nevím. Nevím teďka.

B: Já... ts... neplet' sem francouzštinu.

‘B: [reading *tež starý žóltý samochód*].

A: [reading *žóltý* with a **nasal**], I think. No, [reading *žóltý* with a **nasal**].

B: [reading *żółty*, pronouncing it after the partner], OK.

A: That ... or wait, no. [reading *żółty* with a **nasal**]? I don't know. I don't know now.

B: I ... ts ... don't mingle this with French here.'

This might be due to the tilde symbol ~ that represents nasal vowels and nasalised consonants in the International Phonetic Association (IPA, 1999) standards, and students might have encountered this symbol during foreign language learning and might therefore associate it with nasalisation.

Pair 6 refers to the ogonek in *q* as *krucánek* which can be translated as 'twisted little thing', knowing that it is pronounced as a nasal, but pronouncing it as /an/ while the actual pronunciation in this position is [ɔŋ]:

P6/9: A: A myśliś, że to **skjand** bude výška?

B: Ne, to jsem jenom tak plácla. **Skad** to zní... **skad** je kdy.

A: Takovej jako **an** by měl být, jakože, jako **an**.

B: Jak se čte to, ten **krucánek**, prosim tě?

A: Jakože jako **an**. *Skand* asi. *Skand*.

B: *Skand*.

'A: And you think that this [reading **skqd**] could be height?

B: No, I was just guessing. **Skad** that sounds like ... **skad** is when.

A: Something like /**an**/ that's what it should be, like, like /**an**/.

B: How do you read that, this **twisted little thing**, please?

A: Like an /**an**/. *Skand* probably. *Skand*.

B: *Skand*.'

The closest sound representing the sound repertoire of the CS language would be something that could have been transcribed as *skond*. All occurrences of the letter *q* in the sentence stimuli, their CS translations and the function of the correspondence are provided in Table 18. The regular PL-CS correspondences that would be applicable if correctly recognised are given in the right column:

PL part of stimulus	CS translation	Function	Correspondence
<i>skąd</i> 'from where'	<i>odkud</i>	stem correspondence	<i>q:u</i>
<i>ręka</i> 'hand [instr]'	<i>rukou</i>	instr	<i>q:ou</i>
<i>rosną</i> 'they grow'	<i>rostou</i>	third person pl	
<i>będą</i> 'they will'	<i>budou</i>		
<i>zagrożających</i> 'harming'	<i>ohrožujících</i>	present participle	<i>q:i</i>
<i>interesujących</i> 'interesting [gen pl]'	<i>interesujících</i> ¹²		
<i>księżek</i> 'books [gen]'	<i>knížek</i>	stem correspondence	<i>iq:i</i>
<i>piątku</i> 'Friday [loc]'	<i>pátku</i>		
<i>pięćdziesiąt</i> 'fifty'	<i>padesát</i>		
<i>obowiązków</i> 'duties [gen]'	only stem correspondence <i>wiqz:váz/vaz</i>		
<i>porządkowe</i> 'order [A pl]'	<i>pořádkové</i> [literal], <i>úklidové</i>	stem correspondence	<i>q:á</i>

Table 18: Words containing *q*, CS cognate translations and applicable correspondences.

As shown in Table 18, the PL letter *q* has various orthographic correspondences with different functions in CS. In the stimuli, it can correspond to the vowel letters *u*, *ou*, *i* or *á* in CS. For instance, the *q:ou* correspondence applies to third person plural endings (also *q:i* in other verb forms apart from those in the stimuli) and feminine instrumental endings. The *q:i* correspondence applies to suffixes in present participle forms. While these morphological correspondences are regular, the stem correspondences, in which *q* often also occurs in the digraph *iq*, are not. Based on the previous finding that diacritics are often times ignored or moved to another possible position by respondents, the *q:á* rule is not expected to pose any bigger problems, whereas the other rules might be problematic. Accordingly, words containing the PL character *ę*, their CS translations and the applicable regular PL-CS correspondences are listed in Table 19:

PL part of stimulus	CS translation	Correspondence
<i>język</i>	<i>jazyk</i>	<i>ę:a</i>
<i>jarzębiny</i>	<i>jeřabiny</i>	
<i>ręka</i>	<i>rukou</i>	<i>ę:u</i>
<i>będzie</i>	<i>bude</i>	
<i>będą</i>	<i>budou</i>	
<i>godzinę</i>	<i>hodinu</i> (accu ending)	<i>wę:u</i>
<i>węzeł</i>	<i>uzel</i>	
<i>mięsa</i>	<i>masa</i>	<i>ię:a</i>
<i>pięćdziesiąt</i>	<i>padesát</i>	
<i>mięta</i>	<i>máta</i>	<i>ię:á</i>
<i>się</i>	<i>se</i>	<i>ię:e</i>
<i>miesięczne</i>	<i>měsíční</i>	<i>ię:i</i>

Table 19: Words containing *ę*, CS cognate translations and applicable correspondences.

- 12 This form of *interesující* 'interesting' cannot be found in the CNC, but it appears in 140 google search results (as of 06 December 2018). The more frequent form would be *zajímavých* 'interesting [gen]' and *interesujících* is chosen to demonstrate the PL-CS correspondence.

In contrast to the PL-CS correspondences with *q*, those with *ę* mainly occur in the stems, except for *ę:u* which is a correspondence in the feminine accusative (accu) noun endings. Another source of difficulty while trying to recognise the correspondences with *ę* is the great variety of possible CS correspondences, also in a digraphic combination as *ię*.

Pair 5 is aware of both nasals *ę* and *q* in *ręka* ‘hand [instr]’, which enables them to correctly identify the morphological correspondence *q:ou* in the instrumental forms of *ręka/rukou* ‘hand’:

P5/8: A: **Renkou** – rukou třeba?

B: Že, jakože nevěděl, že jeho žena *pokazuje renkou*.

‘A: [reading *ręka*] – with the hand maybe?’

B: Right, like, didn’t know that his wife [reading *pokazuje ręka*].’

Among the cognates containing *ę*, there was *języka* ‘language [gen]’ which apparently did not cause any complications. Only pair 16 pronounced it /*jezika*/ when reading the stimulus aloud, while all other respondent pairs directly transferred it to its CS cognate *jazyka*:

P16/11: A: [...] **jezyka** polskjego, to znamená znalost polského jazyka.

‘A: [...] /*jezika* polskjego/, that means knowledge of the Polish language.’

The same pair and also pair 2 successfully applied the *q:i* correspondence in the present participle suffix:

P16/7: B: [...] **Interesujacich**, tak to bude zajímavých praktik, ne? **Inte... interesujacich. Interesujících** bude zajímavé, zajímavých praktik, zajímavých zvyků v kraji?

‘B: [...] [reading], well, that will be interesting practice, or not? *Inte... [reading]. [reading with a CS suffix and ending]* will be interesting, interesting practices [gen], interesting habits in the region?’

P2/7: A: **Interasujících**, to budou zájmových, zajímavých.

‘A: [reading *interesujících*] that is going to be interest [A], interesting [gen pl].’

Table 20 gives a comparative overview of the actual pronunciations of the letters *ę* and *q* by the respondents. The characters are bold in words which contain both characters to clarify which of them is analysed in the row. During the analysis of the respondents’ pronunciation, it was attempted to distinguish between the actual reading of the stimuli and the subsequent translation process.

PL	As if without diacritics	%	Nasal (or alike)	%	As if diacritics on other letter	%	Palatalization or jotation	%
skąd	/skant/	73.68	/skant/	21.05	n/a	0	/skjant/	5.26
ręka	/reka/	85.19	/renkou/	7.41	/reka/	7.41	n/a	0
ręka	/rekou/	10.00	/renkou/	10.00	/reka/	75.00	/rjeka/	5.00
rosną	/rosna/	97.37	n/a	0	/rosna/	2.63	/rosna/	2.63
będą	/bjeda/	75.00	n/a	0	/benja/	8.33	/bjendza/	16.67
będą	/beda/	16.67	/benja/, /bjendza/	8.33	n/a	0	/bjeda/, /bjendza/	75.00
zagrożających	/zagrazajatsix/	60.00	/zagrazajoutsr:x/ /interesujantsix/, /interasujotsix/	40.00	n/a	0	n/a	0
interesujących	/interesujatsix/	60.00		40.00	n/a	0	n/a	0
książek	/kjażek/	66.67	n/a	0	/k(t)a3-/	33.33	n/a	0
piątku	/piatku/	100.00	n/a	0	n/a	0	n/a	0
pięćdziesiąt	/pre'tj-, piedzjessants/	100.00	n/a	0	n/a	0	n/a	0
pięćdziesiąt	/precdzessat/	20.00	/pje'tj[ɔ]3jants/ /pretsdzjessat/	20.00	/pje'd3reJanti/, /pied'3eJat/, /pretsdzjessat/	60.00	n/a	0
obowiązków	/obovjaskof/	100.00	n/a	0	n/a	0		0
węzeł	/vezel/	91.70	n/a	0	n/a	0	/vezej/	8.33
mięty / mięta	/mjet-/	73.30	/mzent-/	26.70	n/a	0	n/a	0
jarzębiny	/jarebny/	75.00	/jarebny/	25.00	n/a	0	n/a	0.09
porządkowe	/pora:d-, /porad/	90.91	n/a	0	n/a	0	/porad-/	9.09
Mean		70.32		11.68		10.98		7.18

Table 20: Words containing q and ę and various pronunciations by respondents.

According to the analysis in Table 20, the most frequent way how respondents handled the unknown diacritics in the letters *q* and *ę* was to ignore the diacritics (about 70% of all read-out stimuli). The letter *q* was mostly pronounced as a regular short /a/, corresponding to the CS letter *a*. Pronouncing the stimuli as if the diacritic was on another letter and nasalisation of the vowels was nearly equally frequent (about 11% each). Respondents palatalised these stimuli in about 7% of all cases.

In the following sub-section, examples for the observations made when respondents encountered (unknown) diacritics are listed:

5.7.1.1. Respondents pronounce letters correctly

While most respondents pronounced *ł* as /l/, a few respondents seemed to be aware of the pronunciation of the *ł*:

P5/8: A: Mhm. *Vidźielam, vidzielam, vidźiauam.*

B: Hm. Tak *ně* bude určitě zápor.

A: Mhm, mhm, to tam není...

B: *Viděu, vidžau, vidžaua.*

A: Hm, nevi- nevidět.

B: Mm, *něvidžaua.*

A: Aha.

B: Nevidí nebo neví. No...

A: Aha.

B: Tam bych dal minulý čas, *něvidžaua!*

‘A: Mhm. [reading *widziałam* three times]

B: Hm. So, *nie* is certainly a negation.

A: Mhm, mhm, it’s not there ...

B: [reading *widziałam* three times]

A: Hm, don’t, don’t see.

B: Mm, [reading *widziałam*].

A: Aha.

B: Doesn’t see or doesn’t know. Well ...

A: Aha.

B: I would put past tense there, [reading *widziałam*].’

Respondent 3A also pronounced *l* correctly as /w/ in *pełna* ‘full’:

P3/10: A: **Peu- peuna** *dispozycyjnośc* od pondělka do pátku, *oferta realne možliwosci avansu* ve firmě. Dvanáct bruto *godzině* plus premie *mjesječně*.

‘A: [reading *pełna dyspozycyjność*] from Monday to Friday, [reading the rest of the sentence].’

Pair 8 correctly recognised the nasal *ę* in *węzeł* ‘knot’:

P8/4: B: *Praga... Praga to ważny*, jak se to čte, co?

A: *Venzel*.

B: *Venzel*?

A: Mhm, e s tím je en.

‘B: Prague ... Prague is an important, how do you read that, huh?’

A: [reading *węzeł* with a nasal].

B: [repeating what partner said]?

A: Mhm, an e with this is an en.’

The rather easy to recognise correspondences are such where a PL diacritic such as the dot on top of *ż* can be simply replaced by the corresponding CS háček. This is usually a correct strategy with cognates containing the regular correspondence *ż:ž*, for instance in the cognate pairs *książek – knížek* ‘books’, *możliwość – možnost* ‘possibility’, *żona – žena* ‘wife’, *że – že* ‘that’, and *zagrożający – ohrožující* ‘threatening’. In order to be able to formulate a valid account of how respondents encounter the letter *ż* in the stimuli, a quantitative overview of the read-out utterances is provided in Table 21 as follows:

Word containing <i>ż</i>	Pronounced as			
	<i>/ʒ/</i> , corresponding to <i>ż</i>	%	<i>/z/</i> , corresponding to <i>z</i>	%
<i>już</i>	<i>/juʒ/</i>	100.00	-	0
<i>książek</i>	<i>/ʒ/</i> in various surroundings	81.25	<i>/ksjzæk/</i>	18.75
<i>możliwość / możliwości</i>	<i>/ʒ/</i> in various surroundings	61.90	<i>/mozlivo-/</i>	38.10
<i>też</i>	<i>/te:ʒ/</i> or <i>/teʒ/</i>	86.36	<i>/tez/</i>	13.63
<i>używano</i>	<i>/ʒ/</i> in various surroundings	100.00	-	0
<i>ważny</i>	<i>/ʒ/</i> in various surroundings	72.73	<i>/vazn-/</i>	27.27
<i>zagrożający</i>	<i>/zagraʒ-/</i>	80.00	<i>/zagraz/</i>	20.00
<i>że</i>	<i>/ʒe/</i>	100.00	-	0
<i>żółty</i>	<i>/ʒ/</i> in various surroundings	82.00	<i>/z/</i> in various surroundings	18.00
<i>żona</i>	<i>/ʒona/</i>	55.56	<i>/zona/</i>	44.44
Mean		81.98		18.02

Table 21: Words containing *ż* and the various pronunciations by respondents.

In almost 82% of all cases, the letter *ż* was pronounced /z/ similar to the CS letter *ž* and only in the remaining 18% of all cases as /z/ which would correspond to the letter *z* without any diacritical sign. This suggests that respondents prefer to interpret and “replace” the diacritic on the *ż* by a háček as in the familiar letter *ž* rather than omitting or ignoring this diacritic. Exceptions were (i) the words *już* and *że* that were pronounced with a /z/ in 100% of all cases and (ii) the word *żona* ‘wife’ that offers two orthographic neighbours in CS – *žena* ‘wife’ and *zona* ‘zone’. Pair 15 discussed whether *ż* in *żona* could be the CS *ž* and thus means *žena*, despite their previous wrong pronunciation of the word as *zona* ‘zone’:

P15/8: A: *Že jeho... no jako to zona bude podle mě určitě žena. Ještě jak tam máš tu, tu tečku nad tím.*

‘A: That his ... well this *zona* in my opinion is certainly **wife**. Also, if you have that, that dot on top.’

This suggests that readers’ behaviour towards unknown diacritics changes with neighbourhood density. When there is an option to omit a diacritic and obtain an existing word, the distribution of cases in which it is pronounced as /z/ or /z/ changes enormously.

Accordingly, Table 22 provides an overview of the read-out instances for words containing the letter *ś*. On the contrary to the cases with words containing the letter *ż*, the correct strategy here would be to ignore the diacritic in most of the cases. The regular PL-CS correspondences applicable here are *ś:s* in the auxiliary verbs (attached to the verb forms in PL) such as in *kupiliśmy* – *koupili jsme* ‘we bought’, in the feminine suffix correspondences *ość:ost* (for singular forms) such as in *znajomość* – *znalost* ‘knowledge’ or *ości:osti* (for plural forms) such as in *możliwości* – *možnosti* ‘possibilities’, and in the stems *świad:svěd* and *śr:stř*. There is only one case where the correspondence *ś:š* should be applied: the Common CS translation of PL *jesteś* ‘you are’ would be *jseš*, which is *jsi* in standard CS.

Word containing ś	Pronounced as			
	/ʃ/, corresponding to ś	%	/s/, corresponding to s	%
<i>chcielibyście</i> ‘would you want’	/-bʃtʃe/	33.33	/-bɪscje/, /-bɪstʃe/	66.66
<i>czynności</i> ‘activities’	/tʃnoʃtʃi/	36.36	/tʃɪnosci/	63.64
<i>doświadczenia</i> ‘experience’	-	0	/dosv-/	100.00
<i>dyspozycyjność</i> ‘disposal’	/dɪspozɪtsɪjnɔʃtʃ/	50.00	/dɪspozɪtsɪjnɔs-/	50.00
<i>gotowość</i> ‘readiness’	-	0	/gotovosts/	100.00
<i>jesteś</i> ‘you are’	/jestɛj/	100.00	-	0
<i>kupiliśmy</i> ‘we bought’	-	0	/kupɪlɪsm-/	100.00
<i>możliwości</i> ‘possibilities’	/-lɪvɔʃ-/	47.62	/-lɪvos-/	52.38
<i>środowisku</i> ‘environment’	/ʃrodovɪsku/	56.25	/srodovɪsku/	43.75
<i>znajomość</i> ‘knowledge’	/znajomɔʃtʃ/	33.33	/znajomɔts/, /znajomɔst/	66.66
<i>żebyśmy</i> ‘that we should’	/ʒebɪʃmɪ/	25.00	/ʒebɪsm-/	75.00
Mean		34.72		65.28

Table 22: Words containing ś and the various pronunciations by respondents.

In about 65% of all instances, ś was pronounced as the CS letter š. One of the results stands out: the ś in the verb *jesteś* ‘you are’ was pronounced as /ʃ/ in all of the instances. This is likely due to the successful recognition of the Common CS translation equivalent *jseš*. The reason why some instances of ś were pronounced as /s/ and not /ʃ/ might also be that the sequences /ʃv/ and /ʃm/ are not as frequent as /sv/ and /sm/ in CS.

5.7.1.2. Respondents ignore diacritics and pronounce stimulus as if without diacritics

Examples:

- *ręka* ‘hand [instr]’ → /reka/

P1/8: B: ...*že jeho žena pokazuje reka*... to nevím, jak se ani čte tyhle ty písma v tom *reka*.

‘B: [reading *že jeho žena pokazuje ręka*] ... I don’t even know how to read these letters in this [reading *ręka* without diacritics].’

P4/8: B: Myslíš, *že reka* je *ruka*? [...] No a to slovo *reka* se teda vykašlem.

A: Přemejšlím, co s tím. A asi bych to...

B: Nějak moc nápady nemám.

‘B: Do you think [reading *ręka*] is **hand**? [...] Well and that word [reading *ręka*], we will skip that.

A: I’m thinking about what to do with it. Maybe I’d ...

B: Somehow I don’t have any ideas.’

The form *reka* exists in CS as a genitive/accusative of *rek* ‘hero’.

- *będq* ‘they will’ → /beda/

P6/9: B: Tědka, co je *beda*?

‘B: Now, what is [reading *będq*]?’

- *rosna* ‘they grow’ → /rosna/

P3/7: B: No dobře. Teď *rosna*...

A: Ta *rosna*, to, to fakt nevím. [...]

B: Ale, jo, jakým slovem bys nahradil v češtině to slovo *rosna*? Dobře, dobře. Teď *rosna*, rovněž, **možnosti**...

A: A provádění... ale hlavně **ta rosna, ta rosna** je důležitý.

‘B: Well, ok. Now /**rosna**/ ...

A: That /**rosna**/, that, I really don’t know that. [...]

B: But, yeah, with which word would you replace the word /**rosna**/ in Czech? Good, good. Now /**rosna**/, also, possibilities ...

A: And conducting ... but particularly **that rosna, that rosna** is important.’

Despite pronouncing *rosna* as if without the diacritic, pair 2 manages to correctly disambiguate this word in the end:

P2/7: B: To *rosna* může být třeba **roste**, že jo. Teď roste rovněž...

A: Ha, je já už vím! Že něco, ja...

B: Jo, to roste.

‘B: That *rosna* could be for example **grows**, right. Now it is also growing ...

A: Ha, I got it! Like something, li ...

B: Yes, it grows.’

5.7.1.3. Respondents move diacritics to another suitable letter in the word

Often, a switch of diacritics from one letter to another can be observed. Respondents pronounced some of the words as if the diacritics would be on other letters, for instance *ręka* ‘hand [instr]’ was pronounced as *řeka* ‘river’, apparently by moving the diacritic from *ę* to *ř*, because CS orthography does not allow for *rě* as a string of letters.

Examples:

- *ręka* ‘hand [instr]’ → *řeka* ‘river’

Pair 14 rejects the possibility that *ręka* could correspond to *řeka*:

P14/8: A: [...] No, já myslím, že **to nebude ř se řekou.**

‘A: [...] Well, I think **this is not going to be a ř as in řeka.**’

This process also occurred in cases in which the pronounced word resulted in an actually non-existing word, such as

- *rosną* ‘they grow’ → /rosɲa/

The diacritic of the nasal vowel letter *ą* of the verb form *rosną* was moved on top of the preceding *n* and turned into a sound that would correspond the CS letter *ň*:

P2/7: A: [...] Co je to **rosňa** potom?

‘A: [...] What is this /rosɲa/ there then?’

- *będa* ‘they will’ → /benɲa/

Likewise, respondent 3A pronounced the combination of the preceding *d* and the diacritic of *ą* in *będa* ‘they will’ as a palatalised /ɲa/ which would be represented by *d’a* in CS orthography. Interestingly, the same respondents also pronounced the *l* in *latali* ‘they flew’ palatal even though there is no diacritic in the word:

P3/9: A: Pjedżešiat lat ludzie nje **bend’a** już **ljatali** latadlem.

‘A: [reading *pięćdziesiąt lat ludzie nie będą już latali latadlem*¹³].’

5.7.2. Handling unfamiliar PL digraphs

Beside the difficulties with the pronunciation of letters with different diacritics, the most problematic situations can be observed with the digraphs *cz*, *sz*, *rz* and consonant strings composed of these. The correct PL-CS orthographic correspondences to be applied here are *cz*:č, *sz*:š, and *rz*:ř. The recordings and transcripts reveal that the respondents tended to wrongly divide syllables in the words containing these digraphs and sometimes therefore failed to recognise cognates.

13 Lexically modified stimulus (lex condition – see section 10)

PL stimulus		Wrong syllabic division	Correct CS cognate	EN
<i>oczu</i>	→	oc-zu /otsu/	<i>očí</i>	'eyes [gen]'
<i>brzozy</i>		br-zo-zy /brzoza/	<i>břízy</i>	'birch [gen]'
<i>suszona</i>		sus-zo-na /suzona/	<i>sušená</i>	'dried'

Table 23: Recognition of cognates might fail due to wrong division of syllables.

Besides the results of the cooperative translation experiments, there is evidence for this type of mistake in other experiments, too. In the cloze translation task with highly predictable target words in context (Chapter VI, section 15), for instance, the target word *poczta* 'post [instr]' was wrongly translated as *pocta* 'honour' or *poctou* 'honour [instr]' by 39% of the respondents (52% responded correctly with a form of *pošta*). When the word *poczta* was presented to the Czech respondents without any context, 70% responded *pocta* or *úcta* 'esteem' and only 24% *pošta*.

The digraph *rz* was either pronounced as /r/, which would be in line with the regular correspondence rule *rz:ř*, or in a way that two syllables were created. Pair 3 demonstrates this with the word *gorzej* 'worse' which they pronounced as /gor-zej/. However, regardless of the wrong pronunciation, they manage to find the correct translation:

- P3/6: A: Víš co, ale to, ale ještě **gorze**.
 B: Dobře.
 A: Ale ještě **gorze** zní jakože **hůře** [...]
 'A: You know what, that, but even /**gorze**/.
 B: Good.
 A: But even /**gorze**/ sounds like **worse** [...].'

Pair 8 discussed the regular correspondences *rz:ř*, *g:h*, and *ž:ž*:

- P8/6: A: **Gořčej** podle mě.
 B: Ne, to je, **to je ř. Rž je ř, ale g se nečte jako ř, ne? Ne, g se čte jako h?**
 A: No, **gořej** prostě. [...]
 B: **Gořej** to je starý...
 A: **Tež, to je ž, myslím.** [...]
 'A: /gortʃej/ as for me.
 B: No, that's, **that's ř. Rž is ř, but g isn't read as ř, is it? No, g is read like h?**
 A: Yeah, simply /gořej/
 B: /gořej/ that's old.
 A: /tež/, **that's a ž, I think.** [...].'

Pair 15 explicitly mentions the orthographic correspondence *cz:č*:

P15/1: B: Četl bych *ci z ocu*.

A: Četl by mi z očí.

B: Kdyby nebylo... no *ocu* budou oči. **Cz je č.**

‘B: I would read /tsi s otsu/.

A: He would read from my eyes.

B: If there were no ... well /otsu/ could be eyes. **Cz is č.**’

Some respondents explicitly raise the question how characters might be pronounced in PL. Respondents were, for instance, unsure about the pronunciation of

- the digraph *sz* in the abbreviation *szt.* for *sztuka* ‘piece’, frequently corresponding to *š* in CS:

P11/12: B: No a citrón jeden *st.* Set?

A: Ne.

B: **Když to přečteš, tak je to s t**, že jo.

‘B: Well and lemon one [reading *szt.*]. Set?’

A: No.

B: **If you read that, then it’s /s - t/**, right.’

Pair 15 discussed whether *poszli* ‘went [pl]’ is pronounced with an /s/ or /ʃ/, choosing between the CS neighbours *pošli* ‘send [imperative]’ and *posly* ‘messengers’:

P12/8: B: No, potom to **pošli**.

A: Možná poslat?

B: Jo, **poš- poš- pošli**. [...] Jo, pošli. **A co kdyby to byli posly** třeba? [...] Že by pošlala... že by poslali posly k rektorovi.

‘B: Well, then this **pošli**.

A: Maybe send?

B: Yeah, **poš- poš- pošli**. [...] Yeah, send. **And what if it is messengers** maybe? [...] That she sent ... that they sent messengers to the rector.’

- the digraph *rz* in *gorzej* ‘worse’, frequently corresponding to *ř* in CS:

P3/6: B: Ještě **goře, gořa**... co to... **goře**, hm. **Goře**... no, tak jako **jak bys to jinak četl?**

A: **Goře**, to je zajímavý nápad, jak to přečíst.

‘B: Also [reading *gorzej*] ... what is ... [reading *gorzej*], hm. [reading *gorzej*] ... well, so how else would you read that?’

A: [repeating /goře/], that’s an interesting idea how to read it.’

The CS cognate translation *hůře* ‘worse’ is orthographically relatively distant (LD = 75%) and might only serve as a transfer base if the regular *rz:ř* correspondence is actually recognised. Although respondents successfully applied the rule *rz:ř* in *porządkowe:pořádkové* (100% of all read out instances) in sentence 11, in 75% of all read out instances of the stimulus word *brzozy* ‘birch [gen]’ respondents made a syllabic division between *r* and *z* pronouncing it /br-zo-za/, which led to wrong translation results.

In the following, the frequencies of how the respondents pronounced the digraphs *cz*, *rz*, and *sz* are given in Table 24-Table 26.

Original word containing <i>cz</i>	Pronounced /tʃ/, similar to <i>č</i>	%	Pronounced as subsequent consonants /ts/ + /z/	%
<i>czynności</i> ‘activities’	/tʃinoʃtʃi/, /tʃinosci/	100.00	-	0
<i>doświadczenia</i> ‘experience’	/dosvjettʃeni:/	82.35	/ts/ in different surroundings	17.65
<i>jeszcze</i> ‘also’	/jeʃtʃe/	92.31	/jestse/	7.69
<i>oczu</i> ‘eyes [gen]’	/otʃu/	47.83	/otsu/	52.17
Mean		80.62		19.38

Table 24: Words containing the digraph *cz* and the various pronunciations by respondents.

Original word containing <i>rz</i>	Pronounced /r/, similar to <i>ř</i>	%	Pronounced as two subsequent consonants /r/ + /z/	%
<i>brzozy</i> ‘birch [gen]’	/brɔ-/	25.00	/brzo-/	75.00
<i>gorzej</i> ‘worse’	/gorej/, /gortʃej/	40.00	/gorzej/	60.00
<i>jarzębiny</i> ‘rowanberries’	/jar-/	83.33	/jarzebɪnɪ/	16.67
<i>porządkowe</i> ‘cleanup [A]’	/por-/	100.00	-	0
<i>przekonana</i> ‘convinced’	/prekonana:/	100.00	-	0
<i>przy</i> ‘in, with’	-	0	/przi/	100.00
Mean		58.06		41.95

Table 25: Words containing the digraph *rz* and the various pronunciations by respondents.

Original word containing <i>sz</i>	Pronounced /ʃ/, similar to <i>š</i>	%	Pronounced as /s/ and/or /z/	%
<i>jeszcze</i> ‘also’	/jeʃtʃe/	90.00	/jestse/	10.00
<i>poszli</i> ‘we went’	/poʃli/	95.65	/posli/	4.35
<i>suszona</i> ‘dried’	/suf-/	83.87	/suz-/ /sus-/	16.12
<i>szt.</i> abbreviation for <i>sztuka</i> ‘piece’	/ʃtik/, /ʃtuk/	61.54	/səzətə/	38.46
Mean		82.77		17.23

Table 26: Words containing the digraph *sz* and the various pronunciations by respondents.

On the average, a correct pronunciation of the digraphs *cz* and *sz* seems to prevail with an 80/20 distribution, whereas there is only a slight preference for the pronunciation of *rz* as /r/ (58%). In about 81% of all read-out instances of *cz* and about 83% of all read-out instances of *sz*, the orthographic correspondences seem to have been correctly recognised, although the shares vary considerably between the few individual examples. The *cz* in *czynności* ‘activities’

was pronounced correctly as /tʃ/ in 100% of the cases, which suggests that the *cz:č* correspondence at word onset might be easier to recognise than at another position in the word, probably also because respondents would not divide syllables at word onset. The preference for the pronunciation of *rz* varies considerably from one stimulus to another. While *rz* was pronounced /r/ in all read-out instances of *przekonana* ‘convinced’ and *porządkowe* ‘cleanup [A, pl]’, it was pronounced /rz/ in all instances of *przy* ‘in, with’. An explanation for why the respondents did not recognise *przy* as what could be *při* ‘at’ in CS might be that *przy* is an orthographic neighbour of the CS adverb *brzy* ‘early’ which is pronounced /brzy/ with a syllabic division between *r* and *z*. In contrast to this, *przekonana* and *porządkowe* are long words with no orthographic neighbours and therefore the possible CS transfer bases *překonaná* and *pořádkové* with the respective correspondences (although not being the correct translations) might have been easy to recognise.

5.8. Talking About Grammar

In some dialogues it could be observed that respondents discussed topics of grammar – examples of this will be presented in the following. Since the respondents were non-linguists, their assumptions cannot be expected to be correct. Nevertheless, it is interesting to observe the topics and the grammatical difficulties they identified during the task.

Pair 2 noticed that sentence 4 is lacking a verb, discussing that every sentence and the translation should contain a finite verb form. They correctly identified that PL *to* ‘this’ can be translated with the CS adverb *tot’* ‘this is’, wondering whether *tot’* is a verb or not:

P2/4: A: [...] Já mám pocit, **že v tý větě chybí jakýkoliv sloveso**. Takže teoreticky by v tom překladu by taky nemělo bejt sloveso. [...] Takže něco jako ta Praha, ten významný uzel, jako Praha, ten významný komunikační uzel, třeba.

B: **Ale jak ty můžeš vědět, že to není sloveso polsky? A navíc začíná to velkým písmenem a končí tečkou. A každá věta snad v každým jazyce musí mít...**

A: No, tak to může bejt větnej... ale... to může bejt větnenej ekvivalent.

B: Ježiš, Maria, hele s tím na mě nechod’, prostě to je věta, to musí bejt věta. [...] Praha, to významný. Praha, toť významný komunikační uzel by šlo, že jo. [...] **A není tot’ taky sloveso?**

A: Není. Ale... **nevím, co je tot’**. Ale hodí se to tam nejvíc.

‘A: [...] I have the feeling that **this sentence is lacking any kind of verb**. So, theoretically there should be no verb in the translation either. [...] So, something like Prague, that important hub, like Prague, that important traffic hub, for example.

B: But how can you know that to is not a verb in Polish? And also, it starts with a capital letter and ends with a full stop. Probably every sentence in every language has to have ...

A: Well, then it can be a sentential ... but ... that could be a sentence equivalent.

B: Gosh, don’t try that on me, that’s simply a sentence, it has to be a sentence. [...] Prague, this important. Prague, this is an important traffic hub would be good, right. [...] **And isn’t *to*’ also a verb?**

A: It is not. But ... I don’t know what *to*’ is. But it fits there best.’

Due to the frequent ignoring of diacritics, the ending *-q* was often mistaken for a typical feminine ending and the majority of respondents mistook stimuli words with this feature for a feminine noun (see also section 15.4.3.3. on (perceived) morphological mismatches). This again had influence on other words in the sentence: Pair 16 discussed the possibility that *odbycia* ‘spending [gen]’ in the sequence *rosną również możliwości odbycia* ‘the possibilities of spending [...] are growing’ of sentence 7 might be a verb which would be congruent to the word *rosną* ‘they grow’ which again was mistaken for a feminine noun:

P16/7: A: ...ale *odbicija*, teďka je klíčové jako *odbicija*, **že to je sloveso**.

‘A: But [reading *odbycia*], now this is a keyword, this [reading *odbycia*], **that this is a verb.**’

Nevertheless, mistaking *odbycia* for a verb form might also be an interference from SK, since there is a third person plural verb ending *-ia* which does not exist in CS, for instance in the SK phrase *ludia robia* ‘people do’. Although not immediately, in the end pair 7 managed to disambiguate *rosną* correctly.

Some respondents were aware of the possibility that word order in PL might differ from the CS word order. Pair 3 points out that the NA linearisation in *miód kwiatowy* ‘blossom honey’ “sounds weird” and affirm that it must be “the other way around” in PL, obviously being aware of the post-modification of nouns by adjectives, which also existed in older varieties of CS. They formulate an alternative of the phrase where *med* ‘honey’ is postmodified by an equivalent prepositional phrase – *z květu* ‘from a blossom’:

P3/12: A: A nemají to Poláci třeba naopak? [...]

B: Cože? **Historicky se to takhle v češtině používalo**, ale máš pravdu, spíš se používá třeba **sušená zelená máta**.

A: ...**nemají to obráceně**, že jako to pořadí těch... **med květový**, zní to divně.

B: **Sušená zelená máta, květový med**, no. [...] Med z květu, řekněme. [...]

‘A: And don’t the Polish have that on the opposite? [...]

B: What? **It was used like this in historical Czech**, but you’re right, we’re rather using *sušená zelená máta*.

A: ... **don’t they have it the other way round**, like the order of these ... **med květový**, that sounds weird.

B: **Dried green mint, blossom honey**, yeah. [...] Honey from the blossom, so to say. [...]

5.9. Problems Caused by Differences in Government Patterns

Difficulties in intercomprehension that arise from differences in government patterns were, among other topics, thematised in a study by Muikku-Werner (2014) who investigated the intercomprehension of Estonian by Finnish students. She points out that despite similarity, “even a familiar lexical item can cause translation problems” in cases where the Lx and the language in the reader’s repertoire differ in rection. She defines rection as “the determination of the form of one word by the presence of another word in a phrase or a sentence” (p. 104) and refers to the term of colligation – the co-occurrence of words with particular grammatical categories. Difficulties with different rection and phenomena that could fall under the category of colligation occurred, for instance, in sentence 2:

W 2000 roku wzrósł do ponad 900 mln. marek obrót towarami, w procesie produkcji których nie używano substancji zagrażających środowisku naturalnemu wilka.

‘In the year 2000, the turnover of goods in the production of which no substances that are harmful for the natural habitat of the wolf are used, rose above 900 million German mark.’

Here, the lexical item that can be expected to be familiar is *towar* or the inflected form *towarami* ‘goods [instr]’ which is a cognate of CS *tovar* ‘commodity [nom sg]’ or *tovarů* [gen pl] and could also have been translated with the more frequent *zboží* ‘goods’. The preceding word *obrot* ‘turnover’ demands the instrumental case, whereas the CS cognate *obrat* collocates with its complements in the genitive case. It is remarkable how often the respondents therefore decided for the translation *továren* ‘factories [gen]’, using the orthographically closer

lexical item *továrnami* ‘factories [instr]’ as a transfer base that differs in only one character and a diacritic from the PL stimulus *towarami* ‘goods [instr]’.

5.10. Problems Caused by Different Prepositions

Not only differences in rection, but also the use of different prepositions to express the same meaning in two languages can confuse readers of the related language. This is even more tricky when the preposition identically also exists in the reader’s L1. Among the stimuli, two such cases that have caused difficulties were especially prominent:

- PL *nad jeziorem* vs. CS *u jezera* ‘at the lake’ in sentence 3:

The expression ‘at the lake’ would be *nad jeziorem* with the local case in PL and *u jezera* with the genitive case in CS. The more similar *nad jezerem* also exists in the CS local case, but means ‘over/above the lake’, which the PL phrase could mean, too. Therefore, both CS translation variants were considered correct. Nevertheless, some respondents managed to identify the different grammar and provide the more likely CS translation. Pair 6 does this despite mistaking PL *jezioro* ‘lake’ for *jez* ‘wier’:

P6/3: B: Ten projekt, který jako **zastřešuje** ten jez, takže je jako **nad**...

A: No, právě, si myslím, jestli **to třeba vůbec neznamená, že by to jako vůbec nebylo jako nad ve smyslu výškově**, ale že by to bylo projekt s je... **s jezem nebo projekt na jezu** nebo víš něco takovýho, že, že by to prostě [...] **něco jako je třeba do rektora, k rektorovi**...

‘B: That project which like **covers** that weir, so it’s like **above** ...

A: Well, exactly, I think, if that might not be that **this isn’t even like over in the sense of height**, but that it would be a project with a ... **with a lake or a project at a weir** or, you know, something like that, that, that it would simply [...] **something like for example do rektora, k rektorovi**...

They even synchronise this with another phrase with a divergent preposition in the stimulus set (*do rektora* – see below), and by doing so, they provide a proof for a learning effect in the competence of tolerating divergent prepositions in NPs.

- PL *poszli do rektora* vs. CS *šli k rektorovi* ‘(we) went to the rector’ in sentence 8

This case of morphosyntactic priming seems to create difficulties, because the verb and preposition *šli do* ‘went into’ together with a complement in the genitive case creates the semantic expectation of *entering a building or an institution* in CS, and not *meeting a person*, as it does in PL. The correct CS equivalents would be either *šli za rektorem* in the instrumental case or *šli k rektorovi* in the local case. This might be a reason why the actually identical lexeme *rektor* ‘rector’ was misinterpreted by the respondents frequently (see also section 5.5). Pair 8 even mentioned the difference in the preposition:

P8/8: B: [...] No, jo, ale jako, že, že **rektorát** myslí se, možná... Jako **instituce** prostě **nebo ředitelství**. No, **rektorát** je oficiální slovo. Nebo aby šel už. Počkej, počkej... ne. Mam šanci... Neviděl jsem [...] A co je rektorát? To je něco? [...] Abychom šli **na rektorát** teda...

A: To zní... to je, to je divné pros... Počkej, počkej ještě. [...]

B: Tak jako, **co je tam za předložku?** Takže, rektorát...

A: **Za rektorem možná, jo, za rektorem.**

B: Že za rektorem by mohlo bejt. To zní, to mi zní dobře, to mi zní hodně dobře. Neviděla jsem jeho ženu ukazovat, abychom šli za rektorem.

‘B: [...] Well, yeah, but like, that, that **rectorate** is meant, maybe ... just like an **institution** or a **head office**. Well, **rektorát** is an official word. Or that he went. Wait, wait ... no. I have the chance ... I haven’t seen [...] And what is a rectorate? Is that something? [...] So, that we went **to the rectorate**.

A: That sounds ... that’s, that’s just weird ... Wait, wait a bit [...]

B: Alright, **what preposition is there?** So, rectorate [...]

A: **To the rector maybe, yeah, to the rector.**

B: That could be to the rector. That sounds, that sounds good to me, that sounds very good to me. I haven’t seen his wife showing that we should go to the rector.’

P6/8: A: [...] Že by **k rektorovi**? M-mm, to bude podle mě k rektorovi.

B: *Do rektora* a nemůže to fakt být něco jinýho?

A: Ale jak chceš, já myslím, že to bude k rektorovi, ale zas nechci na tom nějak trvat.

‘A: [...] Could it be **to the rector**? M-mm, that is to the rector I think.

B: [reading *do rektora*] and could that really be something different?

A: As you wish, I think that this is to the rector, but I don’t want to insist on it.’

Some mistakes cannot be classified as being of a certain type. In some cases, interferences can be a mix of wrong pronunciation, morphological differences, divergences in word order or the source of the misinterpretations cannot be clearly identified. The following discussion of pair 1 is a mix of many mistaken pronunciation rules, (wrong) associations and wrongly interpreted keywords:

P1/8: B: No, potom *to pošli*.

A: Možná poslat?

B: Jo, poš- poš- pošli. [...] A co kdyby to byli **posly** třeba?

A: Že bysme mohli po...

B: Že by *poslala*... že by **poslali posly** k rektorovi.

A: To je divný, já myslím, že to bude, že, že... možná. A já bych řekla, že tam je, jakoby, nevidím nebo něco takovýho.

B: ...**zóna**...

A: Že tato ne asi zóna... možná jo. Že tato zo... [...]. Zóna... Že tato zóna...

B: Není to něco ve smyslu, jako, že to... že... že oni chtějí jít někam k němu, k nějakýmu rektorovi a že ta **řeka** tam očividně nevede?

A: Možná. Nevidím, že toto... ta zóna je divná, že tato oblast. [...] Anebo jakože nebo jakože tam překračuje řeka nebo že tam **pokračuje řeka**.

B: Jo, jo, jo, jo, no a že ta řeka neteče k tomu rektorovi.

A: Nevidí... já bych řekla, že jakoby nevidím, že...

B: No, tato řeka...

A: Nebo nemyslím si, že tudy, tudy teče ře- řeka a mohli bysme poslat **pro** rektora.

B: Jo, to zní hodně dobře.

A: Takže nemyslím si, že tudy **poteče** [...]

A: Poteče ře... [...]. No, jakoby, měli bysme poslat pro rektora nebo něco takovýho. [...] Měli bysme, ne, měli bychom, co? [...]

A: Poslat pro... ten rektor je divnej ale.

B: A nebude to třeba **ředitel**?

A: Jo, ale to je pravda. Měli bychom poslat pro... [...] Ředitele... [...] Nemyslím si, no je to docela takový divný, ale... možná, že i jo.

‘B: And then that [reading *poszli*].

A: Maybe to send?

B: Yeah, [reading as imperative of *poslat* ‘to send’] [...] And what if these are **messengers** maybe?

A: That we could see ...

B: That she [reading] ... that they would **send messengers** to the rector.

A: That’s weird, I think that this is, that, that ... maybe. I would say that there is I don’t see or something like that.

B: ... **zone** ...

A: That this probably not zone ... maybe yes. That this zo... [...]. Zone ... That this zone.

B: Something in the sense that, like, that ... that ... that they want to go somewhere to him, to some rector and that this **river** obviously doesn’t lead there?

A: Maybe. I don’t see that this ... this zone is weird, that this area. [...] Or like or like there it is crossing a river or a **river continues** there.

B: Yeah, yeah, yeah, yeah, and that this river doesn’t flow to the rector.

A: She doesn’t see ... I’d say like I don’t see that ...

B: Yeah, this river ...

A: Or I don’t think that here, here a ri- river and we could send **for** the rector.

B: Yeah, that sounds very good.

A: So, I don’t think that a river **will flow** here. [...] Well, like, we should send for the rector or something like that. [...] We should, no, we should, right? [...]

A: Send for ... but that rector is weird.

B: And isn’t that a **headmaster** maybe?

A: Yeah, that’s true. We should send for ... [...] The headmaster. [...] I don’t think so, well that is quite a bit weird, but ... maybe even yes.’

Pair 1 tried to pronounce *poszli* ‘[we] went’ in different ways. First, the correspondence *sz:š* is rejected in favour of *sz:s* which led to weighing if it was a form of the verb *poslat* ‘send’ or the noun *posly* ‘messengers’.

The correspondence *ż:ż* in *żona* ‘wife’ was not recognised and instead, the diacritic was ignored and *ż:z* was applied, which led to a wrong interpretation of *żona* as *zona* ‘zone’. Then, the diacritic of the *ę* in *ręka* ‘hand [instr]’ was moved and that of *ą* was ignored so that *ręka* was interpreted as *řeka* ‘river’. The preposition *do* ‘to’ in *do rektora* ‘to the rector’ was consequently changed into *pro* ‘for’ while trying to meaningfully connect the already translated words. At last, the actually identical *rektor* ‘rector’ was dismissed in favour of the more frequent *ředitel* ‘headmaster’.

5.11. Summary

This section intends to qualitatively evaluate the transcripts of the audio recordings of an intercomprehension experiment in which Czech readers were supposed to translate different PL sentences cooperatively into CS. The analysis was conducted along different categories. It revealed the strategies respondents used, the sources of transfer and over-transfer, the reasons for distrust in already understood items and the handling of unfamiliar orthography.

Respondents used two basic techniques when they encountered difficult to understand language material: i) Leaving unknown words open and using placeholder words for them, mostly in the correct POS, and subsequently trying to infer their meaning from the context. ii) Repeated reading of critical words aloud with varying ways of pronunciation. The most difficult items in the stimuli could be identified not only when respondents applied these techniques, but also by the order they were translated – the most difficult parts were disambiguated last, if at all.

It could be shown that readers use not only their L1 as a transfer base, but also dialects of their L1 as well as other Ln. However, it is not guaranteed that respondents are always able to find an L1 or Ln transfer base for comprehension, even though it is available. Also, evidence was presented that respondents are aware of the meaning of certain non-cognates and false friends, although they had never learnt PL, and use this awareness successfully in this task. On the other hand, it was shown that even words that have identical translation equivalents in CS can be discarded in favour of more frequent translation variants. This is particularly true for internationalisms with infrequent CS cognate translations. In some cases, already correctly identified words were revised and substituted for wrong translations where respondents found that the more similar translations do not fit the context.

Regarding the unfamiliar PL orthography, the most problematic features for the Czech respondents proved to be the digraphs *cz*, *sz*, and *rz* as well as letters with diacritics that do not exist in CS. One of the common mistakes reflected

in pronunciations was the syllabic division of the digraphs. In about 80% of all read-out instances of *cz* and *sz*, the orthographic correspondences seem to have been correctly recognised by the way they were pronounced, while no clear preference for the pronunciation of *rz* could be found. The recognition of these digraph correspondences seems to be easier at word onset. Also, the recognition of digraph correspondences highly depends on the number of available translation options with only a minimal difference (neighbourhood density) – the more neighbours the unknown word has in the reader's L_n , the less likely it is that the word is translated correctly.

Respondents were not consistent in the way they pronounced digraphs or words containing unfamiliar diacritics. In some cases, respondents seemed to be aware of the pronunciation of the PL diacritics, since they pronounced them in line with regular PL-CS correspondences, i.e. correctly in the broadest sense. The most problematic PL letters with diacritics proved to be *q*, *ę*, and *ś*. The failing recognition of the applicable correspondences with these letters often led to wrong comprehension and even to wrong assignment of POS. The latter was mainly the case for the correspondence *q:ou* applicable in feminine instrumental endings and in third person plural verb endings. The pronunciation of the read-out stimuli revealed that often times diacritics were ignored or moved to another suitable base letter in the word, sometimes in order to pronounce the word as an existing CS word. This again highly depended on the neighbourhood density of the stimuli items. Words ending in *q* were therefore frequently mistaken for feminine nouns. The letter *ż* was only problematic when CS neighbours with a *z* at the position of the *ż* exist.

Other sources of mistakes could be identified in differences in government patterns and the different use of prepositions, although often a single source of mistakes could not be determined, since several factors, also less obvious and sometimes respondent-specific associations, interplay when respondents tried to formulate meaningful translations out of bits and pieces of the stimuli they understood.

CHAPTER III: ON-LINE EXPERIMENTS

6. Hypotheses

6.1. Pronunciation-Based Orthographic Distance

Similarity in orthography does not always coincide with phonetic similarity. In some language combinations, cross-lingual similarities might be better preserved in their written forms, which for instance applies for the case of Danish and Swedish (Gooskens & Swarte, 2017), whose spoken forms have diverged further apart than their orthographies. This, however, does not apply to the pair PL-CS, where on the contrary orthography has developed further apart than the actual pronunciation of many words.

As Vanhove points out, transfer from a known Ln to an Lx might not only depend on the objective distance, but is rather a matter of how the reader *perceives* the distance (Vanhove, 2014, p. 5). This is in accordance with Ringbom's (2007, p. 11) distinction between *objective* (symmetrical) and *perceived* (not necessarily symmetrical) cross-linguistic similarities. This might concern individual attitudes towards encountering other languages in general as well as exposure to a particular language. When encountering PL for the first time, one might be overwhelmed by the many consonants representing sibilants that a reader's eye might not be accommodated to. Consider, for instance, the word *pięćdziesiąt* 'fifty' – this was one of the words in stimulus sentence 9 in the cooperative translation experiment (see CHAPTER II). Although it has a relatively large objective distance to its CS translation equivalent *padesát*, it did not pose any comprehension problem for the respondents. After some exposure to PL, this effect of overwhelming might change and the reader can segment the code into individual syllables or morphemes to make understanding possible.

The objective linguistic distance might be measurable with standardised methods between language pairs. However, considering that readers might try to pronounce what they read, successful recognition of cognates highly depends on how they *assume* that words or characters in the Lx are pronounced. Hence, it is desirable to design a metric of linguistic distance which takes into account the respective human decoding process. In the case of this thesis, this metric can be developed with the insights from the cooperative translation experiments in CHAPTER II. Figure 14 visualises the idea of the pronunciation-based Levenshtein distance (pron LD) as a distinct measure between orthographic (trad LD) and phonetic distance.

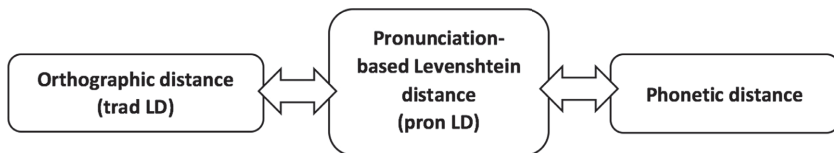


Figure 14: Trad LD vs. pron LD vs. phonetic distance.

Most readers will try to pronounce the unfamiliar language material silently – a phenomenon referred to as “inner speech” (cf. Harley, 2007) – or aloud. There is a number of PL-CS correspondences that can be treated as obviously transparent – those that were evidently pronounced “correctly” or as their respective CS counterparts in accordance with applicalbe correspondences. Consequently, it would be more appropriate not to charge any substitution costs for these characters in the Levenshtein alignment.

According to the insights from the cooperative translation experiments, the PL-CS character and digraph correspondences in Table 27 were assigned a cost of 0 (in addition to the alignment of identical characters):

PL	a	e	er	g	i	ia	ie	ie	ja	ja	ł	ń	o	ó	u	w	y	y	y	ż
CS	á	é	r	k	í	ie	ě	e	ie	e	l	ň	ó	o	ú	v	ý	i	i	ž

Table 27: Additional PL-CS alignments that cost 0 for pron LD.

Some of these alignments are simply cases in which the two languages use different characters to represent the same sound, such as *w* and *v* in PL *woda* and CS *voda* ‘water’ – these characters would not appear in the other L, except in named entities of foreign origin. Some of the correspondences are (nearly) identical sounds in only some words, e.g. *g* and *k* in PL *gdzie* and CS *kde* ‘where’ (the *k* in *kde* regressively assimilates to [g], the PL *g* is palatalised as [ǰ] – therefore not entirely identical to [g]), while these are different in, e.g., PL *gitara* and CS *kytara* ‘guitar’ or PL *gabinet* and CS *kabinet* ‘cabinet’ as /ǰ/ vs. /k/. Even if they are different as in the latter two examples, I assume the difference irrelevant for intelligibility. Other correspondences, particularly long vowels that are written with a *čárka* in CS but without a diacritic in PL are e.g. *y:y* in *dym* and *dým* ‘smoke’. Here, the pronunciation aspect (short vs. long) does not play such a big role, but rather the fact that Czech readers are used to read text without diacritics, e.g. in chats, text messages and *suchlike*. Therefore, such correspondences will probably not pose a problem, because the diacritics can simply be ignored.

Vanhove also showed in intercomprehension experiments with Swiss multilinguals that a combined distance, when calculated towards the closest DE or EN cognates (“Germanic distance”), was a better predictor than the respective monolingual distances (cf. method in Vanhove, 2014 on “Germanic distance”, p. 139; Vanhove & Berthele, 2015, p. 112). This suggests that multilingual readers rely on more than only their L1 when they try to understand words in a related Lx (Vanhove & Berthele, 2015, p. 21). Czech respondents are regularly exposed to SK and therefore can be expected to have receptive skills in SK. Hence, I hypothesise that the same principle might apply when measuring the distance of PL towards a Czechoslovak (CSK) distance, unifying the closest CS or SK variant in the calculation.

The correspondences listed in Table 27 can be ascribed to the Czech readers’ regular exposure to SK (Nábělková, 2007) and its differences in relation to CS orthography. Czechs apply these correspondences more or less unconsciously in CS-SK intercomprehension. Thus, they are likely to tolerate noisy code with differences at the same position when reading PL. Möller & Zeevaert (2015) observed this principle in intercomprehension experiments with GER cognates presented to German native speakers. Examples for such PL:(SK:) CS correspondences are *ie:ě* or *ie:e* as in SK/PL *nie* and CS *ne* ‘no’, *ja:(ia:)ie* in PL *policja* and CS *policie* ‘police’, *ia:ie* in PL *akademia* and CS *akademie* ‘academy’, *ja:(ia:)e* in PL *informacja* and CS *informace* ‘information’, and *ie:i* SK/PL *papier* and CS *papír* ‘paper’ as shown in Table 28.

	<i>ie:e</i>	<i>ja:ie</i>	<i>ia:ie</i>	<i>ja:e</i>	<i>ie:i</i>
PL	<i>nie</i>	<i>policja</i>	<i>akademia</i>	<i>informacja</i>	<i>papier</i>
SK	<i>nie</i>	<i>policia</i>	<i>akadémia</i>	<i>informácia</i>	<i>papier</i>
CS	<i>ne</i>	<i>policie</i>	<i>akademie</i>	<i>informace</i>	<i>papír</i>

Table 28: Correspondences that Czech readers are likely to handle through exposure to SK.

Table 29 demonstrates the difference between the calculation of trad LD (to the left) and pron LD (to the right). Consequently, the alignment of the cognates *człowiek* and *člověk* ‘human’ would result in the following calculation:

# slots	1	2	3	4	5	6	7	8	Trad LD	1	2	3	4	5	6	7	8	Pron LD
PL stimulus	<i>c</i>	<i>z</i>	<i>ł</i>	<i>o</i>	<i>w</i>	<i>i</i>	<i>e</i>	<i>k</i>	4.5/8 = 56.25%	<i>c</i>	<i>z</i>	<i>ř</i>	<i>o</i>	<i>w</i>	<i>i</i>	<i>e</i>	<i>k</i>	1.5/8 = 18.75%
Closest CS	<i>č</i>		<i>l</i>	<i>o</i>	<i>v</i>		<i>ě</i>	<i>k</i>		<i>č</i>		<i>l</i>	<i>o</i>	<i>v</i>		<i>ě</i>	<i>k</i>	
Costs	0.5	1	0.5	0	1	1	0.5	0		0.5	1	0	0	0	0	0	0	

Table 29: Calculation of trad LD of a cognate pair in comparison to pron LD.

In the pronunciation-based calculation in Table 29 (to the right), no substitution cost is charged for the alignment of *l:l*, *w:v*, and *ie:ě*. In the traditional way for calculating the LD for this word pair, a cost of 0.5 for *l:l*, 1 for *w:v* and 1.5 in total for *ie:ě* would have been charged. Consequently, the hypothesis is that the CSK pron LD will correlate better and explain more of the variance in the data than the traditionally calculated orthographic distance trad LD.

6.2. Surprisal as a Predictor Variable for Context in Intercomprehension

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 trigram LMs). In the current multilingual setup, target words that have low linguistic distance to the reader's L1 and are predictable in context are expected to be understood 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 better correlate with LD and surprisal than only with LD.

Of course, the amount of correctly perceived sentential context plays a crucial role in such an intercomprehension task, too. If the context is 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 recognise even non-cognates and maybe even false friends in sentences. However, the effects of semantic priming, which might make some of the target words predictable, are not expected to be predictable by the trigram 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 context?**
2. **If so, do surprisal values obtained from trigram LMs correlate with the intelligibility scores of the target words?**

7. Empirical Base

In order to test the hypotheses and answer the questions formulated in section 6, translation experiments with different kinds of stimuli representing the different linguistic levels were conducted in the framework of this thesis. The experiments and results build upon one another successively in order to make a systematic analysis of the sentence stimuli as the core part of the thesis possible. Only after looking at the role of orthography and morphology separately can the complex phenomena taking effect in sentence material be examined.

7.1. Online Experiments

The online experiments (CHAPTERS III-VI) as well as the cooperative translation experiments (CHAPTER II) were conducted on the experiment website <http://intercomprehension.coli.uni-saarland.de> developed in the INCOMSLAV project. The website interface was translated into 11 Slavic languages (Belarusian (BEL), BG, CS, HR, MK, PL, RU, SR, SK, SL, and UK) as well as into EN and DE, targeting respondents who are native speakers of at least one of these Ls. Not only experiments that are subject to this thesis were conducted on the website, but also experiments in other stimulus languages with respondents from other language backgrounds relevant to INCOMSLAV project were tested. As of 15th February 2019, 1559 respondents have already taken part in at least one of the experiments available on the website.

I refer to the experiments discussed in this thesis in the past tense, even though some of the experiments in different language combinations are ongoing and might be subject to future investigations. All experimental stimuli for the tested language-reader combinations had to be uploaded as .xlsx files to the website's admin panel which is not visible for the public.

Before the actual experiment, the informants clicked to agree on the informed consent form and then created an account on the website with their own user login. After they had entered their standard sociodemographic information (see Figure A 1 in the appendix), their L1(s), Ln(s), and exposure to languages, they were asked for a self-assessment of skills for all languages they had indicated. The self-assessment scale was designed as a drag-and-drop bar with a continuous 7-point scale, ranging from 0 to C2, oriented on the Common European Framework of Reference for Languages (CEFR). For each indicated language, the skills for speaking, hearing, reading, and writing were inquired separately.

Having completed the self-assessment, respondents were automatically assigned one of the experiments in a foreign language, depending on their language background (L1) and the priority of the experiments that was entered in the admin panel. Respondents were not tested in a language that they had indicated in the language background questionnaire¹⁴. They were asked to confirm to have understood the task and to set their keyboard to CS.

Different time limits were set for the different kinds of experiments. The allocated time was meant to be sufficient for typing even the longest words, but not long enough for using a dictionary or online translation tools. When clicking on the Next button on the screen or pressing Enter on the keyboard, the next stimulus was displayed. All stimuli were displayed in random order. The system automatically switches to the next stimulus after the time limit has expired, regardless of whether a respondent has entered anything into the solution field or not. The expected correct answers were entered into the system beforehand, so that the respondents' results were automatically categorised as correct or wrong via pattern matching and the respondents could receive immediate feedback in form of emoticons. Some stimuli had more possible correct translations than was thought of beforehand and therefore all answers analysed in this thesis were checked manually for correctness and for typographical errors. If participants had entered a correct solution that was not fed to the website beforehand, it was subsequently counted as correct. The system tolerated missing diacritics and made no distinction between uppercase and lowercase letters and it saved anything that was entered by an informant, regardless of whether an informant confirmed the translation by pressing enter (or clicking *pokračovat* 'proceed') or not. The emoticon was displayed at the left bottom of the page (see Figure 16) – a thumbs up for a correct translation or a sad face for a wrong or missing translation.

In the following, some distinctive features of the three kinds of web-based experiments discussed in this thesis will be explained.

14 This is not true for Ukrainian and Belarusian respondents who all know RU and/or live in an area where RU is spoken. The RU-UK and RU-BEL combinations are not discussed in this thesis, but in the thesis of Irina Stenger who investigates written intelligibility in the Slavic languages with Cyrillic script.

- **Free translation of individual words**

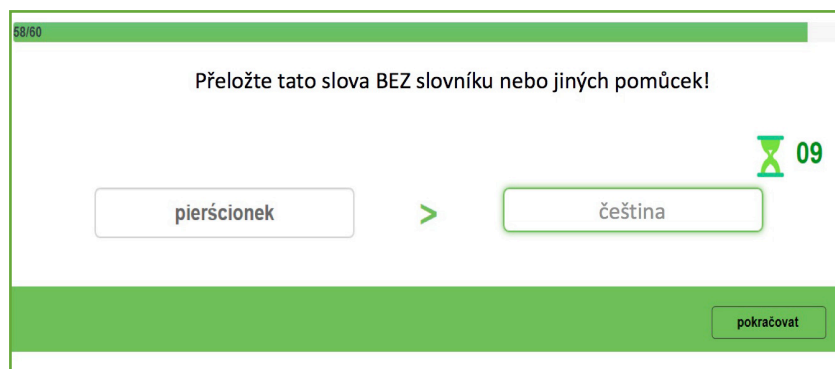


Figure 15: Experimental screen in the free translation experiments.

Figure 15 presents a screenshot of the free translation experiment with individual words. The correct translation of the stimulus *pierścionek* ‘ring’ would be *prstýnek* in CS. The instruction on top says: ‘Translate these words without a dictionary or other aids!’ Respondents had exactly 10 seconds time to enter their CS translation. The time limit for the free translation task was adapted from the limit in similar translation experiments within the Micrela experiment (van Heuven et al., 2015) conducted at the University of Groningen. During the experiment, a window with the message *Time for a break* with a 3-second countdown timer appeared after a certain number of stimuli, depending on the overall number of stimuli per block. In a block of 50 individual word stimuli, for instance, the break appeared after the 10th, 30th, and 40th stimulus.

- **Translation of NPs**

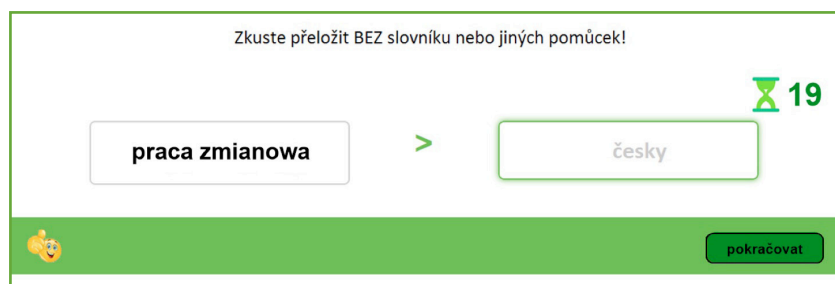


Figure 16: Experimental screen in the NP translation experiments.

Figure 16 is a screenshot of the NP translation experiments as seen by Czech respondents. The time limit in these experiments was 20 seconds – twice the time limit in the free translation experiments with individual words. The correct translation of the stimulus NP *praca zmianowa* ‘shift work’ in Figure 16 would be *směnná práce* or *práce na směny* in CS.

- **Cloze translation experiments**

The cloze translation experiments were the most complex type of experiments. Participants were introduced to the experimental task by a short video demonstration. 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. This method ensured that participants read each sentence word by word. Only after they had clicked on the last word in the sentence, the cloze gap (uniform length of 100 pixel) with the target word for translation was displayed. The target word was displayed on top of the frame, the assumed translation was entered inside the frame. Figure 17 shows a screenshot after a respondent clicked through the whole sentence and entered the response *prstýnek* ‘ring’ as a translation of the PL target word *pierścionek* ‘ring’ into the gap. The instruction on top says: ‘When you click on the last word, a marked word will appear. Then translate this marked word.’ There were 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.

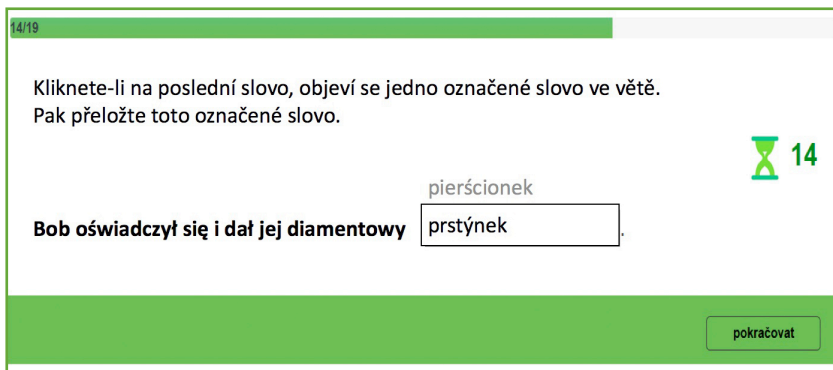
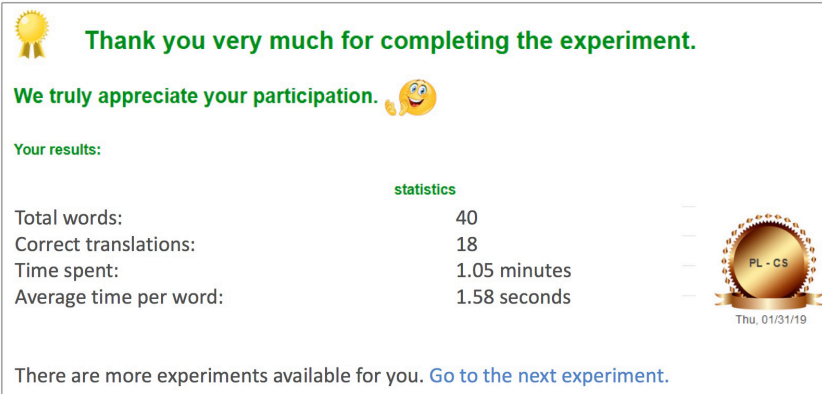




Figure 17: Experimental screen in the cloze translation experiments.

In all experiments, the stimuli were presented automatically in random order. The random order of stimuli was supposed to counterbalance uncontrolled variables, such as learning effects or a loss of concentration after a large number of stimuli. However, it was hardly possible to fully exclude learning effects that may arise when partaking in several experimental blocks. The response to the first stimulus was likely to be somewhat slower than the subsequent responses, because the respondents took more time to become acquainted with the layout and with the experimental design. Nevertheless, given the fact that the stimuli appeared in random order for each participant, an equal distribution of stimuli is expected so that all responses were treated equally, regardless of their within-participant order.

Initial hesitation time (before typing), time spent typing, submission hesitation time (time between the last keystroke and pressing the *enter* or clicking the *next* button) and total time spent on the stimulus was recorded for each translation. For practical reasons, only the total time spent on the stimulus (henceforth referred to as *processing time*) is evaluated with regard to NPs in section 14 in this thesis.




 **Thank you very much for completing the experiment.**

We truly appreciate your participation. 

Your results:

	statistics	
Total words:	40	—
Correct translations:	18	—
Time spent:	1.05 minutes	—
Average time per word:	1.58 seconds	—



There are more experiments available for you. [Go to the next experiment.](#)

Figure 18: Brief statistics shown to respondents after a completed experiment.

At the end of each experimental block, participants saw their results on a brief statistics page, displaying the number of correct translations, total time and average time per stimulus (Figure 18). The respondent in Figure 18 (probably a Czech native speaker) gained a bronze medal for the PL to CS translation experiment. The language of the website could be selected by the respondents (the respondent in Figure 18 chose EN).

The respondents could participate in another experiment by clicking *Go to the next experiment* underneath the statistics. Then they could choose another experiment from a list. If available, the next experiment could be in the same language combination, but with a different block of stimuli. It was not possible to do the same experiment more than once¹⁵.

7.2. Overview of Experiments and Data Collected

Table 30 provides an overview of all experiments discussed in this thesis, the number of stimuli, experimental conditions, blocks (parts) of the stimuli and the time limit for each stimulus.

Experiment	Discussed in section	n stimuli	n conditions	n blocks	Time limit to enter response
Cooperative translation experiments	4 + 5	12	12	1	5 mins
Free translation of individual words					
• With PL-CS correspondences	12	353	1	6	10 s
• 100 most frequent Ns	13	57	1	1	
• From cooperative translation experiments	16.5	103	1	2	
• Highly predictable target words from cloze translation	15.4	118	1	2	
Free translation of PL NPs					
• Czech readers	14	37	2	6	20 s
• German readers	14.7	42	2	2	
Cloze translation in PL sentences					
• Sentences with highly predictable target words at sentence final position	15	149	1	7	depending on number of words/sentence and number of gaps/sentence
• Sentences from cooperative translation experiments	16	12	1	1	
• Sentences containing false friends	16	10	1	1	
Baseline cloze experiments (monolingual)	16.2	30	up to 4	1	no

Table 30: Overview of experiments conducted, sorted by topic and section in this thesis.

The column *n conditions* displays the number of experimental conditions in which the stimuli were presented. For instance, in the free translation of NPs experiment, the NPs were presented in two different conditions – AN vs. NA.

15 A try again feature was added to the experiment website later.

7.3. Participants

Experiment	n stimuli	n participants	n female	n male	n other ¹⁶	Mean age
Free translation of individual words	353					
Block 1	45	21	9	12	0	26.05
Block 2	40	5	4	1	0	26.20
Block 3	53	35	14	20	1	25.74
Block 4	53	8	4	4	0	26.75
Block 5	53	5	3	2	0	25.80
Block 6	53	35	12	22	1	24.71
Block 7	53	32	15	16	1	22.13
Block 8	51	34	14	20	0	24.71
Block 9 (TOP 100)	56	30	9	21	0	25.83
Block 10	60	33	7	26	0	26.82
Block 11	60	30	6	24	0	24.50
Free translation of NPs						
• Czech respondents						
Block 1	37	14	6	8	0	36.00
Block 2	37	18	6	11	1	32.50
Block 3	37	5	2	3	0	27.00
Block 4	37	16	3	13	0	31.13
Block 5	36	17	4	13	0	32.71
Block 6	36	2	1	1	0	33.00
• German respondents						
Block 1	42	42	18	22	2	32.88
Block 2	42	34	20	13	1	25.85
Cloze translation						
Block 1	12	33	15	18	0	23.55
Block 2	10	32	18	14	0	20.91
• Words with high predictability						
Block 1	19	30	4	26	0	25.03
Block 2	22	31	5	26	0	24.73
Block 3	22	31	8	23	0	26.42
Block 4	22	30	6	24	0	24.73
Block 5	23	30	6	24	0	25.33
Block 6	24	30	9	21	0	26.63
Block 7	17	32	9	23	0	24.56
• Words with low predictability	22	30	7	23	0	26.27
Cooperative translation experiments	12	32 (16 pairs)	14	18	0	22.20
Baseline cloze experiments (monolingual)						
• Czech respondents						
Condition 1	31	34	14	19	1	26.49
Condition 2	24	34	7	27	0	24.94
Condition 3	18	33	14	19	0	26.35
Condition 4	6	32	11	20	1	25.28
• Polish respondents						
Condition 1	31	32	12	20	0	25.75
Condition 2	25	32	11	21	0	25.91
Condition 3	18	32	11	21	0	31.79
Condition 4	6	29	12	17	0	29.24
Total		1015	350	656	9	

Table 31: Overview of main demographic characteristics in the experiments.

16 This third gender option was labelled *neurčuji* ‘I do not define’ in the CS translation of the drop-down menu of the survey.

Table 31 gives an overview over all participants in the experiments discussed in this thesis. In total, there were 1015 respondents who took part in at least one of the experiments discussed here. The sociodemographic factors age and gender were elicited, but are not evaluated in this thesis.

8. The Principle of the Closest Possible Translation of Sentences

The basics of the closest possible translation principle of individual words were introduced in section 1.3 already: cognates are preferred over non-cognates and orthographically closer cognates are preferred over more distant ones. The translations do not have to be ideal or frequent, as long as the cognates share meaning in at least one possible context.

Similar, although not identical, methods of translation with the purpose to determine linguistic distance were applied in the field before. In a study on the predictors of intercomprehension between Germanic languages, Gooskens & Swarte (2017) used translations of stimulus sentences that were as literal as possible without being ungrammatical in order to measure syntactic distance of sentence material. In the studies summarised in this thesis, however, I use the closest possible translations even if the translations might be ungrammatical in the readers' language. The simple reason is that this ungrammaticality is expected to cause additional cognitive effort for the reader, which should be represented by higher surprisal scores.

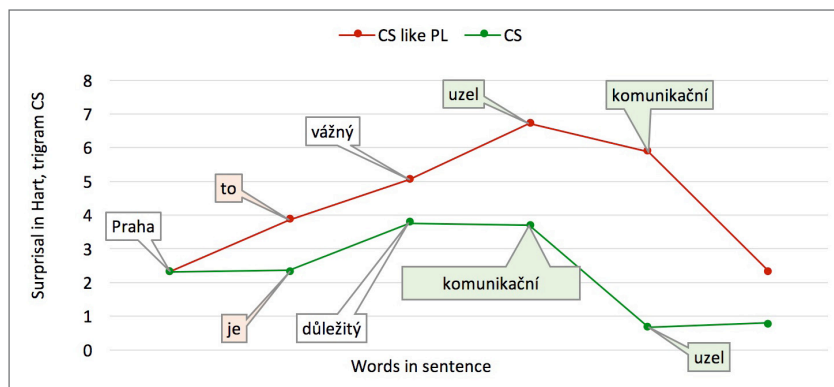


Figure 19: Surprisal of the closest CS translation vs. a good CS translation of a PL stimulus.

Figure 19 shows two surprisal graphs that should represent a Czech respondent's surprisal reading the closest CS translation of the stimulus sentence

Praga to ważny węzeł komunikacyjny. ‘Prague is an important traffic hub.’ (and identifying its constituting words as such) as a transfer base vs. a good CS translation of that sentence. The red graph represents the surprisal curve obtained for the closest CS translation of the sentence with the word-to-word correspondences as presented in Table 32:

PL	<i>Praga</i>	<i>to</i>	<i>ważny</i>	<i>węzeł</i>	<i>komunikacyjny.</i>
Closest CS	<i>Praha</i>	<i>to</i>	<i>vážný</i>	<i>uzel</i>	<i>komunikační.</i>

Table 32: Closest translation principle demonstrated on a PL stimulus sentence.

The green graph in Figure 19 represents a good CS translation of this sentence: *Praha je důležitý komunikační uzel.* The closest CS translation encloses three difficulties that are likely to cause additional cognitive load with the Czech readers:

- Instead of a verb form, there is only the demonstrative pronoun *to* ‘this’. The pronoun *to* also exists in CS, but cannot replace a finite verb in a sentence as in the example here. Acceptable CS translations would in this case be *Praha je ...* ‘Prague is a ...’ or *Praha, to je ...* ‘Prague, that is a ...’.
- The adjective *vážný* ‘serious’ is a cognate to PL *ważny* in other contexts, while in this sentence, the CS adjective *důležitý* ‘important’ would fit better.
- There is divergent word order in the NP *węzeł komunikacyjny* ‘traffic hub’. While PL prefers NA linearization here, a correct CS translation would be in AN: *komunikační uzel.*

These three difficulties are reflected by the higher surprisal values of the red graph as opposed to the green graph in Figure 19. The same principle can apply to smaller units, such as NPs – these are discussed in section 14 and 15.

9. Measures not Considered

For each response to a stimulus, initial hesitation time, typing time, and submission hesitation time were elicited, but these are **not** evaluated in this thesis. Also, the importance of other linguistic features of the stimuli, such as the initial letter of a word, letter shape similarity to L1, the neighbourhood density of words (availability of minimal pairs) as well as non-linguistic factors (age, gender, experience, exposure, intelligence, language awareness) were subject to previous research on intercomprehension, but are **not** examined in this thesis.

10. Scoring Policy Throughout the Experiments

The scoring procedure described subsequently applies to the scoring of responses elicited in the free translation experiments in this thesis. For the most part, it complies with the principles applied by Vanhove (2014, pp. 56-58) in free translation experiments of individual words.

The experiment software was fed with possible correct answers and alternatives beforehand in order to guarantee a quicker automatic classification of answers as correct, wrong or no answer and to provide immediate feedback (smiley) to the respondents. In addition, all responses collected were manually checked for correctness. Responses with obvious orthographic mistakes or typos were not counted as wrong (e.g. *zyvot* instead of *život* ‘life’ as a response to *życie* ‘life’). Responses entered without diacritics were automatically tolerated by the software, i.e. when respondents entered a correct response without diacritics, they still saw a happy smiley. Capitalization was disregarded entirely. Furthermore, the following criteria were applied during scoring of responses:

- If the target words were verbs, forms in **both perfective and imperfective aspect** were accepted.
- If a respondent entered two or more words and one of them was the correct response, this was accepted as correct.
- Both **plural and singular forms of nouns** were counted as correct.
- Responses in the free translation experiments that were not the same POS as the stimulus were counted as wrong. Only if a form of a stimulus could belong to more than one POS, then all possible forms and translations were considered correct. For instance, for PL *raz*, the possible correct responses were CS *rána* ‘stroke, blow’, *ráz/raz* ‘one’, *jedna/jeden* ‘one’, *jednou* ‘once’.
- Nouns that were translated with the equivalent **nominalized CS form**, e.g., if *życie* ‘life’ was translated as *žití* ‘living (N)’ instead of the more appropriate *život* ‘life’, were counted as correct.
- **Responses given in EN are counted as correct.** For instance, there was a case where a respondent entered *I have not seen* where the correct CS translation would have been *Neviděla jsem* ‘I have not seen [fem]’. From the EN response, it is not sure whether the respondent has correctly identified the grammatical gender of the PL stimulus. However, there is no evidence that the respondent did not understand it correctly and therefore the response was counted as correct.

- **Diminutiveness:** when respondents entered non-diminutive forms, e.g. *kniha* ‘book’ for the PL diminutive *książka* ‘book’, these responses **were accepted as correct**.
- Responses consisting of only one letter, a question mark, *nevím* ‘I don’t know’ or a similar expression were counted under the category of *no response given* (manual change from *wrong answer* to *no answer* in the data gathered).
- **Considered wrong:**
 - simple re-types of the stimulus,
 - past tense if stimulus verb was in present tense, and
 - hyponyms and hyperonyms of stimuli, e.g. *příjmení* ‘last name’ instead of *jméno* for *imię* ‘name’.

11. Relevant Statistical Methods in Brief

Statistical correlations between individual predictors and intelligibility scores are estimated by means of the Pearson correlation coefficient r . The higher r is, the stronger is the correlation between two variables. Multiple linear regression models are used to explain the relationship of intelligibility with more than one predictor. These correlations are indicated by the adjusted R^2 . The R^2 indicates how much of the variance in the data can be explained by the model.

In those experiments where two data sets were compared (e.g., two conditions), a one-tailed t -test of independent samples (because there were different respondents in each condition) was performed in order to examine if the two data sets are significantly different. The higher the t value, the greater is the difference between the two data sets.

For all measurements (correlations and t -tests), p values are provided as an indicator of significance. The alpha level is set to 0.05, meaning that results with a $p \geq 0.05$ are considered not significant (ns), $p < 0.05$ is considered significant, $p < 0.01$ very significant, and $p < 0.001$ highly significant. In some tables, the significance levels are indicated by a colour code. Depending on space and layout, an asterisk is added in some cases: $p < 0.001^{***}$, $p < 0.01^{**}$, $p < 0.05^*$. A value of $p < 0.001$, for instance, means that the likelihood for a certain variable to be coincidental is lower than 0.1%.

In order to find combinations of predictor variables which together could best explain intelligibility in the different experiments, the predictors were analysed in multiple linear regression models by adding or removing variables accordingly. The results of the multiple linear regressions are indicated in tables containing the following values:

- Coefficient (relative importance of the predictor for the model): a higher value indicates a greater relative influence of the predictor on intelligibility, a negative value indicates a negative influence of the predictor on intelligibility.
- SE (standard error of the coefficient): indicates the range in which the actual coefficient lies.
- t value: a higher t value indicates a greater influence of the predictor in the data.
- p value (significance of the predictor in the model);
- Adjusted R^2 (estimation of how much of the variance in the data can be explained by the model);
- F crit (significance of the model); and
- F (goodness of fit of the model).

The decision for a certain model was taken according to the F score: a model with a higher F score provides a better fit to the data than one with a lower F score.

Values for standard deviation – SD – are added to calculations of statistical means in order to specify the dispersion of values in the data. A low SD indicates that the data points are close to the mean, while a high SD indicates that the data points are dispersed over a wide range of values.

CHAPTER IV: FREE TRANSLATION OF WORDS WITHOUT CONTEXT

This chapter focuses on the intercomprehension of individual PL words as experimental stimuli presented to Czech readers. In related research, the experimental setting in which isolated cognates in Lx were presented to readers or listeners without context was referred to as cognate guessing task (e.g. Vanhove, 2014), but is hereinafter referred to as *free translation* or *free translation of individual words*. The aim of such an experimental setting is to gain insight into the mainly orthographic factors that influence participants' performance. The absence of context, be it only another word, a sentence or an entire text, should as far as possible exclude the influence of several other linguistic factors. Being provided only with individual words, readers can only rely on cross-linguistic similarities and correspondences in order to correctly guess the meaning of the cognates, for they cannot make use of any contextual clues. The following stimuli were tested in the free translation experiments:

- cognate stimuli containing regular PL-CS correspondences (section 12),
- the 100 most frequent PL nouns (section 13),
- individual words that were part of the sentence stimuli (CHAPTER II and section 15),
- Target words from the cloze translation experiment (section 16).

Except the cognates with applicable PL-CS orthographic correspondences, the stimuli tested in this experimental setting included also non-cognates and false friends. The complete lists of these stimuli and their intelligibility scores are provided in the appendices (Table A 3, Table A 4, and Table A 7).

12. Cognates with Regular PL-CS Orthographic Correspondences

This section presents the findings of a free translation experiment in which PL words containing regular PL-CS orthographic correspondences were translated by Czech readers in a web-based experiment. The stimuli for this experiment were extracted in a computational transformation of parallel word sets in two Slavic language pairs – PL-CS and BG-RU. The experiment aimed at investigating to what extent these closely related languages are mutually intelligible, concentrating on their orthographies as linguistic interfaces to the written text. Besides analysing orthographic similarity, the aim was to gain insights into the applicability of correspondences based on traditional linguistic assumptions

for the purpose of understanding intercomprehension in these language pairs. These were published in the paper *An Orthography Transformation Experiment with Czech-Polish and Bulgarian-Russian* by Fischer et al. (2015) and are summarised under section 1.2. The hypothesis resulting from this is that the more regular the cross-lingual correspondences are, the easier the correspondences should be recognised. Word pairs containing the most regular and frequent correspondences are expected to be translated correctly more often than other stimuli. The concrete research question here is how Czech readers perform when translating PL cognate stimuli containing these regular correspondences.

The computational application of the regular cross-lingual correspondences (Fischer et al., 2015) resulted in a list of cognate pairs from which 296 PL cognates were selected as stimuli for the free translation experiment with Czech readers. In this section, the results of the translation experiment are interpreted together with the results from the computational application.

12.1. Orthographic Distance of the Stimuli

In order to find predictors for the intelligibility of these cognates, the cognates were statistically analysed for the predictors trad LD, pron LD (explanation in section 6.1) – both non-normalised and normalised –, and word length in both languages.

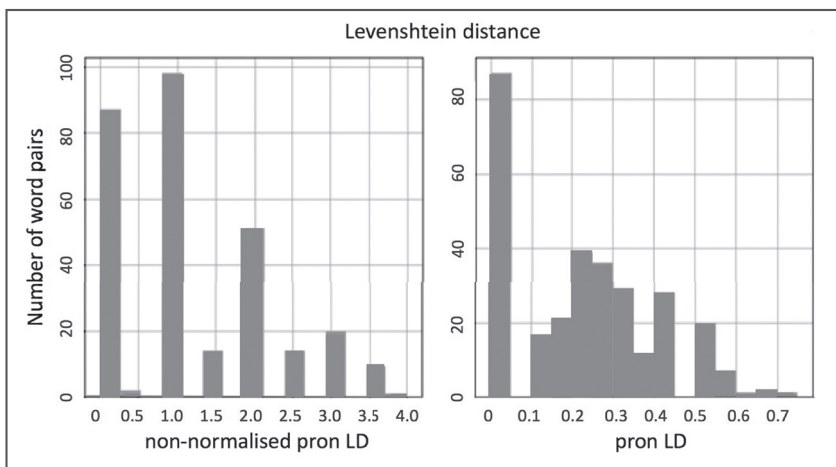


Figure 20: Comparison of non-normalised pron LD (left) and pron LD (right).

The histograms in Figure 20 show a comparison of the distribution of pron LD in total values (non-normalised) on the left side and pron LD when nor-

malised by alignment length (right side). The pron LD values of 0 represent words that differ only in such characters that readers are likely to pronounce correctly, such as PL *w* in the stimulus *woda* ‘water’ which would be *voda* in CS. There is a substantial share of words that do not contain other than these easily pronounceable correspondences and therefore have a pron LD of 0. The major proportion of the cognates have a non-normalised pron LD of less than 2.5. There are only few words with a normalised pron LD of more than 50% and the mean pron LD of the stimuli is 21.7% (SD = 17.9).

Table 33 gives an overview over the mean length and orthographic distance of the cognates.

Length CS	Length PL	Non-norm trad LD	Trad LD	Non-norm pron LD	Pron LD
4.9 (SD = 1.4)	5.3 (SD = 1.5)	1.7 (SD = 1)	32.3% (SD = 16.2)	1.2 (SD = 1)	19.14% (SD = 17.9)

Table 33: Word length and orthographic distance of cognates with regular PL-CS correspondences.

As expected, the CS words are on average shorter than their PL cognates (4.9 < 5.3 characters), which is most likely due to the frequent presence of digraphs in PL.

12.2. Results

The intelligibility scores for the 296 PL stimuli with applicable PL-CS correspondences are presented in Table 34. The scores range from 0% (n = 12) to 100% (n = 85) with a mean intelligibility of 66.7% (SD = 34%). The LDs of the words with 0% intelligibility range from 20% trad LD / 0% pron LD, e.g. for *jesień* ‘autumn’ (CS cognate *jeseň* ‘autumn [literary]’), to a maximum of 75% for *duć* ‘to blow’ (CS *dout*). The LDs of the words with an intelligibility score of 100% range from 6.5% trad LD (0% pron LD) with very similar internationalisms such as *aligator* ‘alligator’ (CS *aligátor*) and *krokodyl* ‘crocodile’ (CS *krokodyl*) to Panslavic vocabulary such as *jarząb* ‘rowan’ with a trad LD of 50% (CS *jeřáb*, pron LD 41.7%,).

∅ Intelligibility	∅ Wrong	∅ No response
66.7% (SD = 33.9)	28.1% (SD = 30.3)	5.2% (SD = 8.8)

Table 34: Intelligibility of cognates with regular PL-CS correspondences.

The intelligibility scores of the individual words are listed in Table A 3 in the appendix.

12.3. Correlations

The correlations of the intelligibility scores of the PL cognates containing regular PL-CS correspondences with the predictors word length, trad LD, and pron LD (non-normalised and normalised) are presented in Table 35. In addition to that, selected correlations between the predictors are presented, too.

		Intelligibility	CS length	Trad LD	Non-norm Pron LD	Pron LD
Word length	CS	0.203				
	PL	0.148	0.888			
Trad	LD non-norm	-0.357				
	LD	-0.466				
Pron	LD non-norm	-0.512		0.852		
	LD	-0.631				
	WAS non-norm	0.108			0.462	0.403
	WAS	0.193			0.477	0.457

Note: The correlations are given as Pearson's *r*.

Table 35: Correlations: intelligibility of cognates with regular PL-CS correspondences and predictors.

The correlation of intelligibility with the normalised LDs and WAS is stronger than that of the non-normalised LDs – this applies to both trad LD and pron LD – and WAS. The normalised pron LD has the strongest correlation of all predictors with intelligibility: $r(296) = -0.631$. Using pron LD instead of trad LD can explain 18% more of the variation in the data ($R^2 = 39.8\% > R^2 = 21.7\%$). In addition, there is a strong correlation in word lengths between both languages ($r(296) = 0.88$), which is not surprising. All of the correlations are significant at the 0.01% level (green colour in Table 35) except the one between intelligibility and PL word length, which is significant only at the 5% level. Figure 21 displays the correlation between intelligibility and pron LD with all stimuli as data points.

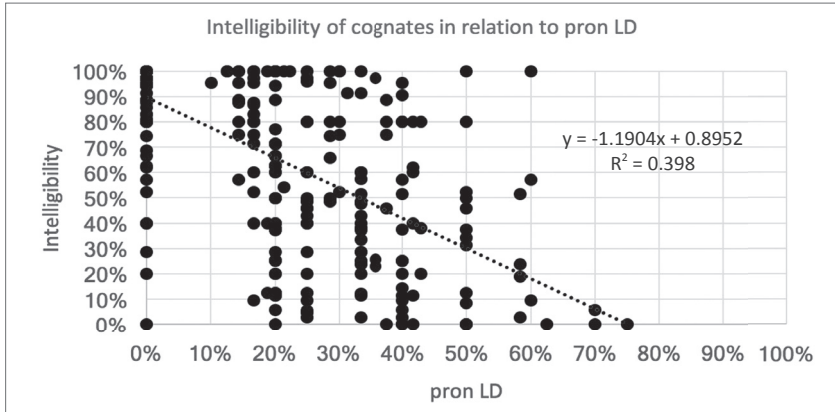


Figure 21: Correlation: intelligibility of cognates with PL-CS correspondences with pron LD.

When adding pron LD and WAS into a model for this scenario, then 40.5% ($R^2 = 0.405, p < 0.01$) of the variation can be explained by these two variables (Table 36). However, it is rather unexpected that the coefficient of WAS in the model is positive, suggesting that a higher WAS leads to more intelligibility, which is counterintuitive.

Model	Coefficient	SE	t	p	Adjusted R ²	F crit	F
Pron LD	-1.294	0.095	-13.589	< 0.0001	0.405	< 0.0001	101.566
WAS	0.093	0.039	2.38	< 0.05			

Table 36: Model for the intelligibility of cognates with regular PL-CS correspondences.

From the most frequently applicable PL-CS correspondences identified in Fischer et al. (2015), the vowel correspondences *a:á* and *y:y* proved to be very easy, as they only require the addition of a diacritical sign. The respondents are likely to be accustomed to this cognitive process, for instance in written communication when using digital devices where diacritics are often dropped for technical or practical reasons. Stimuli containing only one of these correspondences, such as *bal* vs. *bál* ‘dance’ or *jasny* vs. *jasný* ‘clear’ have very high intelligibility scores (80% and 94.3%; ceiling effect). The results for stimuli containing (only) the correspondences *w:v* and *l:l* are similar. Czech readers most probably know that the letter *w* corresponds to the sound /v/ as the letter *w* is also used in foreign and loan words in CS. For a detailed discussion of the rule *l:l*, see section 5.7 in CHAPTER II on the pairwise cooperative experiments.

12.4. Error Analysis

When viewing the results from an error-analytical perspective, the rule *ć:t* seems to cause the greatest problems. Among the stimuli, this rule occurred only with infinitive verb forms (n = 45) plus in the noun *lokieć* ‘elbow’ (CS *loket*) and the numeral *pięć* ‘five’ (CS *pět*). It is very prominent that the infinitive verb ending *-ć* is frequently mistaken for a masculine noun ending corresponding to the CS *-ě* or *-c*. Out of the 45 infinitive verb forms, 22 were translated wrongly with a noun more often than they were translated correctly. Depending on the available possible other options that are still similar enough to the stimulus (neighbourhood density), the responses show different degrees of interferences. For instance, for the stimulus *kopać* ‘to kick’ neighbours with both *-ě* and *-c* ending exist and were among the responses: *kopáč* ‘navvy, digger’ (37.5%) or *kopec* ‘hill’ (20%). Table 37 gives an overview about the infinitive verb forms among the stimuli and the various interfering nouns:

Stimulus	Frequent N translations (wrong)	Wrong Ns %	Correct CS	Correct %
<i>bić</i> ‘to beat’	<i>bič</i> ‘whip’	80.0	<i>bít</i>	11.4
<i>biegać</i> ‘to run’	<i>běžec</i> ‘runner’, <i>běh</i> ‘run’, <i>běhař</i> ‘runner’, <i>bordel</i> ‘disorder’, <i>byt</i> ‘flat’	17.1	<i>běhat</i>	42.9
<i>boleć</i> ‘to hurt’	<i>borec</i> ‘athlete’, <i>bolest</i> ‘pain’, <i>palec</i> ‘finger’, <i>boltec</i> ‘auricle’	37.1	<i>bolet</i>	37.1
<i>bronić</i> ‘to protect’	—	0.0	<i>bránit</i>	40.0
<i>chodzić</i> ‘to walk’	<i>chodec</i> ‘pedestrian’	31.4	<i>chodit</i>	65.7
<i>dąć</i> ‘to blow’	<i>děšť</i> ‘rain’, <i>pláč</i> ‘cry’, <i>táč</i> ‘tray’	8.6	<i>dout</i>	0.0
<i>dawać</i> ‘to give’	<i>prodavač</i> ‘shop assistant’, <i>podatel</i> ‘shipper’	25.0	<i>dávat</i>	12.5
<i>dumać</i> ‘to think’	<i>blbec</i> ‘moron’, <i>máslo</i> ‘butter’, <i>dýmka</i> ‘pipe’, <i>duna</i> ‘dune’, <i>buben</i> ‘drum’, <i>myslitel</i> ‘thinker’	20.0	<i>dumat</i>	60.0
<i>dychać</i> ‘to breathe’	—	0.0	<i>dýchat</i>	75.0
<i>gonić</i> ‘to chase’	<i>chytač</i> ‘catcher’, <i>koně</i> ‘horses’, <i>dub</i> ‘oak’	37.5	<i>honit</i>	20.0
<i>grać</i> ‘to play’	<i>hráč</i> ‘player’, <i>hrad</i> ‘castle’, <i>hra</i> ‘game’, <i>grácie</i> ‘grace’	60.0	<i>hrát</i>	34.3
<i>kąsać</i> ‘to bite’	<i>kasa</i> ‘till’, <i>pokladní</i> ‘casheer’, <i>złoděj</i> ‘thief’, <i>kaše</i> ‘porridge’, <i>kartáč</i> ‘brush’, <i>kapsa</i> ‘pocket’, <i>kasar</i> ‘cracksman’	42.9	<i>kousat</i>	8.6
<i>kopać</i> ‘to kick’	<i>kopáč</i> ‘navvy’, <i>kopec</i> ‘hill’	62.5	<i>kopat</i>	25.0
<i>kosić</i> ‘to mow’	<i>kosa</i> ‘scythe’, <i>košík</i> ‘basket’	75.0	<i>kosit</i>	12.5
<i>kupić</i> ‘to buy’	<i>kupec</i> ‘buyer’, <i>kupující</i> ‘buying (person)’, <i>obchodník</i> ‘businessman’, <i>lupić</i> ‘burglar’	26.0	<i>koupit</i>	57.4
<i>łapać</i> ‘to catch’	<i>łapač</i> ‘catcher’, <i>chytač</i> ‘catcher’, <i>kroupy</i> ‘hail’	62.5	<i>lapat</i>	0.0
<i>lecieć</i> ‘to fly’	<i>list</i> ‘sheet’	12.5	<i>letět</i>	0.0
<i>lepić</i> ‘to glue’	<i>chytač</i> ‘catcher’, <i>lepidlo</i> ‘glue’	25.0	<i>lepit</i>	50.0
<i>leżeć</i> ‘to lie’	<i>lezec</i> ‘climber’, <i>postel</i> ‘bed’	25.0	<i>ležet</i>	50.0

Stimulus	Frequent N translations (wrong)	Wrong Ns %	Correct CS	Correct %
<i>mazać</i> 'to smear'	<i>mazec</i> 'massacre', <i>krém</i> 'cream'	25.0	<i>mazat</i>	37.5
<i>milczeć</i> 'to keep quiet'	<i>miłdżec</i> 'darling', <i>milenc</i> 'lover'	50.0	<i>miłcet</i>	12.5
<i>myć</i> 'to wash'	<i>mić</i> 'ball', <i>myś</i> 'mouse'	75.0	<i>mýt</i>	25.0
<i>prać</i> 'to wash clothes'	<i>práce</i> 'work'	48.6	<i>prát</i>	45.7
<i>rozdzielić</i> 'to divide'	<i>rozdily</i> 'differences'	2.9	<i>rozdělit</i>	71.4
<i>obuć</i> 'to shoe'	<i>obuv</i> 'shoes', <i>obruč</i> 'hoop'	80.0	<i>about</i>	20.0
<i>padać</i> 'to fall'	<i>padák</i> 'parachute'	29.2	<i>padat</i>	62.9
<i>palić</i> 'to burn'	<i>palič</i> 'arsonist'	20.0	<i>pálit</i>	40.0
<i>pić</i> 'to drink'	<i>bič</i> 'whip'	20.0	<i>pít</i>	20.0
<i>plakać</i> 'to cry'	<i>plakát</i> 'poster'	20.0	<i>plakat</i>	60.0
<i>rzezać</i> 'to cut'	<i>řezač</i> 'cutter'	40.0	<i>řezat</i>	40.0
<i>siać</i> 'to sow'	<i>měsíc</i> 'moon', <i>silák</i> 'strongman', <i>sít</i> 'net'	60.0	<i>sít</i>	0.0
<i>siekać</i> 'to chop'	—	0.0	<i>sekat</i>	60.0
<i>skakać</i> 'to jump'	—	0.0	<i>skákat</i>	100.0
<i>slyszec</i> 'to hear'	—	0.0	<i>slyšet</i>	80.0
<i>solić</i> 'to salt'	<i>solíč</i> 'salter'	40.0	<i>solit</i>	40.0
<i>spać</i> 'to sleep'	<i>spáč</i> 'sleeper'	20.0	<i>spát</i>	60.0
<i>stać</i> 'to stand'	<i>stáž</i> 'internship'	20.0	<i>stát</i>	0.0
<i>sypać</i> 'to pour'	<i>sypać</i> 'spreader'	20.0	<i>sypat</i>	60.0
<i>trzeć</i> 'to rub'	<i>teř</i> 'target', <i>trhovec</i> 'marketeer', <i>trhač</i> 'shredder', <i>košík</i> 'basket', <i>trh</i> 'market', <i>plot</i> 'fence'	28.6	<i>třit</i>	5.7
<i>umierać</i> 'to die [impf]'	<i>umíráček</i> 'death knell', <i>hrcbnik</i> 'gravedigger', <i>umělec</i> 'artist'	17.1	<i>umírat</i>	74.3
<i>umrzeć</i> 'to die [perf]'	<i>mrtvola</i> 'corpse', <i>umělec</i> 'artist'	8.6	<i>umřit</i>	51.4
<i>wiedzieć</i> 'to know'	<i>wědec</i> 'scientist', <i>wěštec</i> 'soothsayer'	20.0	<i>wědět</i>	48.6
<i>wstać</i> 'to get up'	<i>tyč</i> 'pole', <i>stav</i> 'state'	25.0	<i>vstát</i>	50.0
<i>wziąć</i> 'to take'	<i>věc</i> 'thing'	12.5	<i>vzít</i>	0.0
<i>zabić</i> 'to kill'	<i>bič</i> 'whip', <i>zajíc</i> 'rabbit', <i>žába</i> 'frog'	14.3	<i>zabít</i>	77.1
Mean		30.0		38.8

Table 37: Verbs with PL-CS correspondences and nouns that respondents translated them with.

On the average, the intelligibility of verbs containing regular PL-CS correspondences was 38.8%, which is substantially lower than the intelligibility of the overall experimental set (66.7%). If the results for the verbs were excluded from the analysis, the intelligibility of the remaining stimuli would be 71.1%. Respondents translated 30.0% of the verbs wrongly with nouns, mostly masculine. In terms of being mistaken for nouns, monosyllabic verbs proved to be more often problematic than polysyllabic verbs, e.g. *bić* 'to beat' (80% nouns), *grać* 'to play' (60% nouns), *myć* 'to wash' (75% nouns). However, this scheme is not consistent for all monosyllabic verbs. For instance, *stać* 'to stand' was problematic (0% intelligibility), but *spać* 'to sleep' was not (60% intelligibility), although the latter would offer the neighbour *spáč* 'sleeper'.

In particular, verbs containing additional PL-CS stem correspondences, such as *q:ou* in *dąć* ‘to blow’, *rz:ř* in *trzeć* ‘to rub’ or *iq:i* in *wziąć* ‘to take’, did not exceed an intelligibility of 6%. In the case of *grać*, the additional *g:h* correspondence seems to have been correctly identified by most of the respondents, since they transformed *g* to *h* in the nouns. Wrong recognition of POS happened also from noun to verb: The noun *grzbiet* ‘back’ was mistaken for a verb in 31.4% of the responses: *drbat* ‘scratch’, *držet* ‘hold’, *hrbit* ‘cower’, *hřmět* ‘rumble’, *mluvit* ‘speak’, *sedět* ‘sit’, *zvracet* ‘throw up’ were among the responses.

The *g:h* correspondence was largely applied successfully, although not in all cases. Again, the recognition and application of this correspondence depended on the neighbourhood density of the stimuli in CS. For instance, *droga* ‘street’ was confirmed to be a false friend of CS *droga* ‘drug’ (82.9% of all responses) and was also translated as *lék* ‘medication’ instead of the cognate *dráha* or the more frequent *silnice*. For *ogon* ‘tail’, although it was translated correctly as *ohon* or *ocas* by 45.7%, the responses also included transformations of *g* to other consonants, such as in *ozón* ‘ozone’ and *okoun* ‘perch’.

Application of the *a:e* correspondence turned out to be no obstacle in feminine noun endings of internationalisms, e.g. *teoria* (CS *teorie*) ‘theory’ with an intelligibility of 88.6% or *energia* (CS *energie*) ‘energy’ with 100%. However, the *a:e* correspondence in stems of rather short words proved to be more difficult, for instance in *las* (CS *les*) ‘forest’ with an intelligibility of only 51.4% or *czajka* (CS *čejka*) with only 11.4%. With *czajka*, 45.7% of the responses maintained an initial *ča-* (*čajka* ‘seagull’, *čárka* ‘comma’, *čaj* ‘tea’, *čajovna* ‘tea-house’). This suggests that at least the *cz:č* correspondence was recognised and successfully applied in these cases, although these translations were wrong. In only two of the responses (5.7%), the *cz* seems to have been transformed into *k*: *kazajka* ‘jacket’ and *krajka* ‘lace’. The phenomenon that stimuli with initial *cz* are translated with words that start with *k* (application of *cz:k* instead of *cz:č*) occurs for other stimuli, too. For instance, *czoło* ‘forehead’ was translated as *kolo* ‘wheel’ or *kouzlo* ‘magic’ in 17.1% of the responses (intelligibility 45.7%).

12.5. Summary

A set of 296 PL cognate nouns, verbs, adjectives and prepositions containing regular PL-CS correspondences were presented to Czech respondents in a web-based free translation experiment. The mean intelligibility of all words tested was 66.7%. It was hypothesised that a pronunciation-based orthographic distance measure (pron LD) would be a better predictor than traditionally calculated orthographic distance (trad LD). It could be shown that pron LD correlates better with intelligibility than trad LD, which confirms the hypothesis.

During the analysis, special attention was also paid to how respondents handled the applicable PL-CS correspondences. Words containing only differences in diacritics were mostly translated correctly (ceiling effect). One of the most problematic correspondences turned out to be *č:t* which is a correspondence in infinitive verb forms. Due to the orthographic and phonetic similarity of PL *č* to CS *č* and *c*, verbs among the stimuli were frequently mistaken for masculine nouns (30% of the responses). The mean intelligibility of the 45 verbs within the stimuli set is only 38.8% and the intelligibility of the stimulus set without the verbs is 71.1%. Monosyllabic verbs, in particular those containing also differences in the stem, proved to be extremely difficult to comprehend. The *g:h* correspondence was largely applied successfully, although this again depended on the available neighbours. The application of the *a:e* correspondence did not pose any problems in feminine noun endings of internationalisms, but proved to be more difficult in stems of rather short words.

13. The 100 Most Frequent PL Nouns

This section analyses the intelligibility of the 100 most frequent PL nouns presented to Czech readers with special attention to their orthographic distance and lexical properties. The list of nouns constituting the stimuli for this experiment appeared as one of the outcomes of Jágrová, Stenger, Marti & Avgustinova (2017) and contains 16 items that are identical with their CS cognates, such as *pan* ‘mister’, *rok* ‘year’. These identical nouns were not tested in the free translation experiment. The remaining nouns (see Table A 4 in the appendices) were presented to Czech readers in this web-based free translation experiment. In previous research in intercomprehension, non-cognates or profile words (Vanhove, 2015) were included into stimuli sets in order to check if the respondents really did not have any knowledge of the language tested. If a respondent was able to translate a non-cognate, she or he was considered likely to have learned the language already. Here, non-cognates from the list were kept in the stimuli set and their intelligibility was evaluated just as those of the

cognates. Accordingly, the answers of respondents who successfully translated a non-cognate are not disregarded for the simple reason that a Czech respondent might know, for instance, about some of the PL-CS false friends without having ever tried to actively learn PL (incidental learning). Evidence that Czech readers know individual PL words are found in the results of the cooperative translation experiments (Chapter II, section 5).

13.1. Results and Correlations

The mean intelligibility of all 84 items from the list of the 100 most frequent PL nouns is 55.03% (SD = 38.83). If the 16 identical nouns were included in the stimuli set, one can speak of an overall intelligibility of the 100 most frequent PL nouns for Czech readers of 71.03%, under the assumption that identical nouns are 100% intelligible.

As the other lists published in Jágrová, Stenger, Marti & Avgustinova (2017), the PL list was translated from PL into CS following the principle of the closest translation (section 8) in order to determine pron LD. Before creating a statistical model for the intelligibility of these nouns, individual predictors are correlated with the intelligibility scores. Figure 22 presents the results with a regression analysis of intelligibility per word in relation to pron LD (blue data points) and trad LD (orange data points). The lower distance values of the data points for the pron LD in comparison to the trad LD manifests itself in a leftward shift of the individual points in Figure 22.

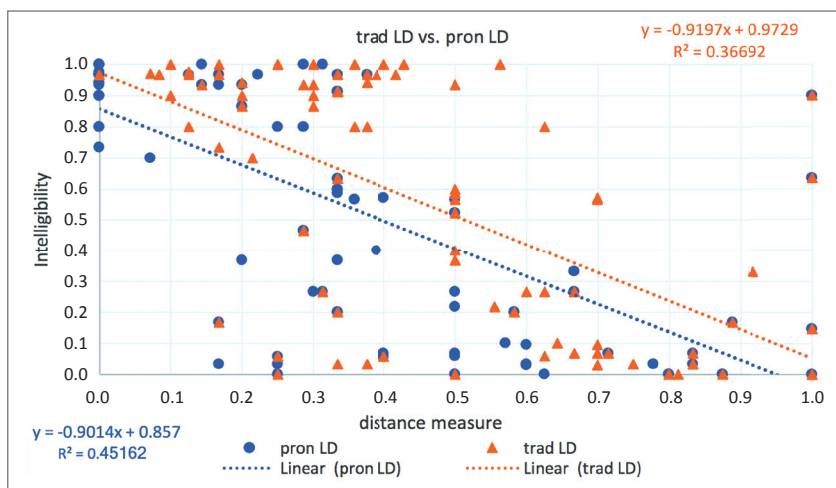


Figure 22: Correlation: intelligibility of the most frequent nouns with trad LD vs. pron LD.

The results reveal that PL-CS distance based on the assumed pronunciations and known CSK correspondences (pron LD) correlates more strongly ($R^2 = 0.45$) with intercomprehension success than trad LD ($R^2 = 0.37$). In the present data set, pron LD can explain 8% more of the variation in the data.

Table 38 presents the correlations (Pearson's r) for the predictors word length, pron LD, WAS, normalised WAS, the binary categories of false friends (FF), non-cognates, and divergent gender for the intelligibility of the 84 non-identical words among the 100 most frequent PL nouns.

	Intelligibility
CS word length	$p > .5$
PL word length	$p > .5$
Trad LD	$-0.61, p < .0001$
Pron LD	$-0.67, p < .0001$
WAS	$-0.43, p < .0001$
FF	$-0.51, p < .005$
Non-cognates	$-0.31, p < .005$
Different gender	$p > .5$

Table 38: Correlations: intelligibility of the 100 most frequent PL nouns and predictors.

No significant correlations could be found for word length in either of the languages. Also, no significant effect of divergent gender could be found here (as opposed to findings on highly predictable target words in section 15). Confirming the findings from the experiment with stimuli containing regular PL-CS correspondences, the negative correlation with pron LD as a normalised measure (as also indicated in Figure 22) is stronger than with the non-normalised pron LD ($-0.67 < -0.59$). A relatively strong negative correlation was found for the category of false friends ($n = 12$; $r(84) = -0.51$). Word adaptation surprisal (WAS) displays a significant, but lower correlation ($r(84) = -0.43$). It has to be kept in mind that in contrast to the stimuli set with applicable regular correspondences in section 02, the present stimuli set consisted not only of cognates, but also of non-cognates, which might have an effect on the importance of the predictors, especially on the failing applicability of regular correspondences when calculating WAS. Also, the stimuli here were only nouns, while in section 12 there were also verbs, adjectives and some prepositions. This might explain why WAS does not have such an impact – the selected multiple linear regression model in Table 39 consists of the variables false friends and pron LD with an adjusted $R^2 = 0.582$, meaning that the two predictors pron LD and the binary

category of false friends can explain 58.2% of the variation in the data. Other possible but less suitable models are listed in Table A 14 in the appendix.

	Coefficient	SE	t	p	Adjusted R ²	F crit	F
Pron LD	-0.787	0.098	-8.064	< 0.0001	0.582	< 0.0001	58.85
FF	-0.424	0.080	-5.288	< 0.0001			

Table 39: Model for the intelligibility of the 100 most frequent PL nouns.

13.2. Error Analysis

When analysing the errors made by respondents, we see that there were a number of cases of L2, L3, Ln interference with certain stimuli. Ln interference as a factor influencing human performance in intercomprehension can hardly be predicted with the usual distance measures in experiments with a large number of respondents where each of them has their own individual Ln repertoire. Table 40 shows all occurrences of obvious Ln interferences with their frequencies. Obvious interferences occurred with only six of the stimuli. For instance, *wiek* ‘age’ was translated as *týden* ‘week’ by some of the respondents, probably influenced by EN *week*, instead of the correct *věk*. The interferences did not only occur from EN (n = 4) or DE (n = 2), but also from other Slavic Ln, as in the case of *godzina* ‘hour’ which was translated as *rok* ‘year’, most likely influenced by BG or BCS *godina* ‘year’.

PL stimulus	Correct %	Wrong %	Ln interference among wrong responses %	Wrong responses	Possible source of interference
<i>bank</i> ‘bank’	86.67	13.33	7.00	<i>lavice</i> ‘bench’, <i>břeh</i> ‘bank’, <i>hrana</i> ‘edge’	EN <i>bank</i> and <i>bench</i> , DE <i>Bank</i> ‘bench’
<i>bóg</i> ‘God’	26.67	53.33	18.75	<i>lod’</i> ‘ship’, <i>batoh</i> ‘rucksack’	DE <i>Bug</i> ‘bow (of a ship)’ and EN <i>bag</i>
<i>wiek</i> ‘age’	80.00	13.33	50.00	<i>týden</i> ‘week’	EN <i>week</i>
<i>głos</i> ‘voice’	26.67	60.00	16.67	<i>sklenice</i> ‘glass’, <i>lesk</i> ‘gloss’	EN <i>glass</i> or DE <i>Glas</i> ‘glass’, EN <i>gloss</i>
<i>godzina</i> ‘hour’	46.67	36.67	9.09	<i>rok</i> ‘year’	BG or BCS <i>godina</i> ‘year’
<i>raz</i> ‘time, stroke’	73.33	26.67	25.00	<i>ted’</i> ‘now’	SK <i>teraz</i> ‘now’
Mean			21.09		

Table 40: Ln interferences among the translations of the most frequent PL nouns.

Regarding the non-cognates among the stimuli ($n = 9$), *sprawa* ‘matter’, *wniosek* ‘suggestion’, *wynik* ‘result’ were not translated correctly by any of the respondents. The non-cognates *kobieta* ‘woman’, *okres* ‘time’, and *rzecz* ‘thing’ did not exceed intelligibility scores of 7%. However, respondents were apparently able to correctly infer the meaning of the internationalisms *decyzja* ‘decision’ (CS: *rozhodnutí*; intelligibility: 14.7%), *punkt* ‘point’ (CS: *bod*; intelligibility: 63.3%), and *numer* ‘number’ (CS: *číslo*; intelligibility: 90%) through DE or EN.

With some cognate stimuli that offered two orthographic neighbours differing only in one vowel letter, it happened that one of the options was more dominant, i.e. chosen more often as a translation. For instance, the stimulus *strona* ‘page’ was translated wrongly as *struna* ‘string’ significantly more often (83.33%) than it should have been (correct: *strana*, intelligibility: 16.67%). This is especially interesting, as when comparing the frequencies of the two concurrent neighbours *struna* and *strana*, *struna* has a corpus frequency of only 11.93 i.p.m.¹⁷ (related to the whole SYN2015 corpus, Křen et al., 2015), which is low compared to *strana* which has a corpus frequency of 671.31 i.p.m. (Křen et al., 2015) and is also among the 100 most frequent CS nouns (Czech National Corpus, 2010).

13.3. Summary and Outlook

Of the 100 most frequent PL nouns published in Jágrová, Stenger, Marti & Avgustinova (2017), 84 nouns that do not have identical cognate translations in CS were presented to Czech readers in a web-based free translation experiment. On the average, about 55% of these nouns were translated correctly by the respondents. Hence, from the 100 most frequent PL nouns, Czech readers should be able to comprehend about 71% on the average. This is in line with the findings from the free translation experiment of cognates containing regular PL-CS correspondences in which Czech respondents were also able to correctly translate about 71% of the stimuli that were not verbs. This again suggests that PL nouns should be easier to understand for CS readers than infinitive verb forms.

In addition to the PL-CS linguistic distances measured in Jágrová, Stenger, Marti & Avgustinova (2017), a pronunciation-based distance (pron LD) of the PL stimuli was calculated and was hypothesised to be a more representative predictor for their intelligibility to Czech readers than traditionally calculated orthographic distance (trad LD), as there is a relatively high orthographic distance in this language pair and the (assumed) pronunciation of PL words

17 Instances per million

might be closer to CS than PL orthography is. It was found that pron LD correlates stronger with the results than trad LD, which confirms the hypothesis. In addition to that, it was found that when adding a variable about whether words are false friends or not (distinct from lexical distance) to the variable of pron LD in a multiple linear regression model, the two variables together can explain 54.5% of the variation in the data. One of the phenomena that cannot be explained by this model are interferences from languages other than CS. In total, about 21% of all wrong responses could be shown to be due to Ln interferences.

Nouns that do not have any cognate translation in CS were also part of the PL frequency list and hence were tested in the experiment just as all other cognate stimuli. Correct translations of such non-cognates were exceptional. However, respondents were able to correctly translate some of these words that are internationalisms through their knowledge of DE or EN, even though these words have no internationalism translations in CS.

The most frequent nouns of BG, CS, and RU published in Jágrová, Stenger, Marti & Avgustinova (2017) have also been uploaded to the experiment website and experimental data is being gathered. As soon as enough data will be available, they can be compared with the distance measures and asymmetries in Jágrová, Stenger, Marti & Avgustinova (2017) and Stenger, Jágrová et al. (2017). In order to examine whether the assumed pronunciation influences intelligibility in the other language-reader combinations, too, pronunciation-based matrices and LD calculations should be established.

CHAPTER V: FREE TRANSLATION OF NPS

This chapter analyses the impact of a canonical grammatical feature of PL – the postmodification of nouns by classifying adjectives – on the intelligibility of PL for Czech readers. I postulate that post-nominal adjectives in PL NPs cause additional processing effort for Czech readers when they attempt to read and understand them, since this feature is not as frequent and typical in CS as it is in PL. As a representation of the predictability of words in NPs, surprisal scores obtained from trigram LMs are correlated with the results of a free translation experiment with PL NPs in the AN (adjective+noun) and NA (noun+adjective) condition. In a subsequent digression, the results are compared to those from an experiment in which PL internationalisms were presented to German respondents.

The main part of this chapter and the stimuli discussed here appeared previously in

Jágrová, K. (2018). Processing Effort of Polish NPs for Czech Readers – A+N vs. N+A. In W. Guz & B. Szymanek (Eds.), *Canonical and non-canonical structures in Polish. Studies in Linguistics and Methodology* (Vol. 12, pp. 123-143). Lublin: Wydawnictwo KUL.

In this previous publication, the intelligibility of the tested NPs ($n = 109$) was correlated with a hypothesised “overall difficulty” (Jágrová, 2018, p. 132). In the present section, a linear regression is applied instead of the overall difficulty, since the regression model should better weight the individual variables that constitute the actual difficulty of NPs, while in the concept of “overall difficulty” as presented in Jágrová (2018), both linguistic distance and surprisal were treated with equal weight (see subsection 14.5 for details). Furthermore, variables for false friends and difference in grammatical gender were added to the regression model. This regression is applied to the 30 most representative NPs (428 data points in each condition, NPs with at least 10 responses in both conditions) from the data set in order to exclude the influence of the different data sizes here, because the data sizes presented in Jágrová (2018) were not evenly distributed over the different NPs. This happened due to the different blocks of NPs that participants were assigned, so that the numbers of translations per NP range from 3 up to 17 translations in each of the conditions (Jágrová, 2018, p. 134).

14. Adjectival Modification in PL

The comparison and distinctive systematization of the AN vs. NA linearisation in PL NPs has been subject to numerous studies. According to Cetnarowska, “the most common position of classifying modifiers in Polish is the post-head position” and “the classifying post-head adjectives are subjective” (Cetnarowska, 2013, p. 19). This feature is generally speaking possible in CS (e.g. in zoological terminology, scientific discourse), too, but it is rather infrequent and often stylistically marked (archaic, literary language). Cetnarowska, Pysz & Trugman (2011) also observe this tendency for PL, stating that there is “a slight difference in the interpretation of AN and NA units containing classifying adjectives in Polish since the AN phrases are perceived as less formal while NA units are typical of scientific discourse” (as cited in Cetnarowska, 2013, p. 20). In both languages, the NA linearisation can also be used to emphasise differences between items or in enumerations. Figure 23 and Figure 24 attempt to quantify the typicality of the two linearisations in PL and CS by means of a comparison of their surprisal values.

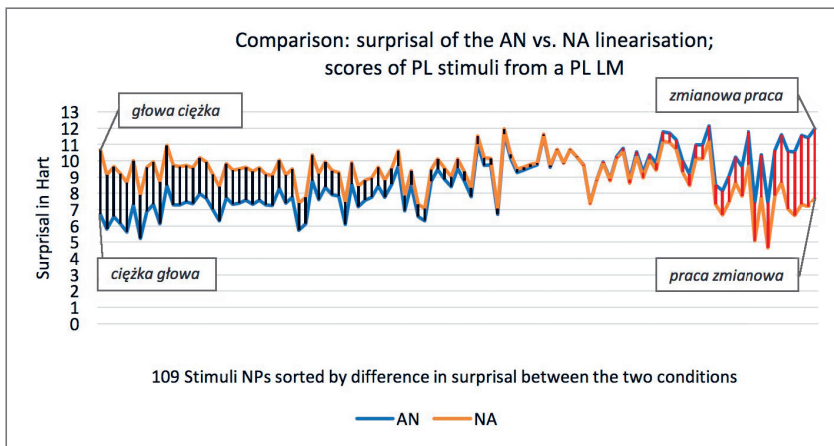


Figure 23: Comparison of typicality of the NP stimuli: AN vs. NA (Jágrová, 2018, p. 130).

Figure 23 compares the typicality of AN vs. NA linearisation of the PL NPs with the help of two surprisal graphs. The surprisal scores were obtained from a PL LM, while those in Figure 24 were obtained from a CS LM. The higher the surprisal score, the more surprising or unpredictable should an NP be and the greater should be the cognitive effort to process this NP during reading. The sums of the surprisal scores per NP (surprisal of noun + surprisal of adjective)

are displayed on the *y*-axis. Each of the 109 NPs is represented by a pair of data points along the *x*-axis (the labels for the complete set of NPs along the *x*-axis are omitted for reasons of readability). The blue data points (connected to a blue line) are the surprisal scores of the NPs in AN linearisation. They are vertically connected to the orange ones – the same NPs, only in NA linearisation. The longer the connecting line between the blue and orange data points in a pair, the greater is their difference in surprisal values. When the connecting line is black, this NP is more typical in the AN linearisation than in NA. When the connecting line is red, this NP is more typical in the NA linearisation than in AN. The leftmost NP pair is *głowa ciężka* / *ciężka głowa* ‘heavy head’ for which the difference in typicality is the greatest in the sample:

$$\text{surprisal}(\text{ciężka głowa}) - \text{surprisal}(\text{głowa ciężka}) = 6.7 \text{ Hart} - 10.7 \text{ Hart} = -4.0 \text{ Hart}$$

According to the surprisal scores obtained from the LM, the NP *ciężka głowa* is much more typical than *głowa ciężka*, hence the higher surprisal value of *głowa ciężka*. The opposite is true for the rightmost NP pair *praca zmianowa* and *zmianowa praca* ‘shift work’ – here, the NA linearisation is more typical:

$$\text{surprisal}(\text{zmianowa praca}) - \text{surprisal}(\text{praca zmianowa}) = 11.96 \text{ Hart} - 7.69 \text{ Hart} = 4.27 \text{ Hart}$$

Overall, 73 of the 109 NPs (67%) in Figure 23 are more likely to appear in the AN order and 36 (33%) in the NA order. However, it has to be noted that in about a fourth of the NPs, the difference in surprisal is negligible. The mean difference in surprisal between the two linearisations is 0.48 Hart.

Figure 24 displays the surprisal values of the closest CS translations of the PL NPs visualised according to the same principle for a comparison of the typicality of the AN vs. NA linearisation.

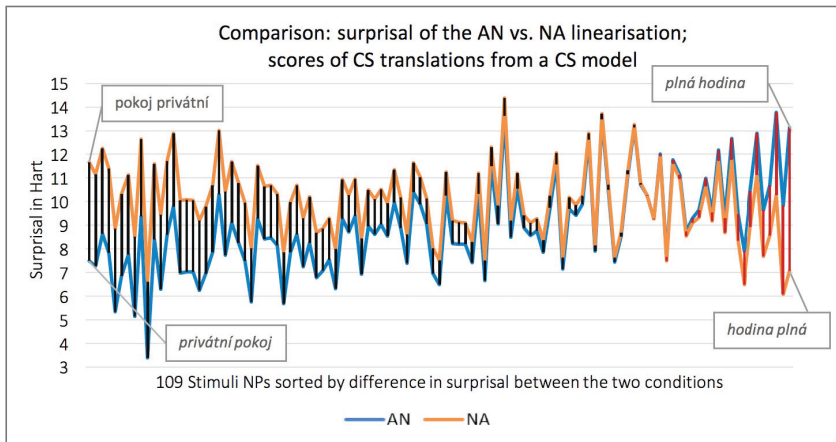


Figure 24: Typicality of closest CS translations of the NPs: AN vs. NA (Jágrová, 2018, p. 130).

First of all, a comparison of Figure 23 and Figure 24 shows that the surprisal scores of the closest CS translations are higher in general: While the highest CS value is around 14, the maximum value for PL is around 12. This is because the closest CS cognate translations are not as natural as what would be considered a good CS translation. For instance, while the closest possible translation of *pokoj prywatny* ‘private room’ is *pokoj privátní*, a good CS translation would be *soukromý pokoj* which would have a lower surprisal score than the closest possible translation. More of the blue data points are beneath the orange points than the other way round in the CS graphs, which confirms the intuition that AN is more usual in CS. According to the scores, only 15 of the 109 NPs (13.8%) should be more typical in the NA than in the AN linearisation, which is less than in the PL sample. Again, the difference in surprisal between the two linearisations is negligible for about a fourth of the NPs. The NP pair *plná hodina / hodina plná* ‘(a) full hour / an hour full of ...’ has the biggest difference in surprisal between the two conditions with NA being more typical. This might be due to the relative frequency with which the combination of the words *hodina* followed by *plná* occurs in the corpus in general, since the model also captures such occurrences as *hodina plná radosti* ‘an hour full of joy’ in which the adjective is followed by a genitive form and both modify the head of the NP – *hodina*. On the other end of the graph in Figure 24, there is the NP *privátní pokoj* vs. *pokoj privátní* ‘private room’ with a clear preference for the AN linearisation.

14.1. Hypothesis

The limited context created by the combination of the adjectives and nouns might influence the intelligibility of these items. The underlying hypothesis is that the unexpectedness of the post-nominal attributes in an NP will cause greater processing effort for CS readers when trying to understand it than in an NP with a pre-nominal attribute. The greater processing effort is expected to manifest itself not only in longer response times, but also in a lower intelligibility of the NPs in the NA condition. This tendency should be reflected in the correlations with the surprisal scores of the two conditions. Also, it is likely that respondents might fail to recognise the POS of the stimuli in NA linearisation more often than in AN linearisation.

14.2. Method

In total, 109 different PL NPs were presented to Czech readers in two different conditions: AN and NA linearisation. This resulted in 218 NPs that were presented in blocks of 4 x 36 and 2 x 37. The experiment software on the website automatically assigned one of the blocks to the participants. After having completed a block, participants could choose to proceed with another block. The stimuli blocks were activated successively in such a way that each NP was presented to a participant in only one of the two conditions. The number of NPs in each condition was evenly distributed among the blocks.

All NPs were constructed out of the most frequent nouns (discussed in section 1.3 and published in Jágrová, Stenger, Marti & Avgustinova, 2017) which were also presented in free translation experiments individually (section 13). These were combined with the most frequent adjectives of PL – both nouns and adjectives were extracted from the readily available frequency list of PL lemmas (Broda & Piasecki, 2016). Magdalena Telus, a linguist and lecturer of PL at Saarland University, looked over the NPs while checking them for plausibility. All possible correct translations for each NP were considered, also considering the differences in meaning that could occur between the two conditions. In addition to these constructed NPs ($n = 100$), 9 other NPs from the sentences in the cooperative translation experiments (CHAPTER II) were added to the stimuli for a possible comparison.

14.3. Distance of the Stimuli

In the first analysis (Jágrová, 2018), orthographic distance of the stimuli NPs was calculated as trad LD. As for lexical distance, only 10 of the nouns and 15 of the adjectives in the NPs are non-cognates (lexical distance score of 1).

14 of the NPs consist of a cognate and a non-cognate. In Jágrová (2018), false friends were counted as having a lexical distance of 2 (double of a regular non-cognate), since these words are expected to be more difficult to translate than non-cognates. This resulted in a mean lexical distance of 1 with 11 NPs that consist either of two non-cognates or of a false friend and a cognate. Only 2 NPs are combinations of two false friends: *ostatni okres* ‘last period’ and *kolejny raz* ‘another time’ – these are not equal in meaning to the very similar CS NPs **ostatní okres* ‘other district’ and *kolejní ráz* ‘rail character’. NPs that consist of 2 cognates (n = 82) were assigned a lexical distance of 0.

The mean trad LD is 40% for the adjectives and 33% for the nouns. A number of nouns are identical – these were not tested in the free translation experiments presented in section 13 (e.g. *rada, projekt, firma*). Nevertheless, there are also such distant cognate pairs among the NPs as *mężczyzna – muž* ‘man’ with a trad LD of 83.33%. There are no identical adjectives, but some differ only in diacritics (e.g. *podobny – podobný*). In addition to the measures evaluated in Jágrová (2018), pron LD of the NPs for which the most representative data was gathered (n = 30) is included in a regression analysis in this section. This is done in order to test the prevailing hypothesis that pron LD correlates better with intelligibility than trad LD. The pron LD of the 30 most representative NPs is 23.56% when counted only as orthographic distance on cognates (as explained in section 6.1.) or 36.17% when counted as total distance. For calculating total distance (only applied in this section), lexical and orthographic distance are summarised by treating non-cognates (units with a lexical distance of 1) as having an orthographic distance of 100%.

14.4. Total Difficulty of the Stimuli

The calculation of the “overall difficulty” (Jágrová, 2018) is demonstrated as follows on the NPs *komunikacyjny węzeł* and *węzeł komunikacyjny* ‘traffic hub’: *węzeł* is a cognate to CS *uzel* ‘knot, hub’. This cognate pair has a pron LD of 40% and makes up one half of the NP. The adjectives *komunikacyjny* and *komunikační* have a pron LD of 19.23%. The mean pron LD of *komunikacyjny węzeł* and *komunikační uzel* therefore is 29.62% (see Figure 25). As demonstrated in Table 41, this value is multiplied by the sum of the surprisal values of the two words (as scored by the CS LM) for each of the two linearisations, resulting in an estimated difficulty score of 2.35 for the AN condition and 3.42 for the NA condition.

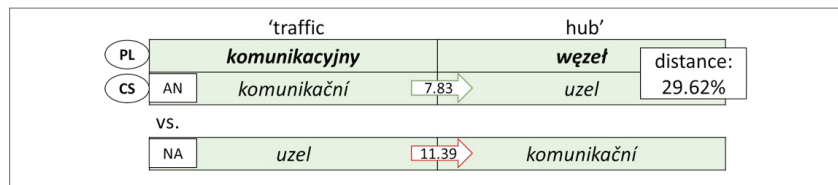


Figure 25: Difference in expected processing effort between the two linearisations.

PL stimulus	In relation to CS		Dist \emptyset	Surp 1	Surp 2	Surp Σ	Diff
<i>komunikacyjny węzeł</i>	<i>komunikační</i>	<i>uzel</i>	0.30	5.10	2.73	7.83	2.35
<i>węzeł komunikacyjny</i>	<i>uzel</i>	<i>komunikační</i>		5.77	5.62	11.39	3.42

Table 41: Example for the calculation of overall difficulty for NP stimuli.

Hence, the somewhat higher overall surprisal value for *węzeł komunikacyjny* predicts that respondents will provide a lower number of correct translations and/or that there will be a higher processing time than for *komunikacyjny węzeł*. Table 42 provides the means of the possible predictors for the intelligibility when deciphering the PL NPs in both conditions – both for the 109 NPs as of Jágrová (2018, p. 132) and of the 30 most representative NPs. The distances of adjectives and nouns are not viewed here separately for the 30 representative NPs.

	AN (n = 109)	NA (n = 109)	AN (n = 30)	NA (n = 30)
Mean surprisal per NP	9.46 Hart	10.02 Hart	8.74 Hart	10.04 Hart
Mean lexical distance per NP	20.18%		33.33%	
Mean lexical distance As	21.10%			
Mean lexical distance Ns	19.27% ¹⁸			
Mean orthographic distance per NP	35.62% (trad LD)		23.56% (pron LD)	
Mean orthographic distance As	39.82%			
Mean orthographic distance Ns	33.08%			
Mean overall difficulty per NP	4.66	5.14	4.80	5.42

Table 42: Comparison of linguistic distance and surprisal scores: AN vs. NA.

18 The lexical distance of the nouns indicated in Table 42 is significantly higher than PL-CS lexical distance as published for instance in Jágrová, Stenger, Marti & Avgustinova (2017) (9%), because false friends are assigned a distance value of 2 here, as explained in 14.3.

The mean surprisal values of the 30 most representative NPs are lower for the AN than for the NA condition (8.74 Hart < 10.04 Hart), but this difference is not significant¹⁹. The same applies to the overall difficulty: The values in the two conditions do not differ significantly ($t(58) = -0.67, p > 0.05$).

14.5. Results

Regardless of the condition, responses given in both AN and NA linearisation were counted as correct only if both of the actual separate words were correctly translated. Intelligibility of the NPs and processing time as experimental results are compared between the two conditions for the whole data set ($n = 1293 / n = 1296$) and for the most representative NPs ($n = 30$) in Table 43 (cf. Jágrová 2018, pp. 134-136).

	AN (n = 1293)	NA (n = 1296)	AN (n = 30)	NA (n = 30)
Correctly translated NPs	49.50%	41.63%	44.00%	41.31%
Correctly translated: only As	66.51%	61.60%		
Correctly translated: only Ns	63.57%	61.99%		
Mean processing time of all NPs	10.08 s	10.04 s	10.20 s	10.37 s
Mean processing time of correctly translated NPs	8.53 s	8.43 s	7.59 s	8.94 s

Table 43: Intelligibility and mean processing time of NPs: AN vs. NA.

14.5.1. Intelligibility

Both for the whole data set as of Jágrová (2018) and for the 30 representative NPs, the intelligibility is slightly higher for the AN condition than for the NA condition (49.5% > 41.63% and 44% > 41.31%). While the difference between the two conditions is almost 8% for all data points collected, this difference, however, is below 3% for the most representative NPs which is not significant ($t(58) = 0.28, p > 0.05$). When viewing the correctly translated adjectives and nouns individually, somewhat more of each are translated correctly in the AN condition. The individual correlations (Pearson's r) regarding the intelligibility of the most representative NPs and the relevant predictors are provided in Table 44.

Condition	Total dist	Pron LD	Lex dist	Surp AN	Surp NA	FF	Gender
AN	-0.74***	-0.27	-0.65***	-0.34	X	-0.54**	-0.30
NA	-0.73***	-0.22	-0.66***	X	-0.45*	-0.54**	-0.28

Table 44: Correlations of predictors with intelligibility of NPs: AN vs. NA.

19 The data in this section can vary slightly from those in Jágrová (2018) because of subsequent corrections and/or rounding up and down. This has no impact on the overall results.

The colour code in Table 44 represents the statistical significance which is additionally indicated by the asterisk. Among all variables tested, the highest correlations for both conditions were found between total distance (unifying lexical distance and pron LD) and intelligibility ($r(28) = -0.74$, $p < 0.0001$ for AN and $r(28) = -0.73$, $p < 0.0001$ for NA). Lexical distance alone has a somewhat lower correlation in both conditions, followed by the variable of false friends. Surprisal only has a significant correlation with intelligibility in the NA condition ($r(28) = -0.45$, $p < 0.05$), but not in the AN condition. Neither pron LD nor grammatical gender correlates with intelligibility here. The missing correlation with divergent gender might be due to the fact that there was only one NP containing a difference in gender among one of the 30 NP pairs here. The lacking correlation of intelligibility and pron LD here could be due to the fact that an NP was only counted as correctly translated when both words were correct. Therefore, the variables incorporating lexical difficulties (lex dist and total dist) correlate stronger with intelligibility in this experiment.

In order to answer the question whether the factors distance and surprisal interplay, the relationship of the possible variables from Table 44 was modelled in a multiple linear regression analysis in Table 45.

AN	Coefficient	SE	t	p	Adjusted R ²	F crit	F
Mean total dist	-1.294	0.252	-5.135	< 0.0001	0.520	0.000	16.685
SURP AN	0.018	0.029	0.625	< 0.05			
NA							
Mean total dist	-1.044	0.196	-5.320	< 0.0001	0.583	0.000	21.307
SURP NA	0.051	0.021	2.391	> 0.05			

Table 45: Regression models for intelligibility of the NPs: AN vs. NA.

The complete regression analysis in which the different combinations of variables were tested for the best fit of the model can be found in Table A 15 and Table A 16 of the appendices. For both conditions, models consisting of total distance and surprisal were selected. The model performs slightly better for the NA condition where it can account for 58% of the variation in intelligibility ($R^2 = 0.58$, $p < 0.0001$) than for the AN condition ($R^2 = 0.52$, $p < 0.0001$). This suggests that surprisal has slightly more influence on intelligibility in the NA than in the AN condition. However, the coefficients of surprisal in the models are positive, which is counterintuitive.

14.5.2. Processing time

Regarding the processing time of all NPs (not only the correctly translated ones), all mean values lie between 10 and 10.5 s (Table 43). When viewing only the correctly translated NPs, the values are somewhat lower on the average (between 7.59 and almost 9 s). As shown in Table 43, the difference in the mean total processing times of correctly translated NPs is minimal between the two conditions and is neither significant for the whole dataset (visualised in light vs. dark grey data points in Figure 26; Jágrová, 2018, p. 134) nor for the representative NPs ($t(22) = -1.28, p > 0.05$; only data of 12 NPs in the two conditions could be compared here, as there were 0 correct translations for the remaining 18 NPs in at least one of the conditions and processing time is only considered for correctly translated NPs).

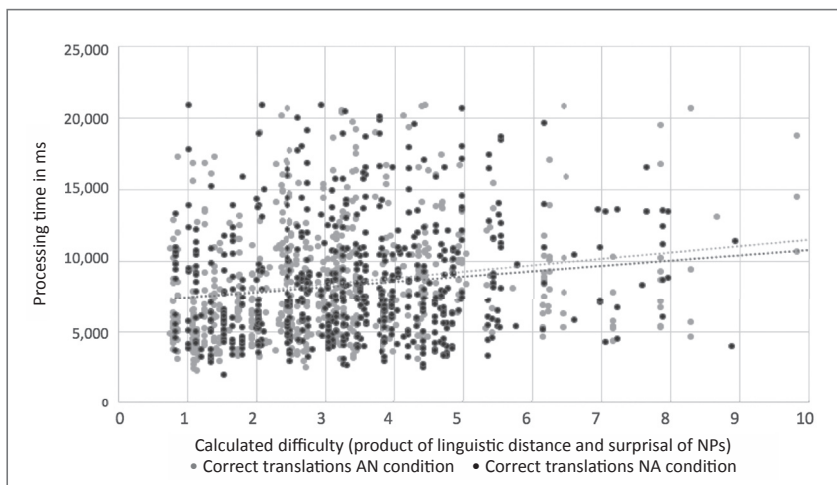


Figure 26: Correlation: processing time in ms for all correct translations and calculated overall difficulty (Jágrová, 2018, p. 135).

Figure 26 (Jágrová, 2018, p. 135) shows a comparison of the processing time of all correctly translated NPs in AN vs. NA condition relative to the total difficulty of the NPs (outliers with a difficulty of more than 10 excluded). Although the correlation of total difficulty and processing time is fairly low, it is significant in both conditions ($r(615) = 0.194, p < 0.001$ for AN and $r(638) = 0.259, p < 0.001$ for NA) (Jágrová, 2018). For all data points of the most representative NPs (AN and NA together), no significant correlation of processing time and surprisal could be found ($r(21) = 0.24, p > 0.05$). When analysed for the

two conditions separately, surprisal has a low, but significant correlation with processing time in the NA condition ($r = 0.105$, $p < 0.01$). Processing time was also correlated with the predictors lexical distance of NPs, orthographic distance of NPs, total distance of NPs, surprisal of NPs in both conditions, and difficulty of NPs in AN and NA condition – the correlations as published in Jágrová (2018, p. 137) are presented in Table 46.

	Lex dist	Trad LD	Total dist	Surprisal	Difficulty
Processing time AN	0.190***	0.140***	0.245***	0.121	0.259***
Processing time NA	0.118**	0.096*	0.157***	0.105**	0.194***

Table 46: Correlations: processing time and predictors in AN vs. NA.

None of the other predictors correlates better with processing time in either of the conditions than the total difficulty does.

14.5.1. Wrong recognition of POS

Two noteworthy types of errors in the translations could be observed: First, adjectives were sometimes mistaken for nouns and nouns for adjectives. Second, the stimuli were translated as NPs with a genitive attribute. For instance, *obca rodzina* ‘foreign family’ was translated as *rodná obec* ‘native village’ or *rodinná obec* ‘family village’ with the adjective *obca* mistaken for the noun *obec* and the noun *rodzina* mistaken for the adjectives *rodná* or *rodinná*. For this specific NP, this error happened 8 times in the AN condition and 13 times in the NA condition. As an example for the second type of error, *zasada zla* ‘bad principle’ was translated as *semeno zla* ‘seed of evil’, *podstata zla* ‘essence of evil’ or *zásada zla* ‘base of evil’ with the adjective mistaken for a genitive, although the latter translation is counted as correct, since PL *zla* can indeed be the genitive form of *zlo* ‘evil’.

The first type of error (mistaking adjectives for nouns and nouns for adjectives) occurred 14 times in AN and 20 times in NA condition of the whole data set. The second type of error (mistaking adjectives for genitive nouns) occurred only 3 times in AN but 35 times in the NA condition. Besides these, there were also other types of errors where wrong recognition of POS is involved, e.g. translations consisted of verb phrases or only adjectives, nouns or combinations of adverbs and adjectives. In total, wrong recognition of POS occurred 37 times in the AN and 98 times in NA condition. Hence, it can be concluded that a greater difficulty of the NA linearisation in PL NPs manifests itself in greater difficulties with recognising the POS of the words constituting the stimuli.

14.6. Summary

A set of 109 PL NPs that had been constructed out of the most frequent nouns and the most frequent adjectives were presented to Czech respondents in a web-based free translation experiment. In order to predict the intelligibility of these NPs to the respondents, Jágrová (2018) calculated the overall difficulty for the NP stimuli in AN vs. NA condition as a product of linguistic distance (trad LD) and the surprisal values of the NPs obtained from a PL LM. This means that in that method, both linguistic distance and surprisal were given the same weights. In this thesis, however, the relation between possible predictors and the data set is analysed in a regression model, since the individual predictors might have different weights in relation to intelligibility, which might deliver a more accurate model. This regression analysis was implemented for the NPs for which the most representative data was gathered ($n = 30$).

Viewing the predictors, the fact that neither surprisal nor the overall difficulty values as used in Jágrová (2018) differ significantly for the 30 NPs with the most representative data in the two conditions suggests that these two variables might be unsuitable predictors for intercomprehension in this scenario. As for the results, the two conditions do not differ significantly in their intelligibility or total processing times. Intelligibility of the NPs in AN was only slightly higher than in the NA condition. Only a low but significant correlation between processing time and surprisal could be found in the NA condition.

According to the regression model, 58% of the variation in intelligibility can be explained by combining total distance (unifying pron LD and lexical distance) and surprisal. This model is somewhat stronger for the NA than the AN condition. This suggests that predictability effects are not of primary relevance for the intelligibility of NPs with AN linearisation, but that they do play a small but significant role for the intelligibility of NPs with NA linearisation. Nevertheless, the greater difficulty of the NA condition seems to manifest itself with regard to POS recognition: Respondents failed to correctly recognise the POS of the stimuli in NA linearisation about 2.6 times more often than in AN linearisation. Consequently, the greater difficulty of the postnominal attribute in NPs does, in comparison to the prenominal attribute, not manifest itself in significantly lower intelligibility scores or processing times, but rather in the type of errors made, specifically in the frequency of wrongly recognised POS.

14.7. Digression: PL NPs Presented to German Readers

Parts of this subsection have previously been published in German in Jágrová, K., Stenger, I. Avgustinova, T. 2018. Polski nadal nieskomplikowany? Interkomprehensionsexperimente mit Nominalphrasen. *Polnisch in Deutschland. Zeitschrift der Bundesvereinigung der Polnischlehrkräfte*, 5, 20-37.

They were translated freely into English and reformulated in this subsection.

14.7.1. Hypothesis

The NA word order in PL NPs is ungrammatical in DE and thus, even more than for Czech readers, unexpected for German native speakers. Therefore, it might be surprising for German readers of PL NP stimuli to encounter a modifying adjective following a noun, although not impossible to decipher, since they might have encountered this linearisation e.g. in the Romance languages or other Ln. The concrete hypothesis is that the lower predictability of postnominal attributes in PL NPs causes more difficulties for readers with DE as L1 than the same NPs with prenominal modification. We can expect that this greater difficulty will manifest itself in a lower number of correct translations and in longer processing times when PL NPs are presented to German readers in NA than in AN linearisation. Also, it is likely that respondents might fail to recognise the POS of the stimuli in NA linearisation more often than in AN linearisation. Consequently, surprisal should be higher for the NA condition than for the AN condition, and the sums of the surprisal values per NP should have a negative correlation with the percentage of correct translations of the NPs: The less a word is expected to follow upon another one, the less correct answers may be expected.

Vanhove & Berthele showed in intercomprehension experiments with Swiss multilinguals that the combined DE and EN distance, referred to as “Germanic distance” (Vanhove & Berthele, 2015, p. 112), was a better predictor for the intercomprehension of other Germanic languages than a monolingual distance measured towards DE. Hence, I hypothesise that GER distance also is a better predictor in an intercomprehension scenario where native speakers of DE read PL stimuli.

14.7.2. Stimuli

42 NPs were presented in two conditions (AN vs. NA word order) to respondents who are DE native speakers in a web-based free translation experiment. The stimuli NPs in the two conditions were evenly distributed in two

experimental blocks so that there were always 21 NPs with AN and 21 other NPs with NA word order in one block. Every respondent saw an NP in only one of the conditions. 37 respondents took part in the first block and 34 persons took part in the second block. The NPs within a block were automatically presented in random order.

The stimuli NPs were manually compiled from the internationalisms and Indo-European PL-DE cognates among the 100 most frequent PL nouns ($n = 42$) and adjectives ($n = 13$) that were extracted from a corpus-based frequency list (Broda & Piasecki, 2013, see also section 13). Since there were more cognate nouns than cognate adjectives in the list, some of the adjectives occur more often in the stimuli, always in combination with another noun. For all NPs, possible translations were gathered and loaded into the web-based system beforehand in order to provide respondents with positive or negative feedback on their responses. For instance, different correct translations for the stimulus *prywatny szpital* ‘private hospital’ were *Privatkrankenhaus*, *privates Krankenhaus*, *privates Spital*, and *Privatspital*. The orthographic distance of the PL NPs to their orthographically closest DE or EN translations was calculated by means of the Levenshtein algorithm (Levenshtein, 1966), since the sample of respondents can be expected to be fluent in both DE and EN.

14.7.3. Orthographic distance

Among the stimuli NPs, the nouns have a mean orthographic distance of 47% and the adjectives 61% (Jágrová et al., 2016, p. 10). The distances range from 0 for several identical nouns, regardless of capitalization, e.g. *projekt* ‘project’, *punkt* ‘point’, *problem* ‘problem’, *minister* ‘minister’, *firma* ‘company’, *film* ‘film’ up to very distant cognate pairs such as *mężczyzna* – *Mann* ‘man’ (78%) or *tysiąc* – *Tausend* ‘thousand’ (73%). Some of the Indo-European cognate pairs are so distant from each other that their common etymological origin is hardly transparent, for instance *rząd* – *Ordnung* ‘order’ (86%) or *ojciec* – *Vater* ‘father’ (83%). The stimuli NPs have the same orthographic distance in both conditions (AN and NA), but they differ in their surprisal values, depending on the underlying linearisation.

14.7.4. Surprisal in context

Two bigram models which were trained on two different DE corpora were used: FraC, a corpus of sentence fragments (380,000 tokens; Reich & Horch, 2017), and the German Wikipedia Corpus (666.5 mio tokens, Sikos et al., 2017). As a result, surprisal values for each NP per condition and word were determined. For a comparison, the analysis of the FraC corpus was based on lemmas, i.e.

the model did not know any inflected forms and therefore counted all occurrences of lemmas regardless of their inflection. According to the hypothesis, surprisal should be higher for the NA condition than for the AN condition and the sums of the surprisal values per NP should have a negative correlation with the percentage of correct translations of the NPs: The less a word is expected to follow upon another one, the less correct answers may be expected.

With some knowledge of the world, one can say it is plausible that the adjective *neues* ‘new [neut]’ can stand before the noun *Haus* ‘house’. On the contrary, it is rather unexpected that the adjective *begeistert* ‘impressed’ would precede the noun *Haus* ‘house’. This unexpectedness is also reflected in the surprisal values of the two phrases: We obtain a sum of surprisal of 6.79 Hart for *neues Haus* ‘new house’ and 10.72 Hart for *begeistertes Haus* ‘impressed house’.

14.7.5. Results

Only those responses for which both words of an NP were correctly translated were counted as correct. The results of the experiment (2,898 responses in total) were compared for the two conditions (1,449 responses each). Some of the respondents did not finish their experimental block, therefore the number of respondents per NP ranges from 32 to 37. The share of correct responses per NP ranges from 0 to 91.43% in both conditions and differs only to a small but significant extent between the two conditions (29.84% (SD = 1.39%) for AN; 26.81% (SD = 1.16%) for NA). The NPs *amerykański ojciec* ‘American father’, *nowy tysiąc* ‘new thousand’, *możliwy punkt* ‘possible point’, *polski rząd* ‘Polish government’, and *możliwa decyzja* ‘possible decision’ were not translated correctly by any of the respondents in either of both conditions. The rather difficult phrase *francuski mężczyzna* ‘French man’ was translated correctly by two respondents (5.41%) in the AN condition. NPs that consist of basically identical internationalisms and/or adjectives of nationality were translated correctly most often: *nowy projekt* ‘new project’ (91.43% in AN), *polski minister* ‘Polish minister’ (86.11% in AN and 91.43% in NA) and *amerykańska firma* ‘American company’ (88.89% in AN and 88.57% in NA). Table 47 presents a comparison of the results between the two conditions in an overview. The results regarding the individual hypotheses are formulated in the subsections.

	AN (n = 1449)	NA (n = 1449)	All NPs (n = 2898)
Correct	29.84% ²⁰	26.81%	28.36%
Incorrect POS	8	45	53
Mean processing time in s (correct answers)	12.40	10.90	11.65

Table 47: NP translation experiment with German readers: AN vs. NA.

14.7.5.1. Relation between intelligibility and orthographic distance

It was hypothesised that the intelligibility of the stimuli would correlate stronger with a distance measure calculated towards the closest DE or EN cognates (“Germanic distance” Vanhove & Berthele, 2015, p. 112) than to the closest DE cognates. Since only 3 of the 71 respondents indicated no or minimal knowledge of EN, the vast majority of them can be considered DE-EN bilinguals. Besides EN, respondents indicated knowledge of French (n = 22), Spanish (n = 10), Latin (n = 7), Bulgarian (n = 3), Russian (n = 2), Italian (n = 2), Hindu (n = 2), Estonian (n = 2), as well as Croatian, Serbian, Portuguese, Japanese, and Hebrew (n = 1 each).

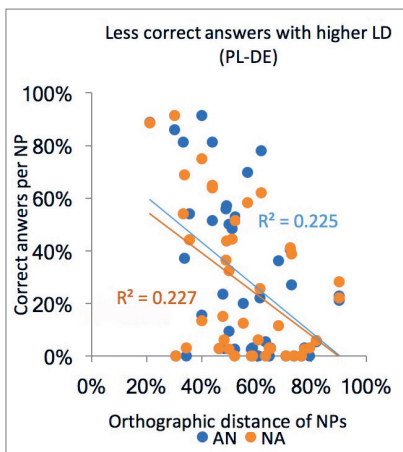


Figure 27: Correlation: correct answers and orthographic distance calculated towards DE.

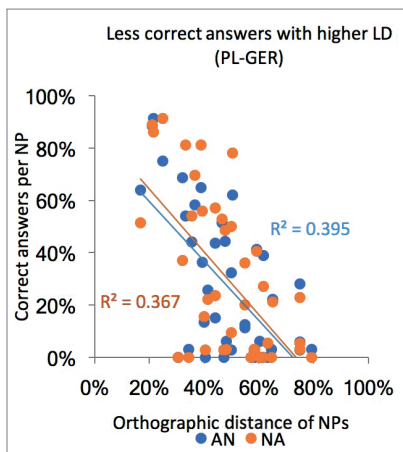


Figure 28: Correlation: correct answers and orthographic distance calculated towards GER.

20 The numbers differ slightly from those published in Jágrová, Stenger & Avgustinova (2017), because the data here are calculated with the means per stimulus NP, while those in Jágrová, Stenger & Avgustinova (2018, p. 26) were calculated on the individual data points.

Figure 27 and Figure 28 show the correlations between the answers for both conditions (AN and NA) and the monolingual orthographic distance between the PL stimuli and their DE translations (Figure 27) vs. the correlations when orthographic distance is calculated towards DE or EN as Germanic distance (GER) (Figure 28). Monolingual orthographic distance can account for only about 23% of the variation in intelligibility in both conditions. As is visible in Figure 28, the shares of correct translations correlate stronger with PL-GER distance ($r = -0.61$, $p < 0.005$ for AN and $r = -0.63$, $p < 0.005$ for NA). The bilingual GER orthographic distance can account for 36% of the variation in intelligibility for the AN and 39% for the NA condition ($R^2 = 0.36$, resp. $R^2 = 0.39$).

14.7.5.2. Relation between intelligibility and surprisal

Regarding the difficulty that results from different linearisation, it was hypothesised that the surprisal values for the NA condition were higher than for the AN condition. This can be confirmed for the data from both corpora: the mean surprisal values from the FraC are 8.76 for AN and 8.96 for NA and from the Wikipedia Corpus 9.07 for AN and 9.82 for NA. The surprisal values from the Wikipedia Corpus have a low correlation with intelligibility ($r = -0.24$, $p < 0.05$) in both conditions – surprisal can account for 6% of the variation in the data ($R^2 = 0.06$). The surprisal values obtained from the FraC correlate on a similar level for the NA condition ($r = -0.24$, $p < 0.05$), but the correlation with intelligibility in the AN condition is close to zero ($r = -0.02$, $p < 0.001$).

When viewing the surprisal values obtained from the Wikipedia corpus and the orthographic distance measures together in a linear regression analysis with intelligibility per NP, we obtain a correlation of $R^2 = 0.466$, $p < 0.01$ for AN and $R^2 = -0.518$, $p < 0.01$ for NA. This means that the factors GER distance and surprisal explain 47% of the variation in the data in the AN condition and 52% in the NA condition. This is more than the factor GER distance alone can explain.

14.7.5.3. Mean processing time

Processing time in ms was measured for every response entered and compared between the two conditions (method cf. Gooskens, 2013, p. 4). This is the time from the moment the stimulus appears till pressing the enter button or clicking continue. The second part of the hypothesis was that the greater difficulty of the NA condition will manifest itself also in the processing times of the NPs. Against our expectations, processing time for AN is on average about 1.5 s higher than that of the NA condition. The analysis of GER orthographic

distance in relation to processing time did not show any significant correlation ($r = 0.105$, $p > 0.05$ for AN and $r = 0.064$, $p > 0.05$ for NA). Also, an analysis of the processing time of all kinds of responses (correct, wrong, nothing entered) in both conditions showed no significant correlation with GER orthographic distance ($r = 0.11$, $p > 0.05$ for AN and $r = 0.125$, $p > 0.05$).

After viewing the stimuli with the lower processing times, e.g. *nowy dom* ‘neues Haus’ with Ø 6.4 s or *nowa noc* ‘neue Nacht’ with Ø 5.4 s, a logical tendency became visible: Processing time depends largely on the length (number of characters) of the DE translation of the stimulus. This suggests that orthographic distance might not be a good predictor of processing time in inter-comprehension experiments that are designed as free translation tasks. The hypothesis that the greater difficulty of the NA condition will manifest itself in higher processing times could not be confirmed. However, one should bear in mind that informants 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.

14.7.5.4. Wrong recognition of POS

As hypothesised, the unexpected word order lead to more misinterpreted POS of stimuli in the NA than in the AN condition: There were 8 instances in AN vs. 45 instances in NA. Table 48 lists all occurrences for wrongly recognised POS in both conditions. The answers are displayed as entered by the participants (Jágrová, Stenger & Avgustinova 2017, p. 30).

AN			NA	
Stimulus PL	Response		Stimulus PL	Response
<i>francuska gmina</i>	<i>Gemeine Franzosen</i>	1	<i>akcja nowa</i>	<i>Alle Jetzt</i>
<i>francuska komisja</i>	<i>Komischer franzo</i>	2		<i>Jetzt handeln</i>
<i>możliwa decyzja</i>	<i>musilma</i>	3	<i>decyzja możliwa</i>	<i>zehn Mäuse</i>
	<i>entscheiden</i>	4		<i>deutsche Muttersprache</i>
	<i>dezente</i>	5		<i>Zehnte (?)</i>
	<i>dezente Mulsima</i>	6		<i>Mozilla Firefox</i>
<i>nowy dzień</i>	<i>heutige Zeh</i>	7	<i>drzwi francuskie</i>	<i>Franzose</i>
<i>specjalna woda</i>	<i>Wasser Spektakel</i>	8		<i>3 franzosen</i>
		9		<i>Ich bin Franzose</i>
		10		<i>schlagende Franzos</i>
		11		<i>zwei franzosen</i>
		12		<i>drei Franzosen</i>
		13		<i>drei Franzosen</i>

AN		NA	
Stimulus PL	Response	Stimulus PL	Response
		14	<i>zehn neue</i>
		15	<i>zehn neue</i>
		16	<i>10 mark</i>
		17	<i>zehn</i>
		18	<i>zehn</i>
		19	<i>Filmnacht</i>
		20	<i>film produktion</i>
		21	<i>einen Film drehen</i>
		22	<i>Filmschauspieler</i>
		23	<i>gemeiner Franzose</i>
		24	<i>deutsche Frau</i>
		25	<i>Informative Veröffentlichung</i>
		26	<i>Informationen publizieren</i>
		27	<i>Informationen Gesellschaft</i>
		28	<i>komischer Franzose</i>
		29	<i>Komische Franzosen</i>
		30	<i>Medezinische F</i>
		31	<i>Mischen</i>
		32	<i>alter Amerikaner</i>
		33	<i>öko strom</i>
		34	<i>ökonomische neue</i>
		36	<i>Öko neu</i>
		37	<i>prozentuales muster</i>
		38	<i>Programm ausstrahlen</i>
		39	<i>merkwürdige Rose</i>
		40	<i>rein privat</i>
		41	<i>reiche Private</i>
		42	<i>Ich bin europäer</i>
		43	<i>Thema funktioniert</i>
		44	<i>Türkisch</i>
		45	<i>türkischer</i>
		46	<i>Ich bin spanier</i>

Table 48: Cases with wrongly recognised POS: AN vs. NA.

When viewing the wrongly recognised POS, we observe that in the NA condition, some adjectives were mistaken for nouns and some nouns were mistaken for adjectives. In many of the cases, the NPs were flipped regarding their POS,

e.g. *reiche Private* ‘rich private people’ was entered as a response for *rzecz prywatna* ‘private matter’. This also happened in the AN condition, but less frequently, e.g. in *Gemeine Franzosen* ‘mean Frenchmen’ for *francuska gmina* ‘French community’ or *dezente Mulsima* (sic!) ‘discreet female muslim’ for *możliwa decyzja* ‘possible decision’. Adjectives were mistaken for nouns, e.g. *europijski* ‘European [A]’ was translated as *Europäer* ‘European [N]’ or *francuski* ‘French [A]’ as *Franzose* ‘French man [N]’.

14.7.5.5. Lexical interferences

Once the noun *dom* ‘house’ was translated correctly, also the accompanying adjectives *nowy* ‘new’ was translated correctly in both conditions, which was not the case with other nouns. The adjectives *nowy*, *nowa* or *nowe* ‘new [masc, fem, neut]’ were also not always translated correctly. Sometimes (7 times in NA and 3 times in AN) they were translated as *jetzt* ‘now’, which is likely due to an interference of EN *now*. Contrary to the advantage of bi- and multilinguals, we can observe a disadvantage here: Interferences can occur not only from the L1, but also from other acquired languages. More examples of such L2/L3 interferences are given in Table 49.

Stimulus PL	Correct DE	Responses (FFs)	Interference DE	Assumed interference from Lns
<i>rzecz</i>	<i>Sache</i> ‘thing’	<i>Reiche</i>		EN <i>rich</i>
<i>teren</i>	<i>Terrain</i> ‘terrain’	<i>Zug</i>		EN <i>train</i>
		<i>Tiere</i>	<i>Tiere</i> ‘animals’	
<i>cel</i>	<i>Ziel</i> ‘aim’	<i>Handy; Telefon</i>		EN <i>cell phone</i>
		<i>Zelle</i>		EN <i>cell</i>
		<i>Himmel</i>		FR <i>ciel</i> ‘sky’
<i>decyzja</i>	<i>Entscheidung</i> ‘decision’	<i>Kinder</i>		BCS <i>djeca</i> ‘children’
<i>pieniądze</i>	<i>Geld</i> ‘money’	<i>Fußgänger; Weg</i>		EN <i>pedestrian</i> /ES <i>pies</i> ‘feet’/FR <i>pieds</i>
		<i>Peanuts</i>		EN <i>peanuts</i>
<i>matka</i>	<i>Mutter</i> ‘mother’	<i>Mathe(matik(er)); Markt; Matte</i>	<i>Mathematik</i> ‘maths’; <i>Markt</i> ‘market’; <i>Matte</i> ‘mat’	
		<i>Angelegenheit</i>		EN <i>matter</i>
<i>nowy / nowa / nowe</i>	<i>neu / neue / neuer</i> ‘new’	<i>Jetzt</i>		EN <i>now</i>
		<i>norwegisch(e)r</i>	<i>Norwegisch</i> ‘Norwegian’	
<i>prezes</i>	<i>Vorsitzender</i> ‘president’	<i>Geschenk</i>	<i>Präsent</i> ‘present’	EN <i>present</i> /BCS <i>prezent</i> ‘present’
		<i>Prinzessin</i>		EN <i>princess</i>
		<i>Prozess; Presse</i>	<i>Prozess</i> ‘process’; <i>Presse</i> ‘press’	EN <i>press</i>

Table 49: Lexical L1/Ln interferences (EN, FR, ES, BCS).

Nevertheless, the correct answers should not be neglected here. For instance, *matka* ‘mother’ was translated correctly more often than there were wrong answers with interferences – see data on the correct responses per NP in Table A 6 of the appendix.

14.7.6. Summary

42 NPs were presented to respondents who are DE native speakers in a web-based experiment with the task to try to translate these phrases. Each NP was presented in one of the two conditions: AN vs. NA linearisation. It was hypothesised that the unexpectedness of the NA linearisation would cause greater difficulties for the German respondents, which would manifest itself in less correct translations, higher processing times and a more frequent wrong recognition of POS in the NA linearisation. Furthermore, it was tested whether surprisal obtained from two DE corpora would correlate with intelligibility of the NPs and whether a distance measured towards the closest DE or EN translations of the stimuli would be a better predictor than when distance was measured only towards the closest DE translations.

The experimental results have shown that the respondents were, on average, more successful in translating NPs with AN linearisation (29.84%) than the same NPs in NA linearisation (26.81%). We can conclude that, generally speaking, AN causes less difficulties for German respondents in a written intercomprehension scenario than NA linearisation. Although the difference between the correct responses in the two conditions is small (3.02%), it is still significant. The fact that almost a third of all responses in both conditions was correct is an argument that learners with DE L1 have good prerequisites as beginners in PL language courses in Germany and that lessons can be held in the target language from the beginning. Besides that, the results suggest that learners can build upon their knowledge not only of other Slavic languages, but also on knowledge of EN or the Romance languages. In this case, familiarity with the grammatical feature of modifying adjectives in the postnominal position, as is common in the Romance languages, seems to help in the present PL-DE intercomprehension scenario. Nevertheless, some cases of lexical Ln interferences from EN, FR and Slavic languages could be observed. These cases were about twice as frequent in the NA condition than in the AN condition. Also, wrong recognition of POS happened about 5 times more often with phrases in NA than with AN linearisation. Nevertheless, both types of errors were rather infrequent: Ln interferences can be confirmed in only 2.00% and wrong recognition of POS in 1.83% of all responses.

The hypothesis that orthographic distance measured between the PL NPs and their closest DE or EN (GER) translations can serve as a better predictor for intelligibility than when the distance was measured only towards the closest DE translations was found to be true. The correlation of Germanic distance with intelligibility has shown that 36% of the variation in AN and 39% of the variation in NA can be explained by PL-GER distance, which is more than PL-DE distance can explain. This goes in line with the findings of Vanhove & Berthele that intercomprehension does not only depend on the L1, but also on the Ln in the multilingual readers' repertoire. When also considering unpredictability (surprisal) as an additional predictor for intelligibility, the factors Germanic distance and surprisal can explain 47% (AN) or 52% (NA) of the variation. Orthographic distance and surprisal, however, turned out to be unsuitable predictors for the processing time of these stimuli. Processing time seems to be directly related to the time respondents take for typing the response, i.e. to the word length of the DE translation.

14.8. Comparison of PL NP Results Between Czech and German Readers

For the experiments described in this section, the hypothesis was that PL NPs in NA linearisation are more difficult to understand for both Czech and German readers than the same NPs in AN linearisation, since postnominal attributes are not as typical (CS) or ungrammatical (DE) in the readers' L1. This difficulty was expected to be reflected in a lower intelligibility and greater processing times in the NA condition. It was also hypothesised that besides linguistic distance, the data would correlate with surprisal scores obtained from language models trained on corpora of the respondents' languages.

Table 50 presents a comparison of the differences between the conditions for the two respondent groups. It has to be kept in mind that the scores cannot be compared directly, since the stimuli sets were different. Here, the differences between the AN and NA condition are of interest. For both respondent groups, intelligibility scores were somewhat higher in the AN condition. The difference between intelligibility scores in the two conditions is statistically significant for the German readers as well as for all data points gathered from the Czech respondents. However, no significance could be found for the NPs with the most representative data set of the Czech respondents (Table 50). As for the differences in processing times, only the mean processing time in the AN condition with the German respondents is slightly higher, which does not confirm the hypothesis. Instead, processing time seems to be directly related to word length (number of characters) of the translation.

\emptyset	Czech respondents		German respondents	
	AN (n = 428)	NA (n = 482)	AN (n = 1149)	NA (n = 1149)
Intelligibility	44.00%	41.31%	29.84%	26.81%
Time spent per NP	10.20 s	10.37 s	12.39 s	10.90 s
R² with LD + surp	0.52	0.58	0.47	0.52

Table 50: Comparison: NP translation experiments with Czech vs. German respondents.

For both Czech and German respondents, it could be confirmed that linguistic distance and surprisal interplay with regard to intelligibility of PL NPs. As for the German respondents, correct answers correlate better with linguistic distance and surprisal than only with linguistic distance in both conditions with a slightly stronger correlation in the NA condition. For the Czech respondents, intelligibility correlates better with linguistic distance and surprisal than with linguistic distance only in the NA condition. This suggests that surprisal as a predictor of intelligibility gains significance for stimuli with rather unexpected word order.

No correlation between processing time and surprisal or processing time and distance was found for the German readers, while for the Czech readers, a low correlation of processing time and total distance (a combined measure of orthographic and lexical distance) was found. Only in the NA condition, a very low but significant correlation could be found for processing time and surprisal with Czech respondents. An observation that speaks for the greater difficulty of the NA linearisation is that both groups of respondents failed to correctly recognise the POS of the stimuli more often in the NA than in the AN condition. The difference between the two conditions regarding wrong recognition of POS is the clearest of all indicators of difficulty analysed here.

Correct answers correlate better with bilingual distance for both respondent groups (CSK or GER) than with monolingual distance. This confirms previous findings that there is an advantage for multilingual readers compared to monolingual readers in intercomprehension and it is an argument for treating adult Czech readers as CSK bilinguals in terms of receptive SK language skills and adult German readers as bilinguals in terms of receptive EN skills at least to a certain extent. Another interesting outcome for both respondent groups is that internationalisms were translated about 3 times more often correctly than other cognates with the same orthographic distance. Although this might not seem surprising, results from the cooperative translation experiment reveal that PL internationalisms that have infrequent CS cognates caused problems with Czech readers when translating sentences (section 5).

CHAPTER VI: TRANSLATION OF TARGET WORDS IN CONTEXT

In previous research on intercomprehension, different types of cloze tests were used as a reliable method for measuring overall text comprehension. In recent studies, individual selected words from the Lx text were placed above the text in alphabetic order and replaced by blanks in the text (cf. Gooskens & Swarte, 2017; van Bezooijen & Gooskens, 2005; Golubović, 2016). Respondents were then asked to put the words back in the text in the correct place. That type of experiment was designed in order to assess the overall comprehension of the text. The present experimental design aims at investigating the intelligibility of **individual** words within sentences. In this section, cloze translation experiments with two different stimuli sets are discussed:

- target words at sentence **final position** in high constraint sentences (proved to be **highly predictable** in monolingual context, source: Block & Baldwin, 2010 – section 15),
- target words in sentences at **random positions** with **random context** (sentences from the cooperative translation experiments (discussed in CHAPTER II) and sentences containing false friends (section 16)).

15. Highly Predictable Target Words in Cloze Translation Task

This section discusses the findings of a cloze translation experiment with PL sentences in which Czech readers were asked to read the entire sentence and translate the highly predictable target word (the last word) in each sentence. Parts of this section were published as a preprint on https://www.coli.uni-saarland.de/%7Etania/ta-pub/CICLing_preprint_Jagrova_Avgustinova_2019.pdf, but some details were corrected in this thesis and might thus differ slightly from those in Jágrová & Avgustinova (2019).

15.1. Experiment Design

The cloze translation experiments were conducted on the experiment website of the INCOMSLAV project (section 7.1). As a baseline, the target word forms from the sentences were also presented without context to other Czech respondents over the same experiment website in order to facilitate a valid comparison of the role of context – see Figure 17 (together with other individual words in the free translation experiment). In the condition without context, the

target words were tested in their base forms, with the exception of nouns that were in plural forms in the sentences – these were also tested in their plural forms without context. This exception was included in order to better represent target words such as *oczy* ‘eyes’ which in their lemma form *oko* ‘eye’ would be identical in PL and CS. Target words with identical base forms in both languages were otherwise not tested in the condition without context. 94 nouns, 14 verbs, 7 adjectives and 3 adverbs were among the target words (total: 118). There were less stimuli in the free translation task than there were sentences in the cloze translation task, because some target words occurred twice in the sentences and some were identical in their base form and therefore were not tested.

15.2. Stimuli

In order to use stimuli with predictive context systematically, sentences from a monolingual cloze probability study by Block & Baldwin (2010) 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. Their study again was based on previous findings of Bloom & Fischler (1980) who provided cloze probabilities for 398 sentences of which 91 turned out to have a high cloze probability, meaning that 67% or more participants provided the same response for a gap in the cloze test (Block & Baldwin, 2010, p. 665). Block & Baldwin extended Bloom & Fischler’s data set of high constraint sentences by adding 398 other constructed sentences. In addition to the cloze experiments, they validated their own dataset as well as Bloom & Fischler’s dataset in psycholinguistic ERP experiments. The study resulted in a new dataset of 400 high-constraint, high cloze probability sentences.

From these 400 sentences, those with the most predictable target words (90%-99% cloze probability) were translated into PL for the present study. Another 22 for which the cloze probabilities were the lowest in the dataset (only 18%-39%) were also translated into PL and added for a possible comparison. A colleague who is a native speaker and professional translator of PL was asked to translate the sentences in such a manner that the target words remain on the last position in the sentences, although a translation variant with another word order might have been more appropriate or more natural in some sentences. The 149 stimuli sentences together with their original EN versions are made available under https://www.coli.uni-saarland.de/%7Etania//ta-pub/CICLing2019_PL_sentences_resource.xlsx. (see also appendix A 4.4.).

Building up on insights from previous research on the role of context in intercomprehension (for instance, Heinz, 2009), the sentences for this study were presented completely in the Lx, i.e. PL. In the original (American) EN

sentence set, there were sentences which contained particular cultural topics and which we assumed to be of no contextual help for readers that are not familiar with the American culture. Such sentences were omitted from the set and not translated into PL. This resulted in a set of 149 sentences. A few translations were modified where it was appropriate, e.g. the original sentence

When Colin saw smoke he called 911 to report a fire.
(Block & Baldwin, 2010)

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.'

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. The translated sentences are published as a resource in the data supplement of Jágrová & Avgustinova (2019) and on www.coli.uni-saarland.de/%7Etania//ta-pub/CICLing2019_PL_sentences_resource.xlsx. They can also be found in Table A 7 in the appendices or in the digital appendices of this thesis.

15.2.1. Closest translation

Linguistic distance and surprisal as predictors of intelligibility were measured for the literal CS translations towards the original PL stimuli as explained in section 8 in detail. These two measures were applied (i) to the whole sentence, (ii) to the final trigram, (iii) to the final bigram, and (iv) to the target word only. All measures were tested as total and normalised values. The closest CS translations are meant to reflect as close as possible how a Czech would read the PL sentence. To score them with an LM trained on the Czech national corpus (CNC, Křen et al., 2015), 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.

If the original PL word was e.g. *przodkach* 'ancestors [loc]', it could not be transformed into the closest possible CS imaginary form **předkách* instead of the translation *předcích*, because this imaginary form does not appear in the corpus. Otherwise, I tried to preserve grammatical forms and phraseological units as close as possible to the PL original, as long as they could be found in the CNC.

Grammatical forms, phraseological units, and prepositions were kept as in the PL original, e.g. *do* 'to' instead of the correct CS *k(e)* in

Poszła do fryzjera, żeby ufarbować włosy.

‘She went to the salon to colour²¹ her hair.’ (cf. Block & Baldwin, 2010)

which was transformed into

**Zašla do kadeřníka, žeby obarvit vlasy.*

for the calculation, instead of a correct CS translation, e.g.

Zašla ke kadeřníkovi / do kadeřnictví, protože si chtěla nechat obarvit vlasy.

Another example would be *genealogiczne drzewo* ‘family tree’ that was transformed into *genealogický strom* ‘genealogical tree’ instead of *rodokmen* ‘family tree’.

Partial cognates, such as *pewny/pewný* ‘stable’, but also ‘sure’ (only in PL) were kept in the sentences and turned into literal translations:

**Był tak pewny, że ten kůň dostihový vyhraje, že зробил сáзку.*

– literally: *‘He was so stable that the racing horse would win that he made a bet’.

A good CS translation would be, for instance,

Był si tak jistý, že ten dostihový kůň vyhraje, že se vsadil.

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.’ (Block & Baldwin, 2010),

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’:

Nemohla koupit košili, protože jí nepasovala.

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 the 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.

21 In the original AE version of the sentence (Block & Baldwin, 2010) it is *color*.

15.2.2. Surprisal

As the trigram 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 ranoek dojąc swoje krowy.

‘The farmer spent²² the morning milking his **cows**.’ (Block & Baldwin, 2010)

or hyponymy such as in

Ellen lubi poezję, malarstwo i inne formy sztuki.

‘Ellen enjoys **poetry, painting**, and other forms of **art**.’ (Block & Baldwin, 2010)

are not expected to be predictable with surprisal obtained from the trigram LMs. Table 51 demonstrates the calculations of surprisal-related predictor variables on a sentence, the final trigram, bigram, and target word. The sentence-final trigram is marked grey.

	Sentence context					w_1	w_2	w_3
PL	<i>dzieci</i>	<i>wyszły</i>	<i>na</i>	<i>dwór</i>	,	<i>żeby</i>	<i>się</i>	<i>bawić</i>
CS cognate	<i>děti</i>	<i>vyšly</i>	<i>na</i>	<i>dvůr</i>	,	<i>žeby</i>	<i>se</i>	<i>bavit</i>
Surp PL	3.81	4.19	0.80	2.00	n/a	1.49	1.29	2.74
Total surp sentence PL	16.32							
Surp trigram PL								5.52
Surp bigram PL								4.03
Δ surp bigram PL								-1.45
Surp lit CS	3.08	4.17	1.28	4.00	n/a	6.06	0.54	4.17
Total surp sentence lit CS	23.30							
Surp trigram lit CS								10.77
Surp bigram lit CS								4.71
Δ surp bigram lit CS								-3.63
Good CS	<i>Děti si šli hrát na dvůr.</i> 'The children went outside to play.' (Block & Baldwin, 2010)							
Note: All surprisal values are given in Hart.								

Table 51: Calculation steps for all surprisal-related variables.

The surprisal scores for commas are not taken into account here, because linguistic distance cannot be assigned to commas in the next step. According to both the PL and the CS LM, the target word *bawić* ‘to play’ is relatively

22 The original source says *spend* (Block & Baldwin, 2010).

unexpected after the reflexive particle *się* ‘oneself’, since there is an increase in surprisal (from 1.29 to 2.74). The differences between the surprisal of the target word itself (w_3) and the word preceding the target word w_2 , indicated by the delta values (Δ *surp bigram*), are calculated as:

$$\Delta \text{ surp bigram} = \text{surp}(w_3) - \text{surp}(w_2).$$

The delta is added as a variable to express the predictability of the target relative to its preceding word, regardless if their individual surprisal values are high or low. Surprisal of the final bigram (*Surp bigram PL / lit CS*) is the sum of the surprisal scores of w_2 and w_3 , surprisal of the final trigram (*Surp trigram PL / lit CS*) is the sum of the surprisal scores of w_1 , w_2 and w_3 .

15.2.3. Linguistic distance

Lexical distance is determined by the number of non-cognates (in the particular sentential context) per sentence and is given as a total and normalised measure per sentence (*NC* and *NC/words* in Table A 15 of the appendix). The *total distance* variables in the appendix (*total dist*) are measures unifying orthographic and lexical distance. This means that if a word pair consists of non-cognates, their total distance is automatically 1, although they might share common features – for instance, a corresponding prefix such as *prze:pře* in *przebrać* ‘change clothes’ and *převléct*. An additional measure *dist* that ignores whether the words are cognates or not is added in the Tables in the appendix – the *dist* value of the pair *przebrać – převléct* would be only 77.3%. A separate variable for the category of false friends was added, since false friends can be both cognates and non-cognates (see section 15.4.2.4).

Orthographic distance was calculated as the CSK to PL pronunciation-based LD (pron LD), i.e. always towards the closest CS or SK translation equivalent under the assumption that the Czech readers have receptive skills in SK (see explanation in section 6.1 and cf. method of Vanhove on “Germanic distance”, (Vanhove, 2014, p. 139)). Table 52 demonstrates the calculations of pron LD in comparison to the traditional orthographic distance (trad LD) on a sentence, the final trigram, bigram, and the target word *bawić* ‘to play’.

	Sentence context					w_1	w_2	w_3
PL	<i>dzieci</i>	<i>wyszły</i>	<i>na</i>	<i>dwór</i>	,	<i>żeby</i>	<i>się</i>	<i>bawić</i>
CS cognate	<i>děti</i>	<i>vyšly</i>	<i>na</i>	<i>dvůr</i>	,	<i>žeby</i>	<i>se</i>	<i>bavit</i>
Trad LD	0.583	0.417	0	0.500	n/a	0.125	0.500	0.400
Pron LD	0.333	0.250	0	0.250	n/a	0	0	0.200
Pron LD bigram			0					0.100
Pron LD trigram			0					0.067
Pron LD sentence								0.148

Table 52: Calculation steps for all orthographic distance-related variables of a sentence.

15.3. Scoring of Responses

Some of the principles for scoring responses in the context-free cognate guessing experiment as described in section 10 were modified during the **scoring of the cloze translation** responses:

- Cognate translations of target words that are only mutual translations in other contexts, but not in the given sentence, were not counted as correct responses. The responses for the condition without context were scored by the same context criterion, accordingly, i.e. only those translations were counted correct without context that were also counted correct in the context condition. This was done for reasons of comparability of the condition with vs. without context. For instance, the target word *broda* in the sentence

Po rozbiciu się okrętu marynarzowi urosła długa broda.

‘While shipwrecked, the sailor grew a long beard.’ (Block & Baldwin, 2010)

means ‘beard’ (CS *vousy*). Without context, it could also mean ‘chin’ (CS *brada*). Only translations of *vousy* or synonyms were accepted as a correct response here.

- **Nominative forms of nouns were accepted** as an alternative to inflected forms. However, forms in other grammatical cases that would change the meaning of the sentence were not accepted as correct. For instance, the response *ryba* ‘fish [nom]’ in the sentence

Przyniósł swoją przynętę nad jezioro, żeby złowić rybę.

‘He brought his bait to the lake to catch fish.’ (Block & Baldwin, 2010)

was accepted as an alternative to *rybu* ‘fish [accu sg]’ or *ryby* ‘fish [accu pl]’, but not *rybě* ‘fish [dat]’, since this would change the meaning in a sense that the subject is fishing something for a fish. In such cases, the fact that only such a grammatical feature was wrong was noted in an extra column for possible later analyses.

- **Wrong gender of verbs was accepted**, e.g. *Neviděl jsem* ‘I haven’t seen [masc]’ for *Nie widziałam* ‘I haven’t seen [fem]’, except if the stimulus was presented in the infinitive (as was the case in the free translation of individual words task) and the translation was provided in an inflected form.
- **Wrong tense was not accepted.**

There were some cases in which a classification of an answer as correct or wrong was not a trivial decision. The target word *zespół* ‘group’ occurs in two of the stimulus sentences in which it has different, but semantically related meanings:

Lubiła grać na gitarze, więc dołączyła do zespołu.

‘She loved playing the guitar so she joined the **band**.’ (Block & Baldwin, 2010)

and

Dana poproszono, aby został nowym coachem zespołu.

‘Dan was asked to be the new coach of the **team**.’ (Block & Baldwin, 2010)

It can be a cognate to the CS word *spolek* ‘association, union’ in contexts other than those presented. For the two different stimulus sentences, only those responses were counted as correct that are correct in the respective context.

15.4. Results

15.4.1. Comparison: with vs. without context

This subsection compares the target words in terms of how frequently they were translated correctly by the Czech respondents in the conditions with vs. without context. I pinpoint some of the prominent difficulties in a subsequent error-analytical section (15.4.3).

The intelligibility scores vary with different categories of target words in both conditions, i.e. with and without context. As mentioned before, among the target words in the sentences with high cloze probability, there turned out to be cognates, non-cognates, and false friends. Some of the target words were non-cognates in the particular context presented, but they can be considered cognates in another context. Some target words could be classified as false friends only after an analysis of the responses from the experiments. An analysis of the target words resulted in a classification scheme as shown in Table 53 which also serves as an explanation for the colour code used in Figure 29.

Category	n	Description	Example
C	C-IB 11	Real cognate with an identical base form that differs only in inflected forms.	The cognate <i>ryba</i> ‘fish’ is identical in its base forms in both Ls. The PL target <i>rybę</i> ‘fish [accu]’ differs from its CS corresponding form <i>rybu</i> ‘fish [accu]’.
	C-C 89	Translation equivalent is a real cognate correspondent that differs in orthography and can differ in morphology and phonetics, too.	PL <i>głos</i> vs. CS <i>hlas</i> —both ‘voice’

Category		n	Description	Example
C	C-OC	3	Non-cognate in the presented context, but cognate in at least one other context .	PL <i>szczotka</i> 'brush, broom' can correspond to <i>štětka</i> 'brush' only in some contexts, e.g. as a brush for shaving, but not a broom (correct CS <i>smeták</i>) for sweeping the floor.
			NC	7
NC	5	Non-cognate , not expected to be intelligible for the reader without context, incidental or Ln knowledge.		
FF	FF-C	14	Cognate that is also a false friend , i.e. frequently mistaken for another more similar CS word.	PL <i>znaczek</i> 'stamp' is frequently mistaken for <i>značka</i> 'sign', while the correct cognate translation would be <i>známka</i> 'stamp'.
	FF-OC	9	Cognate translations in other contexts that are frequently mistaken for another more similar CS word and therefore are false friends.	PL <i>zdanie</i> 'opinion' was frequently mistaken for CS <i>zdání</i> 'appearance' instead of the correct translation <i>názor</i> 'opinion' in the particular sentence. These cognates can be mutual translations, e.g. in <i>Mam zdanie, že ...</i> and <i>Mám zdání, že ...</i> 'It seems to me as if ...'.
	FF-A	6	Non-cognate frequently mistaken for another more similar CS word (FF), but allows for associations with a correct translation.	PL <i>drzewo</i> 'tree' is frequently mistaken for CS <i>dřevo</i> 'wood', which at the same time can provide a correct association in the respective context.
	FF	5	Non-cognate frequently mistaken for another more similar CS word (FF).	PL <i>gwóźdź</i> 'nail' is frequently mistaken for CS <i>hvozď</i> 'forest', while the correct translation would be <i>hřebík</i> .

Table 53: Overview of (sub-)categories of target words included into the statistical model.

The full list of the stimuli sentences (including target words) together with their categories according to those shown in Table 53 can be found in Table A 7 in the appendix.

Overall, the mean intelligibility of target words improved significantly from 49.71% without context to 67.99% in highly predictive contexts ($t(295) = 4.39$, $p < 0.001$). This means that the hypothesis that 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. Figure 29 contains a trend line at $f(x) = 1x$ 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.

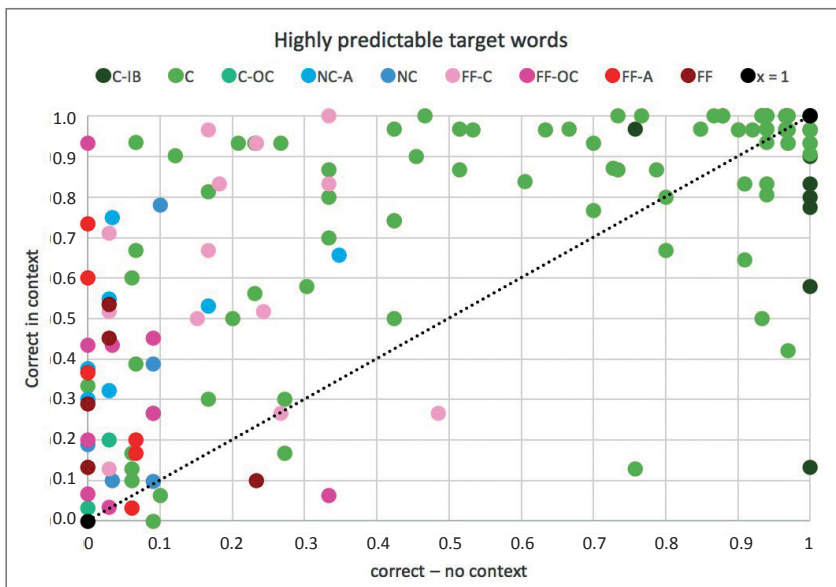


Figure 29: Intelligibility of target words with vs. without context.

In the condition with context, an intelligibility score of 100% could be observed for 26 target words, and 18 other target words were correctly translated by 96.67% of the respondents. In the condition without context, there were only 19 target words with an intelligibility score of 100% and 11 with $\geq 96.67\%$.

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’ (Block & Baldwin, 2010).

When presented in this sentence, 90% translated the PL target *pierścionek* ‘ring’ correctly, while in the condition without context, only 45.5% entered the correct CS cognate *prstýnek*. Both the CS and the PL trigram LM confirms that the target *pierścionek* ‘ring’ is highly predictable after *diamentowy* ‘diamond [A]’ (Block & Baldwin, 2010), which is indicated by the dropping surprisal curve after *diamentowy* in Figure 30 (red graph – surprisal from the PL LM; green graph – surprisal from the CS LM).

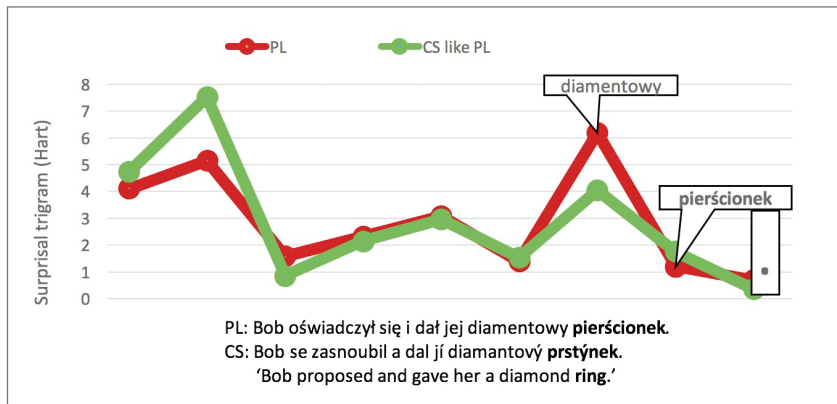


Figure 30: Surprisal graph of a sentence with a low-surprisal target word.

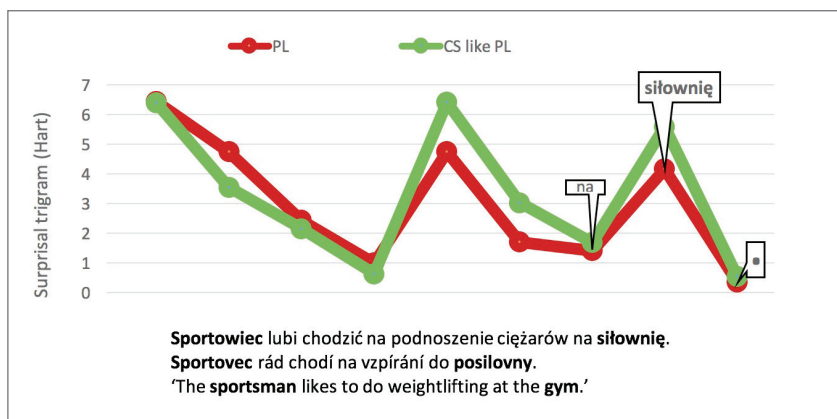


Figure 31: Surprisal graph of a sentence with a high-surprisal target word.

In contrast to the sentence in Figure 30, there is an increase in surprisal in Figure 31 at the target *siłownię* 'gym [accu]' for the sentence

Sportowiec lubi chodzić na podnoszenie ciężarów na siłownię.
 'The sportsman likes to do weightlifting at the gym.'²³

In the monolingual cloze completion task (Block & Baldwin, 2010), 95% of English native speakers provided the response *gym*, which suggests that the

23 The original EN version is 'The athlete is enjoying lifting weights at the gym.' (Block & Baldwin, 2010).

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 [accu]’ with the sentence-initial *sportowiec* ‘athlete, sportsman’ rather than with its directly preceding words *ciężarów na* ‘weights [gen pl] at’.

For a more detailed analysis and with respect to the different lexical and cross-lingual properties of the stimuli, the target words are categorised by their lexical characteristics in an overview with examples and separate graphs in subsection 15.4.2. Figure 29 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 15.4.2.4). The effect of the predictable context seems to be especially striking with such not clear-cut cases of false friends. Consequently, the first question to pose was how to define false friends and how to distinguish them from regular non-cognates for a reliable statistical model. This question is discussed in detail in subsection 15.4.2.4. Consequently, additional variables concerning the lexical relations were added to the statistical model that is presented in section 15.5.

Surprisal might also explain our perception of humour. If something is rather unexpected, it usually can also be amusing. While going through the responses of the cloze experiments with high-constraint sentences, there were some cases that made me laugh:

i. *Poszła do fryzjera, żeby ufarbować włosy.*

‘She went to the salon to colour her hair.’ (Block & Baldwin, 2010)

The correct CS translation of *włosy* should have been *vlasý* ‘hair’, which 93% of the Czech respondents translated correctly. Only 1 person had the idea to translate it as *vosý* ‘wasps’, resulting in ‘She went to the salon to colour her wasps’.

ii. *Ponieważ błyskało, nie mogła iść na basen pływać.*

‘Because there was a lightning she could not go to the pool to swim.’ (Block & Baldwin, 2010)

Again, the target word *pływać* ‘swim’ was translated correctly as *plavat* by 90% of the Czech respondents. Only one respondent apparently thought it is more likely to ‘go to the pool to cry’ by responding *plakat* ‘cry’. However, one should contemplate that this response might as well simply be a typo.

15.4.2. Different lexical categories of target words

A one-tailed t-test of independent samples is performed in order to analyse whether the differences between the conditions with vs. without context are significant.

Category	No context	Context	t-test	Significance
C-IB	94.50%	81.40%	$t(20) = -1.50$	<i>ns</i>
C-C	65.90%	80.10%	$t(176) = 3.05$	$p < 0.01$
C-OC	4.00%	16.60%	$t(4) = 1.67$	<i>ns</i>
NC-A	8.70%	49.80%	$t(12) = 5.07$	$p < 0.001$
NC	6.30%	31.10%	$t(8) = 1.90$	$p < 0.05$
FF-C	20.61%	64.79%	$t(26) = 5.24$	$p < 0.001$
FF-OC	2.73%	32.01%	$t(16) = 3.06$	$p < 0.01$
FF-A	3.23%	34.98%	$t(10) = 2.85$	$p < 0.01$
FF	30.17%	5.88%	$t(8) = -2.53$	<i>ns</i>

Table 54: Intelligibility of target words with vs. without context in the different categories.

As shown in Table 54, the differences between the intelligibility of target words with vs. without context are significant for all categories except for C-IB, C-OC, and FF, which for the latter two are most likely due to the low number of these items.²⁴ Also, the frequency of the ceiling effect (maximum scores in both conditions) in C-IB could be a reason for the insignificance of the differences. The greatest and highly significant difference between the two conditions was found for target words that are FF-C. Interesting examples from these categories are shown as follows. The distinction between words that allow for associations and those that actually do have real cognates also seems to play a role in the results.

24 The values may slightly differ from the ones published in Jágrová & Avgustinova (2019) because of their subsequent correction, mainly regarding the false friends' subcategories, in this thesis.

15.4.2.1. Cognates (C)

Cognates can be identical in their base form in both languages (C-IB), but not in the inflected forms as presented in the context. This also applies to such target words that can be cognates only in a certain context, for instance *punkt* ‘point’ which can be *punkt* in CS only in some contexts (cf. Vavřín & Rosen, 2015). Target words that are identical in their base forms in both languages were not tested in the condition without context and an intelligibility of 100% was entered for comparison. Some of them that only stand in a different grammatical case in the stimulus sentence than in CS seem to be easily identifiable as their CS equivalents and thus cause ceiling effects. For instance, *roku* ‘year’ in

Wiosna była Jo ulubioną porą roku.

‘Spring was Jo’s favorite season of the year.’ (Block & Baldwin, 2010)

was successfully translated by 100% of the Czech respondents, although the equivalent CS construction of this phrase would not be a modification by the genitive, but an adjectival premodification – *roční období* ‘season of the year’. However, this success might as well be caused by the identical genitive (also dat, loc, and voc) form of *rok* ‘year’ in CS.

Others were not translated correctly by 100% of the respondents when presented in context. This might be due to the morphological distance and different or unknown inflectional endings of the forms, e.g. PL *punktów* vs. CS *punktů* ‘points’ (PL *testamencie* vs. CS *testamentě/testamentu* ‘testament [loc]’), especially endings of feminine target words in the accusative case, e.g. PL *rybę* vs. CS *rybu* ‘fish [accu]’.

Cognates (C) differ in orthography and can additionally differ in morphological and phonetic features. Ceiling effects can also be observed with target words with very small orthographic distance that differ only in diacritics and thus were translated correctly by all respondents, such as PL *mokry* and CS *mokrý* ‘wet’ in

Potrzebowałbyś płaszcz przeciwdeszczowego, żebyś nie był mokry.

‘You would need a raincoat to avoid getting wet’ (Block & Baldwin, 2010).

The same applies to target words with easily identifiable pronunciation, for instance, in *czasu* – CS *času* – ‘time [gen]’ that was translated correctly by 96.8% in

Jej praca była łatwa większą część czasu.

‘Her job was easy most of the time.’ (Block & Baldwin, 2010).

Interestingly, there are target words with a relatively high LD, e.g., PL *obiad* ‘lunch’ with a LD of 40% to the 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.’ (Block & Baldwin, 2010).

When the category of cognates is viewed separately, their intelligibility correlates significantly with the pron LD of the target word ($r = 0.549, p < 0.001$), but no significant correlation with surprisal could be observed ($r = 0.043, p < 1$). For better visibility than in Figure 29, Figure 32 provides a separate overview over the results for cognates with vs. without context.

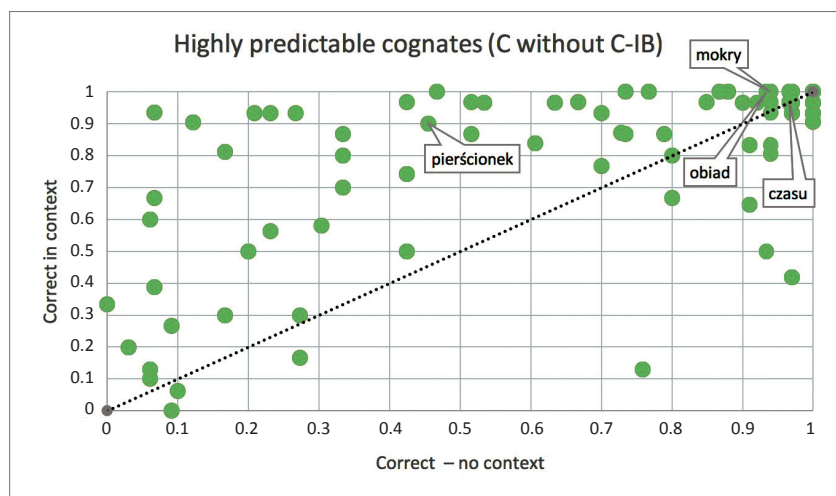


Figure 32: Comparison: target cognates (C without C-IB) with vs. without context.

On one hand, context seems to be helpful in understanding target cognates. On the other, the target cognates were in base forms in the context-free condition. Those words that have a lower distance in their base forms than in their non-base forms were more often translated correctly without context. This also confirms the recent results of a monolingual benchmark study of eye movements in a with RU sentences by Sekerina et al. (2018). There it was found that “mean fixation durations were higher for the non-base-form words” (Sekerina et al., 2018, p. 15), which suggests that a higher cognitive effort is needed for processing non-base forms than for base forms. However, in the present study, this does not apply to all POS. No greater success of translating the base forms of **verbs** could be observed, since a morphological interference seems to have prevailed with the PL infinitive ending *-ć* which was frequently mistaken for a nominal ending (see section 15.4.3.3).

15.4.2.2. Cognates in other contexts (C-OC)

Within the category of cognates, words can be considered cognates because they can be mutual translations in a particular context, but are no cognates in the context presented. In the present set of stimuli, this was the case for target words in only three of the sentences: One was *szczotką* ‘broom [instr]’ in

John zamiótł podłogę szczotką.

‘John swept the floor with a broom.’ (Block & Baldwin, 2010),

which would be translated as *smeták* or *koště* in CS. A cognate to PL *szczotka* would be CS *štětka*. These two words, however, can be mutual translations only in a context in which *szczotka* signifies a *brush* (*paint brush, shaving brush, toilet brush* etc.), but not a *broom*.

In the other two cases, the target word *zespołu* [gen of *zespół*] could be translated either as *kapely* ‘band [gen]’ in

Lubila grać na gitarze, więc dołączyła do zespołu.

She loved playing the guitar so she joined the band. (Block & Baldwin, 2010)

or as *týmu* ‘team [gen]’ in

Dana poproszono, aby został nowym coachem zespołu.

Dan was asked to be the new coach of the team. (Block & Baldwin, 2010)

A possible CS cognate translation of *zespół* could be *spolek* in another context with the meaning of *club, association*. The results for these three cases are displayed together with the non-cognates in Figure 33.

15.4.2.3. Non-cognates (NC)

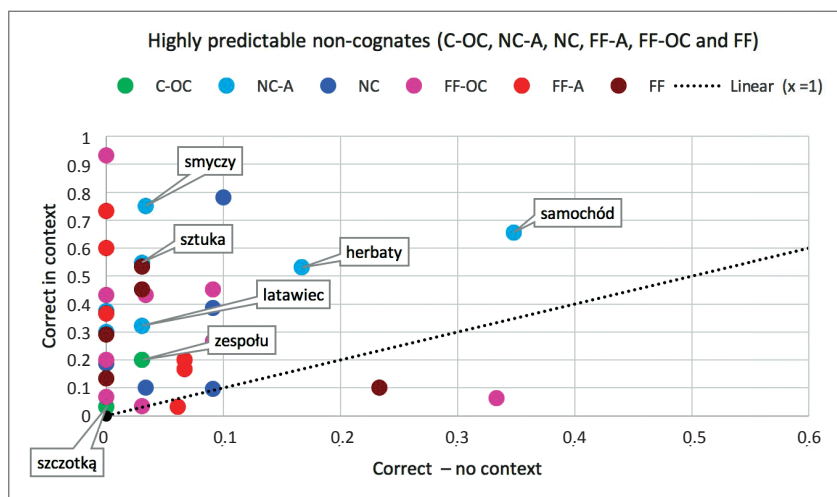


Figure 33: Comparison: target non-cognates (incl. C-OC, FF-OC, FF-A, and FF) with vs. without context.

The CS translations of these target words are not cognates of the PL targets, but they do share some common features. Respondents might associate the stimulus with a concept in their L1 and then come up with the correct translation in context. For instance, PL *latawiec* ‘kite’ might be associated with the CS verb *létat* ‘to fly’ or the concept of flying in general and then lead to the correct CS translation *drak* (which besides ‘kite’ can also mean ‘dragon’). The data points in Figure 33 reveal that associations, which non-cognates can provide, lead to an even greater improvement in intelligibility in context than if the target words are cognates in another context. Table 55 lists the possible associations that some of the non-cognates from Figure 33 might evoke:

Target	Possible association	Explanation
<i>smycz</i> 'leash'	<i>smyč</i> 'sling'	There can be a sling at the end of a leash, e.g. instead of a collar.
<i>samochód</i> 'car'	<i>samo-</i> 'self' + <i>chod</i> 'walk'	A car is a means of transport, so people do not have to walk.
<i>herbata</i> 'tea'	EN <i>herb</i>	Tea can be made out of herbs (only applies to readers with knowledge of EN).
<i>biurko</i> 'desk'	<i>byro</i> 'office', EN/FR <i>bureau</i> , DE <i>Büro</i> 'office'	Every office usually has at least one desk (only applies to readers with knowledge of these foreign words, resp. the CS loan word <i>byro</i> 'office').
<i>plusk</i> 'splash'	<i>plesk</i> 'smack'	Similar onomatopoeic words.
<i>latawiec</i> 'kite'	<i>létat</i> 'to fly', masc nominal suffix <i>-ec</i>	A kite is a flying object.
<i>sztuka</i> 'art'	<i>štuka</i> 'stucco'	This architectural element is also considered art.

Table 55: Overview of target non-cognates that offer associations with correct CS translations.

15.4.2.4. False friends (FFs)

Regardless of what linguists define as false friends, respondents might perceive certain stimuli differently than expected by the experimenters. Target cognates that were expected to be easy to guess might turn out to be false friends. On the other hand, some words that were expected to be false friends were translated incorrectly in various ways without a particularly prominent wrong response. For example, *znaczek* – CS *známka* 'stamp' – was frequently mistaken for *znak* or *značka* 'sign' (93.9% wrong) when presented without context. In the predictive context of

Wysłał list bez znaczka.

'He posted the letter without a stamp.' (Block & Baldwin, 2010),

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.

For what is considered false friends in the following analysis, I define the term as follows: As a threshold, the percentage of the particular wrong type of response must have been higher than the sum of no responses and correct responses and a particular wrong response must have been more frequent than the sum of all other wrong responses. In addition, the share of a particular

wrong answer and the missing answers must have been higher than the variation of responses in order to consider it as a false friend in the present study.

In order to calculate the variation of the responses, all different types of responses were counted and divided by the number of respondents for this particular stimulus, i.e. the variation of answers is the number of different answers per stimulus divided by the total number of responses. Based on the criteria for scoring of the responses described in section 15.3, the following principles were applied in order to determine the variation of the responses:

- Inflected forms and synonyms were counted as one type of response, for instance *droga*, *drogy* and *lék* ‘drug(s)’ as a wrong response to the PL stimulus *droga* ‘road’.
- Obvious typos were not counted as another response, but as the word that the respondent evidently meant to write.

If all respondents entered the same response and only one respondent entered a different one, the variation of this stimulus would be $2/30 = 0.067$, because there were exactly two types of answers. A high variation can be due to high neighbourhood density (number of available options with minimal differences). The higher the variation, the more difficult the stimulus should be, i.e. the less correct responses can be expected. The variance values are indicated in Table A 8 – Table A 11 in the appendix for all types of false friends.

Of course, such an existing cognate can be a cognate in the context presented or in another context. Consequently, a distinction has to be made between those false friends for which no other correct cognate exists in the reader’s language and such for which it exists. Therefore, the false friends among the target words were classified according to the four categories shown in Table 53.

When all four categories of false friends are counted together, they are on average translated correctly by 11.62% when presented without context and by 45.76% when presented in sentential context. The differences are significant at the 1% level ($t(35) = 6.77$). Figure 34 displays the results for the different subcategories of false friends.

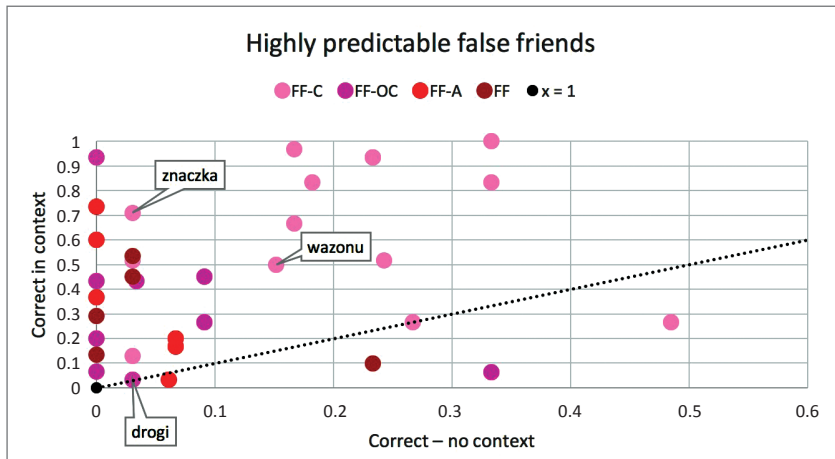


Figure 34: Comparison: target false friends (FF-C, FF-OC, FF-A, and FF) with vs. without context.

In Table 56-59, all target words in the four subcategories of false friends are listed and information on how frequently they were mistaken for which word is provided. If a target word appeared in two sentences, the respective values appear in connected cells in the tables.

PL target word	Inflected form if different from base	Correct CS cognate	FF	Mistaken (%)	
				No context	In context
<i>znaczek</i> 'stamp'	<i>znaczką</i>	<i>známka</i>	<i>značka</i> 'sign'	69.70	12.90
<i>plywać</i> 'to swim'		<i>plavat</i>	<i>plivat</i> 'to spit'	51.52	0.00
<i>królowa</i> 'queen'	<i>królową</i> [accu, instr]	<i>královna</i> (<i>královnou</i>)	<i>králova</i> 'the king's'	24.24	40.00
<i>czek</i> 'cheque'		<i>šek</i>	<i>čech</i> 'Czech person'	43.33	0.00
<i>wazon</i> 'vase'	<i>wazonu</i> [gen]	<i>váza/vázy</i>	<i>vagon</i> 'wagon'	48.48	36.67
<i>niebo</i> 'sky'		<i>nebe</i>	<i>nebo</i> 'or'	73.33	0.00
	<i>niebie</i> [loc]			0.00	
<i>nić</i> 'thread'	<i>nici</i> [gen]	<i>nit/nítě</i>	<i>nic</i> 'nothing'	90.91	6.45
<i>nocą</i> 'at night'		<i>v noci/noci</i>	<i>noc</i> 'night'	81.82	76.67
<i>poczta</i> 'post (office)'	<i>pocztą</i> [instr]	<i>pošta</i>	<i>poceta</i> 'honour'	63.64	38.71
<i>ręka</i> 'hand'	<i>rękę</i> [accu]	<i>ruka/ruce</i>	<i>řeka</i> 'river'	90.91	64.52
<i>kwitnąć</i> 'to bloom'	<i>kwitną</i>	<i>kvést/kvetou</i>	<i>květináč</i> 'flowerpot'	73.33	0.00
<i>liście</i> 'leaves'	<i>liści</i> [gen pl]	<i>listí/listí</i>	<i>liška</i> 'fox'	36.67	43.00
<i>miecz</i> 'sword'		<i>meč</i>	<i>míč</i> 'ball'	40.00	13.33
<i>ściana</i> 'wall'	<i>ścianie</i> [loc]	<i>stěna/stěně</i>	<i>scéna</i> 'scene'	50.00	0.00

Table 56: Comparative overview of FF-C with vs. without context.

PL target word	Frequently mistaken for	Mistaken (%)		Cognate in other context and association	Response was cognate (%)		Intelligibility (%)	
		No context	In context		No context	In context	No context	In context
<i>broda</i> 'beard'	<i>brod</i> 'ford'	0.00	36.67	<i>brada</i> 'chin'—a beard grows on a chin	33.33	46.88	0.00	6.25
<i>ciasto</i> 'cake'	<i>často</i> 'often'	73.33	6.67	<i>těsto</i> 'dough'	6.66	50.00		
<i>droga</i> 'road'	<i>droga</i> 'drug', <i>lék</i> 'medicine'	78.79	53.33	<i>dráha</i> 'track, lane'	0.00	0.00	3.03	3.33
			0.00			13.33		20.00
<i>ogrzewanie</i> 'heating'	<i>ohřívání</i> 'heating up [N]'	63.63	29.03	<i>ohřívání</i> 'heating up [N]'	63.63	29.03	9.09	45.16
			33.33			33.33		26.67
<i>przebrać</i> 'to change clothes'	<i>přebrać</i> 'to pick over'	60.61	16.67	<i>přebrać</i> 'to pick over'	60.61	16.67	0.00	6.67
<i>wiadomości</i> 'news'	<i>vědomost(i)</i> 'knowledge'	75.76	87.10	<i>vědomost(i)</i> 'knowledge'	75.76	87.10	6.06	3.23
<i>zdanie</i> 'opinion'	<i>zdání</i> 'appearance'	56.67	10.00	PL <i>Mam zdanie, że ...</i> and CS <i>Mám zdání, že ...</i> 'It seems to me as if ...'	56.67	6.67	3.33	43.33
<i>zmiana</i> 'shift'	<i>změna</i> 'change'	83.33	0.00	<i>změna</i> 'change'	83.33	0.00	0.00	93.33

Table 57: Comparative overview of FF-OC with vs. without context.

For instance, the word *spodnie* 'trousers' turned out to be a false friend, as 80.0% of the Czech readers mistook it for *spodky* 'underpants' when presented without context. The immense improvement in correct responses from 0% without context to 60.0% when presented in the context

Cid potrzebował paska, żeby przytrzymać swoje spodnie.

'Cid needed a belt to hold up his pants.' (Block & Baldwin, 2010)

could be explained by the association of trousers as underpants as both belong to the category of clothing for the legs. Table 58 gives an overview of the words that are likely to provide associations helpful for the correct translation.

PL target word	Frequently mistaken for	Association	Mistaken (%)	
			No context	In context
<i>spodnie</i> 'trousers'	<i>spodky</i> 'underpants'	Underpants are a subcategory of trousers, <i>spodní</i> 'under' [A].	80.00	6.67
<i>drzewo</i> 'tree'	<i>dřevo</i> 'wood'	Wood as a material is produced from trees.	96.97	56.67 15.00
<i>miłość</i> 'love'	<i>milost</i> 'mercy'	<i>milý</i> 'dear' [A]	83.33	60.00
<i>doniesienie</i> 'message'	<i>donesení</i> 'bringing'	Messages can be brought.	43.33	6.67

Table 58: Comparative overview of FF-A with vs. without context.

PL target word	Inflected form if different from base	Frequently mistaken for	Mistaken (%)	
			No context	In context
<i>gwóźdź</i> 'nail'	<i>gwóździa</i> [gen]	<i>hvozd</i> 'forest'	45.45	0.00
<i>nastrój</i> 'mood'	<i>nastroju</i> [gen, loc]	<i>nástroj</i> 'instrument'	96.97	38.71
<i>stroić</i> 'to tune'	<i>nie stroiło</i>	<i>stroj</i> 'machine'	53.33	3.33
<i>wyznaczony</i> 'appointed'	<i>wyznaczonym</i> [instr, loc]	<i>vyznačený</i> 'characterized'	54.55	48.39
<i>zakład</i> 'bet'		<i>základ</i> 'base'	100.00	70.00

Table 59: Comparative overview of false friends with vs. without context.

Consequently, when the results should be analysed in a multiple linear regression model, it is necessary to include separate lexical variables into the model, i.e. to have a separate column for the criterion cognateness (y/n) and false friend (y/n) (section 15.5.), as not all false friends belong to the category of cognates, while others do.

15.4.3. 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:

15.4.3.1. Differences in government pattern

In some sentences, the target words seem to have been more difficult, probably because of the 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.’ (Block & Baldwin, 2010).

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 *žízeň* ‘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.

15.4.3.2. Ln interferences

Effects of another language (Ln) interference occurred 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 [gen]’ was translated as *skla* ‘glass [gen]’ by another respondent living in Great Britain. One respondent translated *biurku* ‘desk [loc]’ 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 [loc]’ in the sentence

Jenny zapalila świeczki na urodzinowym torcie.

‘Jenny lit the candles on the birthday cake.’ (Block & Baldwin, 2010),

was translated as *svícnu* ‘candlestick [gen/dat/loc]’ 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’. Except for this last example, none of the wrong responses due to interferences would fit the context of the sentence better than the correct translations, so that they cannot be expected to be a context-driven decision.

15.4.3.3. (Perceived) morphological mismatches

- **PL feminine accusative nouns ending in -e**

PL feminine nouns ending in *-e* were frequently wrongly translated with words ending in *-e*, *-é* or *-ě* or with plural forms. For instance, *swójq rolę* ‘her role [accu]’ was translated as *role* [nom sg or nom/accu pl] when the correct equivalent would have been *roli* [accu] in CS. Nevertheless, *role* was counted as a correct answer since 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.’ (Block & Baldwin, 2010),

with words ending with an *-e*, *-é* or *-ě*: *přírodě* [dat of *příroda* ‘nature’], *tance* ‘dances’, *hřiště* ‘sport field, playground’, *sondě* [dat of *sonda* ‘sond’], *laně* [loc of *lano* ‘rope’], *poprvé* ‘for the first time’, *zkoušce* [dat of *zkouška* ‘test, rehearsal’] for which the correct CS translation would have been *zkoušku* [accu of *zkouška*]. The target word *próbę* could be correctly identified as *rehearsal* through the DE cognate *Probe* ‘rehearsal, specimen’ by those subjects who knew DE. Table 60 provides an overview of the target words ending in *-e* and the frequencies of correct accusative forms vs. wrong plural forms. If no frequency is indicated behind a response, then the response was given only once. Not all replies are given in the column *other replies*.

Target word in sentence	Correct accu singular CS (frequency)	CS plural (frequency)	Other replies (frequency)
<i>rolę</i> ‘role’ [accu of <i>rola</i>]	<i>roli</i> (0.5)	<i>role</i> (0.43)	<i>dílo</i> ‘work, piece of art’ (0.03), <i>postavení</i> ‘position’ (0.03)
<i>próbę</i> ‘test, try, rehearsal’ [accu of <i>próba</i>]	<i>zkoušku</i> (0.33)	<i>zkoušky</i> (0)	<i>zkoušce</i> ‘test, try, rehearsal [dat/loc]’, <i>přírodě</i> ‘nature [dat/loc]’, <i>tance</i> ‘dance [gen sg, nom/accu pl]’, <i>hřiště</i> ‘playground’, <i>sondě</i> ‘sond [dat/loc]’, <i>laně</i> ‘rope [loc]’, <i>poprvé</i> ‘for the first time’ (total 0.23)
<i>książkę</i> ‘book’ [accu of <i>książka</i>]	<i>knižku, knihu</i> (0.37)	<i>knižky, knihy</i> (0.5)	<i>knižka</i> (0.03)
<i>patelnię</i> ‘pan’ [accu of <i>patelnia</i>]	<i>pánev</i> (0.09)	<i>pánve</i> (0.06)	<i>rohliky</i> ‘croissants’, <i>buchty</i> ‘pastries’, <i>hřebíky</i> ‘nails’ (total plurals 0.09), <i>těsto</i> ‘dough’ (0.31)
<i>rękę</i> ‘hand’ [accu of <i>ręka</i>]	<i>ruku</i> (0.1)	<i>ruce</i> (0.03)	<i>řece</i> ‘river [dat/loc]’ (0.35), <i>řeku</i> ‘river [accu]’ (0.16), <i>řeky</i> ‘rivers’ (0.1), <i>řekne</i> ‘he/she/it will say’, <i>řeči</i> ‘languages/comments’, <i>velké</i> ‘big [pl]’

Target word in sentence	Correct accu singular CS (frequency)	CS plural (frequency)	Other replies (frequency)
<i>różę</i> 'rose' [accu of <i>róża</i>]—high predictability	<i>růži</i> (0.4)	<i>růže</i> (0.57)	n/a
<i>różę</i> 'rose' [accu of <i>róża</i>]—low predictability	<i>růži</i> (0.06)	<i>růže</i> (0.77)	<i>lůže</i> 'lodge/box' (0.06), <i>ryže</i> 'rice'
<i>trasę</i> 'road' [accu of <i>trasa</i>]	<i>trasu</i> , <i>cestu</i> 'road [accu]' (0.07)	<i>trasy</i> (0)	<i>trase</i> [dat/loc] (0.33), <i>cestě</i> 'road [dat/loc]' (0.23), <i>turné</i> 'tour' (0.1)
<i>rybę</i> 'fish' [accu of <i>ryba</i>]	<i>rybu</i> (0.03)	<i>ryby</i> (0.68)	<i>rybě</i> 'fish [dat]' (0.16)
<i>siłownię</i> 'gym' [accu of <i>siłownia</i>]	<i>posilovnu</i> (0.03)	<i>posilovny</i> (0.03)	<i>posilovně</i> 'gym [dat/loc]' (0.26), <i>tělocvičně</i> 'gym [dat/loc]' (0.06), <i>svaly</i> 'muscles' (0.06), <i>podnose</i> 'tray [loc]' (0.03)
Mean	19.8%	30.7%	

Table 60: PL *-ę* mistaken for a plural marker or a marker of other grammatical forms ending with *-e* or *-ě* in CS.

On the average, the frequency of responses in plural forms of the actual correct CS translation (30.7%) is about 10% higher than the frequency of correct responses in the accusative case (19.8%).

• PL feminine instrumental nouns ending in *-ą*

The PL instrumental ending of feminine nouns *-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', *szcnotką* as *šotka* 'Scottish woman', *pocztą* as *pocsta* 'honour' – see Table 61. Czech readers are rather unlikely to identify the PL ending *-ą* as an instrumental marker similar to the CS *-ou*.

Target word in instrumental case	Correct CS	%	CS nominative	%	Selected responses	%
<i>królową</i> 'queen'	<i>královnou</i>	3.33	<i>královna</i>	23.33	<i>králova</i> 'the king's'	43.33
<i>szczotką</i> 'broom'	<i>smetákem</i>	3.13	<i>smeták</i>	0	<i>šotka</i> 'Scottish woman'	12.50
<i> pocztą</i> 'post'	<i>poštou</i>	9.68	<i>pošta</i>	38.71	<i>pocta</i> 'honour'	32.26
					<i>poctu</i> 'honour [accu]'	6.45
<i> nocą</i> 'at night'	<i>v noci</i>	20.00	<i>noc</i>	23.33	<i>noci</i> 'night [gen]'	53.33

Table 61: Target words in instrumental case mistaken for words ending in -a.

An additional difficulty at the target word *szczotką* 'broom [fem, instr of *szczotka*]' is the divergent grammatical gender of the translation equivalent *smeták* or *smetákem* [masc, instr] (see below).

- **Target words with different grammatical gender**

Among the target words, there were 11 cases with divergent grammatical gender between PL stimulus and correct CS translation. Only two of these target words have CS cognate translations – *napiwek* 'tip' (CS *spropitné*) and *wazon* 'vase' (CS *váza*). At first glance, a dominance of the stimulus gender is prominent in the condition without context. In all 11 cases, the greatest percentage of the responses is of the same gender as the stimulus in the condition without context. This changes drastically in the condition with context: Target words with different grammatical gender were translated correctly significantly more often when presented in context than without any context. The difference in correct responses between the two conditions ranges from 3.1% to 73.3% with a mean increase by 28.3%.

This confirms the findings of Muikku-Werner (2014) who pointed out that sentential context can restrict the decision of respondents to fit their translation into a syntactic frame (p. 105). If, for example, a noun is preceded by a congruent adjective that has divergent gender than the one in the reader's L1 or Ln, the readers would have to revise also the presented adjective with the correct decision on the target noun.

Table 62 presents the 11 cases with a visualised comparison of the grammatical gender of the target words, the percentage of correct translations and the distribution of the grammatical gender among the responses for both conditions.

Target word in sentence	Correct CS	Distribution of grammatical gender among responses	
		No context	In context
<i>herbaty</i> [gen of <i>herbata</i> 'tea']	<i>čaje</i> [gen of <i>čaj</i> 'tea']		
<i>szczotką</i> [instr of <i>szczotka</i> 'broom']	<i>smetákem</i> [instr of <i>smeták</i> 'broom']		
<i>broda</i> 'beard'	<i>vousy</i> [pluralia tantum, base form identical]		
<i>sztuki</i> [gen of <i>sztuka</i> 'art']	<i>umění</i> [base form identical]		
<i>samochód</i> 'car'	<i>auto</i> [base form identical]		
<i>napiwek</i> 'tip'	<i>spropitné</i>		
<i>napiwku</i> [gen of <i>napiwek</i> 'tip']	<i>spropitné</i> [base form identical]		







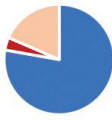

Target word in sentence	Correct CS	Distribution of grammatical gender among responses	
		No context	In context
<i>drzewo</i> 'tree'	<i>strom</i>		
<i>drzew</i> [gen plural of <i>drzewo</i> 'tree']	<i>stromů</i> [gen plural of <i>strom</i> 'tree']		
<i>biurku</i> [loc of <i>biurko</i> 'desk']	<i>(psacím) stole/stolu</i> [loc of <i>(psaci) stůl</i> 'desk']		
<i>wazonu</i> [gen of <i>wazon</i> 'vase']	<i>vázy</i> [gen of <i>váza</i> 'vase']		

Table 62: Target nouns that differ in grammatical gender between PL and CS.

The grammatical gender is indicated by a colour code in the two left columns (background colour) and in the pie charts (segments): blue for masculine, red for feminine, green for neuter and grey for other POS than the stimulus. The background colour of the cells in the two left columns represents the grammatical gender of the PL target word and the grammatical gender of the CS translation, accordingly. Likewise, the colour code in the segments of the pie charts represents the percentages of the grammatical gender among the responses given. All responses with the same grammatical gender are summarised under a segment with the same colour. For instance, for the target word *biurko* 'desk', respondents have entered a number of neuter nouns, such as *pero* 'pen', *pírko* 'little feather', and *horko* 'hot weather'. These are summarised under the neuter category represented by the green segment in the pie chart. The correct responses from the column *Correct CS* are marked in lighter shades of the colours in the pie chart segments, respectively: light blue for correct translations in the masculine gender, light red for correct feminine translations, and light green

for correct neuter translations provided by the respondents. Missing responses are not included in the pie charts. The experimental data with all details and responses listed can be found in Table A 7 of the appendix.

Concerning potential misinterpretations of inflectional endings, only the form *napiwku* [gen (+loc)] 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*.

- **Verb forms in third person plural**

For instance, *kwitną* ‘they bloom’ in which the ending *-ą* would correspond to the CS verb ending *-ou* were also frequently mistaken for a feminine noun ending: 13% translated it with a feminine noun, e.g., *teplota* ‘temperature’, *květina* ‘flower’ or *kytky* ‘flowers’ [colloquial] instead of *květou*.

- **Infinitive verb forms mistaken for nouns**

This subsection broaches the issue of the notable frequency with which PL infinitive verb forms were mistaken for nouns. In 12 of 13 cases, the infinitive verbs were more often mistaken for nouns when presented without context than in context and in one case, the intelligibility scores were the same in both conditions (see also Table 63). The reason for this becomes apparent in an error-analytical view of the responses provided. Among the responses, there were some prominent cases in which the respondents perceived the PL infinitive ending *-ć* as 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 (and orthographically) close.

Figure 35 visualises the intelligibility of infinitive verb forms with vs. without context within all other data points. The infinitive verb forms are more intelligible in context than without context.

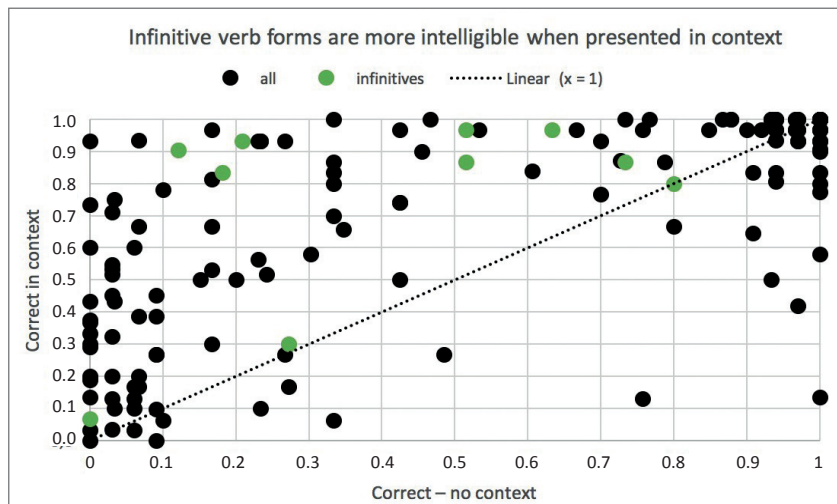


Figure 35: Comparison: infinitive verb forms with vs. without context.

The share of infinitive forms mistaken for nouns ranges from 0 in both conditions to 26.8% for target infinitives in context and 76.7% without context. On the average, 26.6% of all infinitives without context and only 6.0% infinitive verb forms in context were mistaken for nouns (including the sentences with target words with low predictability). When counting only the stimuli presented in high constraint sentences, the mean difference between the shares of nouns among the responses is 23.6% ($t(16) = 4.15, p < 0.01$). If the results from sentences with low constraint context are included, the mean difference between the two conditions is even slightly higher – 24.4% ($t(30) = 4.492, p < 0.01$).

One of the prominent examples is the target word *bawić* ‘to play’ that was translated as *bawić* ‘entertainer’ by 39.4% when presented without context. Also, other nouns which the respondents most probably associated with the concept of *bawić* were among the responses: *komik* ‘comedian’ and *zábava* ‘amusement’. The verb appeared in two of the sentences, where it was translated as *bawić* significantly less often – 13.3% and 3.2% respectively.

When *padać* was presented without any context, only 63.3% 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.’ (Block & Baldwin, 2010), however, 96.7% translated it correctly as *padat* ‘to fall’ or *pršet* ‘to rain’.

Table 63 presents a comparative overview of the target verbs together with the frequencies of their misinterpretations as nouns and correct responses with and without context. The frequency of wrong responses per stimulus is indicated for all nouns among the responses in total, although some of the individual responses might have been more frequent than others.

Target word in sentence	Correct CS	Intelligibility (%)		Wrong responses (N) without context	%	Wrong responses (N) in context	%
		No context	In context				
<i>bawić</i> 'to play'	<i>bavit (se), hrát (si)</i>	51.5	96.8	<i>bavič</i> 'entertainer', <i>komik</i> 'comedian', <i>zábava</i> 'amusement'	45.5	<i>bavič</i> 'entertainer'	3.2
			86.7				13.3
<i>dawać</i> 'to give'	<i>dávat</i>	80.0	80.0	<i>vdova</i> 'widow', <i>dárek</i> 'present', <i>důvod</i> 'reason'	10.0	n/a	0
<i>kwitnąć</i> 'to bloom'	<i>kvést</i>	20.0	66.7	<i>květináč</i> 'flowerpot', <i>květ</i> 'blossom'	76.7	<i>teplota</i> 'temperature', <i>květiny</i> 'flowers', <i>kytky</i> 'flowers'	20.0
<i>latać</i> 'to fly'	<i>létat</i>	21.7	93.3	<i>letec</i> 'pilot, aviator', <i>leták</i> 'flyer', <i>chytač</i> 'catcher'	30.4	<i>křídlo</i> 'wing'	3.0
<i>padać</i> 'to fall'	<i>padat</i>	63.3	96.7	<i>padák</i> 'parachute'	30.0	<i>padák</i> 'parachute'	3.3
<i>pasować</i> 'to fit'	<i>pasovat</i>	60.6	83.9	<i>opravář</i> 'repairman', <i>pás</i> 'belt', <i>pas</i> 'passport', <i>pásovec</i> 'armadillo', <i>pasovač</i> 'smuggler'	15.2	n/a	0
<i>plakać</i> 'to cry'	<i>plakat</i>	73.3	86.7	<i>plakát</i> 'poster', <i>plac</i> 'place'	26.7	<i>plášť</i> 'cape', <i>pekáč</i> 'pan'	6.7
			100.0				n/a
<i>plywać</i> 'to swim'	<i>plavat</i>	18.2	83.3	<i>plyvátko</i> 'spitting bowl', <i>plivač</i> 'spitter', <i>jazyk</i> 'tongue'	12.1	<i>kopec</i> 'hill'	3.3
<i>przebrać</i> 'to change clothes'	<i>převléct</i>	0	6.7	<i>žebřák</i> 'beggar', <i>hrbáč</i> 'hunchback', <i>změna</i> 'change'	15.2	<i>pohrabáč</i> 'poker', <i>hráč</i> 'player', <i>příbytek</i> 'abode', <i>žebřík</i> 'ladder', <i>žebřák</i> 'beggar', <i>hlavu</i> 'head [accu]'	26.7

Target word in sentence	Correct CS	Intelligibility (%)		Wrong responses (N) without context	%	Wrong responses (N) in context	%
		No context	In context				
<i>przetrzymywać</i> 'to hold aptive'	<i>zadržovat</i>	0	43.3 ²⁵	<i>promítač</i> 'projectionist', <i>vypínač</i> 'switch', <i>myčka</i> 'dishwasher', <i>přesmyčka</i> 'anagram', <i>jízdní řád</i> 'schedule', <i>předpírka</i> 'pre-treatment', <i>záclona</i> 'curtain', <i>umyvadlo</i> 'sink', <i>prezentace</i> 'presentation', <i>zahradník</i> 'gardener'	36.7	n/a	0
<i>rosnąć</i> 'to grow'	<i>růst</i>	12.1	90.0	<i>ručník</i> 'towel', <i>mravenečník</i> 'anteater', <i>rosa</i> 'dew', <i>moucha</i> 'fly', <i>rosomil</i> [made up word], <i>rostlina</i> 'plant', <i>rosník</i> 'frog', <i>rosnička</i> 'tree frog', <i>žába</i> 'frog'	42.4	n/a	0
<i>stroić</i> 'to tune'	<i>ladit</i>	0	10.0	<i>stroj</i> 'machine', <i>strýc</i> 'uncle', <i>strach</i> 'fear'	63.3	n/a	0
<i>tonąć</i> 'to sink'	<i>tonout</i> , <i>topit se</i>	27.3	30.0	<i>tón</i> 'tone', <i>tanec</i> 'dance', <i>toner</i> 'toner', <i>tác</i> 'tablet', <i>tonic</i> 'tonic', <i>tuna</i> 'ton'	27.3	No Ns, only verbs and <i>dnes</i> 'today'	0
<i>usłyszeć</i> 'to hear'	<i>uslyšet</i>	90.9	64.5	<i>uklízeč</i> 'cleaner', <i>uzlíček</i> 'little knot'	9.1	n/a	0
Mean all sentences		37.1	71.7		29.4		5.0
Mean infinitive in sentence		38.6	75.0		26.6		6.0
Mean high constraint		39.9	71.7		28.9		5.3

Table 63: Target verbs mistaken for nouns with vs. without context.

While the target words in the free translation task without context were presented in their infinitive forms, the verb forms in context might have been in other forms – for instance, *plakać* 'to cry' without context vs. *plakalo* 'it cried' in context. The grey cells in Table 63 mark all the cases in which the target verbs in sentential context were not in their base forms. The cell *mean infinitive in sentence* in Table 63 indicates the mean values for all target words that were infinitive verb forms in the sentences, too. Although, in general, target words

25 This verb form was tested only in one of the low constraint sentences and not in one of the high probability, high constraint sentences.

in base forms were expected to be translated more often correctly, because base forms were shown to be easier to process in monolingual context (cf. for instance Sekerina et al., 2018), this proved to be correct only for the infinitives in sentential context, but not for PL infinitive verb forms when presented to Czech readers without any context.

15.5. Correlations and Model

As a result of the error analysis, the following variables were added to the statistical analysis of possible predictors:

- 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 (see section 15.4.3.3).
- A binary variable for the category of false friends (*FF* and *FF/words* in Table A 17 of the appendix) was added, since false friends can be cognates or non-cognates (see section 15.4.2.4).

Also, the number of words per sentence was added as an additional predictor. The following correlation analysis serves to determine whether correlations between any of the predictors linguistic distance, surprisal, the variables added after the error analysis (15.4.3), and the intelligibility scores for both conditions could be found. First, a simple linear regression is performed with the individual predictors. Second, those predictors that turned out relevant are unified for a multiple linear regression model. The full correlation matrix can be found in Table A 19 in the appendix.

First of all, none of the surprisal values (target, bigram, trigram, or mean) correlates with the cloze probability of the sentences. This might be due to the fact that the cloze probabilities in the given sentences do not vary strongly, as they are high-constraint cloze sentences and range from 0.90 to 0.99 (Block & Baldwin, 2010). There is a highly significant intercorrelation between the corresponding surprisal measures (for target, bigram, trigram, and sentence) from the two LMs (the CS and the PL one), the strongest correlation being that of the total surprisal per sentence in both languages ($r = 0.732, p < 0.001$).

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 intelligibility results obtained in the context condition. The correlation of the CS target words' surprisal and target word intelligibility is only slightly higher than that of the PL surprisal of the target words ($r = -0.191 > r = -0.186, p < 0.05$). 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 PL stimuli sentences, not in the case of their closest CS translations. However, when leaving the

cognates out of the analysis, the correlation with the total surprisal of the PL sentence increases to $r = -0.411$ ($p < 0.01$), even more when correlating only the false friends and intelligibility ($r = -0.443$, $p < 0.01$).

For the linguistic distance measures, all correlations are highly significant for the target words in context. The correlations are somewhat stronger for all total distance measures (unifying orthographic and lexical distance) as opposed to their corresponding pron LD. This also applies to the correlations of the results for the target words without context and their distance measures: Their total distance has a slightly higher correlation with the results than pron LD only ($r = -0.772 > r = -0.767$, $p < -0.001$). The correlations are the highest with the linguistic distance of the target words and not of the bigram, trigram or sentence distances. The longer the 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 ($r = -0.772$, $p < 0.001$) than in context ($r = -0.680$, $p < 0.001$).

All lexical distance and false friend variables proved to be highly significant, the strongest correlation being the total number of non-cognates per sentence ($r = -0.508$, $p < 0.001$). Both number of non-cognates and false friends correlate stronger with the results ($r = -0.353$, $p < 0.001$ for the category of false friends) when they are counted as a total score per sentence than when normalised 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 ($r = -0.272$, $p < 0.001$). Without context, the correlation of grammatical gender and intelligibility is only slightly higher ($r = -0.281$, $p < 0.001$). 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 $R^2 = 0.496$ ($p < 0.001$), i.e. this model can account for 49.6% of the variation in the data for all sentences with highly predictive context. For the condition without context, a model with the predictor variables distance (not total, but pron LD), gender, false friends, and non-cognates has an adjusted $R^2 = 0.644$ ($p < 0.0001$), i.e. this model can account for 64% of the variation in the data. The more detailed overview with coefficients for each predictor in the models for both conditions can be found in the appendix (Table A 19 with context and Table A 20 without context).

15.6. Summary and Discussion

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). The correlation of surprisal and target word intelligibility also depends on the lexical category of the target word: no correlation with surprisal could be found for target words that are cognates. However, a significant correlation of surprisal was found for target words that are non-cognates or false friends. 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. This is confirmed by the multiple linear regression models: The gender category is not of relevant in the best fitting model for target words in context, but contributes to the best fitting model for target words without context.

Czech readers proved to be unlikely to identify the PL ending *-q* 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 infinitive verb forms in sentences, since they were significantly more often mistaken for nouns when presented without context.

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

An analysis of intelligibility for the different lexical categories of target words reveals different levels of importance of the predictors in 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 (C-IB), cognates in other contexts (C-OC) and “true” false friends (FFs) that do not have correct cognate translations or do not offer any possibility for a correct semantic association. The lack of significance

for the latter two categories might be caused by the low number of these items in the data set. The difference between the two conditions was the greatest for false friends that are cognates (FF-C) and for non-cognates that offer possible associations with the correct translations (NC-A).

For real cognates (C-C), 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.

16. The Impact of Random Context on the Understanding of Particular Words in Sentences from the Cooperative Translation Task

In this section, a quantitative error-analytical approach is chosen to evaluate the results of a web-based cloze translation experiment with the sentences from the cooperative translation task (CHAPTER II). The intention of the experiment is not only to obtain a more representative sample of responses to critical words within the stimuli, but also to compare whether the random context helps disambiguating the critical words in a similar way as it does for most of the target words with high cloze probabilities as shown in section 15. In addition to the sentences from the cooperative translation task, 10 other sentences (hereafter referred to as other sentences) were added to the stimuli in order to have a bigger set of stimuli and for possible later analyses. Hence, the underlying hypothesis is that also random context can improve the intelligibility of target words even if they do not necessarily have a high cloze probability in sentences. The types of errors that occur are compared throughout the sentence stimuli with target words in different contexts.

16.1. Method

The web-based cloze translation experiments were conducted only after the written results of the cooperative translation experiments were evaluated. The 12 stimuli sentences were presented again to another, bigger sample of Czech native speakers (same method as described in section 15.1) in order to collect a more representative data set for certain problematic words in the sentences. These problematic words were placed in the cloze gaps for translation. 10 other sentences containing false friends, some of which were part of my state examination thesis (Jágrová, 2010) or were inspired or copied from signs and advertisements in the streets of Gdańsk and Warsaw, were added in a second stimuli block for further analyses. The stimuli were divided into two blocks – one with 12 and another one with 10 stimuli sentences. Both blocks were presented to 33 respondents each. One of the two blocks was assigned to a respondent automatically. The sentences and possible CS and EN translations are listed in Table 64. In the sentences 13-22, words that are non-cognates are marked **bold**, words that can be cognates in another context are marked **bold and blue**, false friends are **bold and red**. Cognates with morphological differences are underlined²⁶. Words that the respondents were asked to translate (gaps) in the cloze task are framed. The EN translations previously provided in Table 13 (CHAPTER II) should assist the comprehension of the differences between the PL stimuli and the possible CS translations and are therefore not always identical with the EN sentences listed in Table 64.

26 The respective marking for sentence 1-12 can be found in Table 13 of CHAPTER II.

PL stimulus—ORIG condition	Possible good CS translation	EN translation
1	Gdyby nie było książek, czytałbym Ci z oczu.	If there were no books, I would read from your eyes.
2	W 2000 roku wzrósł do ponad 900 mln. marek obrót towarami, w procesie produkcji których nie używano substancji zagrażających środowisku naturalnemu wilka.	In the year 2000, the turnover of goods in the production of which no substances that are harmful for the natural habitat of the wolf are used, rose above 900 million German mark.
3	Kolegium dało mi pozwolenie, aby zrealizować ten projekt nad jeziorem.	The council gave me the permission to realize the project at the lake.
4	Praga to ważny węzeł komunikacyjny.	Prague is an important traffic hub.
5	Czy pani będzie głosowała? Czy chcielibyście, aby stały się one gwiazdami?	Madam, are you going to vote? Do you want [pl] that they [group of females] become stars?
6	Kupiliśmy nie tylko czerstwy chleb, ale jeszcze gorzej – też były żółty samochód.	Not only did we buy stale bread, but even worse – also an old yellow car.
7	Teraz [rosną] również możliwości odbycia interesujących praktyk w kraju.	Right now, also the possibilities of undergoing interesting internships at home are growing.
8	Nie widziałam, że jego żona pokazuje rękę, zebymy poszli do rektora.	I have not seen that his wife is showing with her hand that we should go to the rector.
9	Skąd jesteś przelotem, że za pięćdziesiąt lat ludzie nie będą już latali samolotem?	Why are you convinced that in fifty years people will no longer fly airplanes?
10	OCZEKIWANIA: doświadczenia] w pracy przy produkcji mięsa; pełną dyspozycyjność od poniedziałku do piątku. OFERTA: realne możliwości awansu w firmie; 12,00 brutto/godzinę + premie miesięczne	EXPECTATIONS: work experience in the meat production; [full] availability from Monday to Friday. OFFER: realistic promotion opportunities within the company, 12.00 gross/hour + monthly boni
11	OBSŁUGA SKLEPU – ZAKRES OBOWIĄZKÓW: znajomość języka polskiego; ekspozycja towarów; gotowość do pracy zmianowej; czynności porządkowe	SALESPERSON IN A RETAIL STORE – SCOPE OF DUTIES DESCRIPTION: knowledge of Polish; display of goods; readiness for shift work; cleaning activities

	PL stimulus — <i>ORIG</i> condition	Possible good CS translation	EN translation
12	NAPÓJ Z MIĘTY I MIODU: mięta zielona suszona: 25 g, miód kwiatowy: 50 g, cytryna: 1 szt., [lód] konsumpcyjny: 5 kostek, sok z brzozy: 100 ml; jarzębiny: 50g.	NAPÓJ Z MĄTY A MIEDU; suszená zelená máta: 25g; květový med: 50g; [letrón]: 1 kus; konzumní led: 5 kostek; [šťáva] z brzozy: 100ml; jeřabiny: 50g.	MINT AND HONEY DRINK: dried green mint: 25 g; blossom honey, 50 g; [lemon], 1 piece; consumable [ice], 5 cubes; [birch] [sap], 100 ml; rowan berries: 50 g.
13	[státek] zabiera max. 65 osób i dysponuje dwoma [pokładami].	[Lod] pobere max. 65 osob a disponuje dvěma palubami.	[The ship] accommodates up to 65 people and has two decks.
14	Poszła do [sklepu] i kupiła [znaczek].	Zašla do [obchodu] a koupila [známku].	She went to the [shop] and bought a [stamp].
15	Zakaz palenia wyrobów [tytoniowych] w pojeździe.	Zakaz [kouření] [tabákových] výrobků ve vozidle.	Smoking [tobacco] products is prohibited in the vehicle.
16	[rusztowania], szalunki, zsypy do gruzu, sprzęt budowlany, sprzęż.	[řešení], bednění, shozy na suť, stavební úklid, prodej	[Scaffolding], formwork, rubble chutes, site cleaning, sales
17	NOWOŚĆ DO PRANIA Czystość, która [pięknie] [pachnie] NOWY ZAPACH	NOVINKA NA PRANÍ Čistota, která [krásně] [voní]. NOVÁ VŮŇ	LAUNDRY NOVELTY Purity that [smells] [nice]. [NEW FRAGRANCE]
18	ZŁOTE MISIE SMIAK RADOŚCI BEZ SZTUCZNYCH BARWNIKÓW	ZLATI MEDVÍDCI PŘÍCHUŤ RÁDOSTI BEZ UMĚLÝCH BARVIV	GUMMY [literally GOLD] BEARS THE TASTE OF JOY WITHOUT ARTIFICIAL COLOURING
19	PROSZĘ NIE WYCIAGAĆ [LIZAKÓW]!	PROŠÍM NEVYTAHOVAT [LÍZÁTKA]!	Please do not tear out the [lollipop].
20	[POWIERZCHNIA] REKLAMOWA DO WYNAJĘCIA	REKLAMOVÁ [PLOCHA] K PRONÁJMU	ADVERTISING [SPACE] FOR [RENT]
21	W czasie pracy klimatyzacji okna są [zamknięte].	Při spuštění klimatizace jsou okna [zavřena].	When the air conditioning is running, the windows are [closed].
22	Biurowo Trojmiasto.pl Wynajem biur – Nowoczesne przestrzenie, Wyjątkowe lokalizacje	Kancelář Trojmiěstí Pronájem [kanceláří] – Moderní prostory, Výjimečné lokality	Tricity Office Rental of [offices] – Modern premises, Exceptional locations

Table 64: Stimuli from the cooperative translation task + 10 other sentences presented in the cloze translation experiments with random context.

As it was not clear to which extent the sentences from the cooperative translation experiments provide helpful context for understanding the critical words, monolingual cloze completion experiments were conducted in order to create a baseline and to obtain cloze probabilities for the target words in the original stimuli.

16.2. Baseline Experiments: Cloze Probabilities in Monolingual Context

The cloze probability tests were conducted over SoSci Survey (Leiner, 2019) – a software package for online surveys which was made available to the respondents through www.soscisurvey.de. The survey was carried out in CS, PL, EN, and DE. The participants were asked to fill gaps in sentences with words they consider most suitable in the respective sentential context. The data gathered in the baseline experiments allows us, *inter alia*, to:

- Estimate the (un-)predictability of the original stimuli words from the intercomprehension cloze experiments.
- Determine the responses that readers consider most likely in a (comprehension) gap.
- Classify the responses in the cloze translation experiments as either context- or similarity-driven (or neither).

As a consequence, the cloze probabilities of (words in) gaps should correlate with their surprisal scores, if the language model is of good quality. This question, however, will not be addressed here. The focus of this subsection lies on the classification of context-driven errors and on the possible role of random context for the intelligibility of target words in random position.

16.2.1. Design

The data was collected in different conditions in order to present only 1 gap in a stimulus sentence at a time. This means that sentences with three gaps in the original PL stimulus had to be tested three times in the survey, each time with the gap placed at a different position. The task presented to the respondents can be found in section 5.1 of the appendix (EN version).

16.2.2. Results

The responses were evaluated in the following categories:

- cloze probability (most frequent answer in %) and
- response equal or synonymous to the word from the original PL stimulus (in %).

The CS and PL versions of the stimuli used in the monolingual cloze probability experiments are listed in the respective conditions under section 5.2 (Table A 12 and Table A 13) of the appendix. Regarding the word from the PL stimulus, synonymous expressions – those that were counted as correct in the cloze translation experiment – were summarised under one category. For instance, the CS responses *výrobků* ‘products [gen pl]’, *produktů* ‘products [gen pl]’, and *zboží* ‘goods’ were summarised for where the PL word *towarów* ‘goods [gen pl]’ in the original PL stimulus was. All other responses that were summarised were not distinguished between synonymous or non-synonymous, as it is often not clear which responses can be considered synonyms and which not. Grammatical or morphological differences in the responses, however, e.g. *knih* ‘books [gen pl]’ vs. *knížek* ‘books [gen pl, diminutive]’ or *myslíš* ‘you think’ vs. *myslel* ‘you thought [masc]’ vs. *si myslel* ‘you thought (to yourself) [masc, reflexive]’ vs. *myslela* ‘you thought [fem]’, were not considered different responses. Cases such as German *Käsestück* ‘piece of cheese’ and *Stück Käse* ‘piece of cheese’ were treated as the same answer. Words and their short synonyms such as *laboratory* and *lab* were also not considered different responses. The same applies for orthographic errors in the responses. Cases in which an additional word was entered, e.g. *práce* ‘work’ vs. *ruční práce* ‘manual work’ were counted as two different responses. No distinction was made between responses in upper and lower case. In cases where respondents have entered numbers in the gaps, all numbers were summarised as one type of answer.

Some responses might be classified as primes due to repetition, for instance the response *znalost* ‘knowledge’ which was the most frequent response for the gap in the stimulus

OBSLUHA OBCHODU – ROZSAH POVINNOSTÍ: znalost polského jazyka, [] zboží, ochota práce na směny, udržování pořádku.

‘SHOP ASSISTANT – SCOPE OF DUTIES: knowledge of Polish, [] of goods, willingness to work in shifts, keep order.’

which might be because the word *znalost* appears in the same sentence before the gap.

Culturally different cases had to be decided upon as how to treat them. One such case was the original PL stimulus *ZŁOTE MISIE*, which is the PL translation for the named entity *Gummy bears*, while the literal translation would be ‘golden bears’. Czech readers were asked to translate *ZŁOTE* ‘golden’, however, also the correct CS equivalent *gumoví medvídci* was considered a correct answer. Therefore, *golden*, *gummy* and *Haribo* were summarised as the same answer in the monolingual cloze tests.

While the top answers for *She went to the shop to buy apples and ...* for the EN speaking respondents was oranges, DE and CS speaking respondents' top answer was *Birnen* and *hrušky* 'pears'. *Pears* was the second most frequent response of EN speaking respondents, while none of the DE speaking respondents answered *Orangen* 'oranges' and only one CS respondent answered *pomeranče* 'oranges'.

16.3. Scoring of Responses

For the scoring and categorization of responses, the same principles as applied to the high-constraint, high-probability sentences were applied to the present set of sentences. In addition, the following decisions were made: For instance, in stimulus sentence 18

ZŁOTE MISIE – [SMAK]RADOŚCI

'GUMMY BEARS – [THE TASTE] OF JOY',

respondents were asked to translate the noun *SMAK* 'taste' into CS. 11 out of the 33 respondents answered *chut'* or *příchut'* correctly, but 3 of them entered the verb *chutnají* 'they taste', so that the translation of the phrase *SMAK RADOŚCI* changes minimally from *chut' radosti* 'the taste of joy' to *chutnají radosti* 'they taste of joy' with the grammatical case of *RADOŚCI* being interpreted as instrumental instead of genitive. A similar case occurred with *POWIERZCHNIA REKLAMOWA* 'advertising space' in sentence 20, where respondents were asked to translate the noun *POWIERZCHNIA* 'surface' (CS: *plocha* or *povrch*). Two respondents entered the translation *povrchová* 'surface [A]', which they in combination with *REKLAMOWA* 'advertising [A]' most probably understood correctly, although turning the noun into an adjective and the adjective into a noun. Such cases were counted under a sub-category of correct: answers with grammatical divergences, which might indicate that there is a more natural way of expressing the meaning in CS, for instance with a different morpho-syntax. This **part of speech tolerance in responses applies only to the cloze experiments**. Another discussable case is *zamknęte* 'closed' in the sentence

W czasie pracy klimatyzacji okna są zamknięte.

'When the air conditioning is running, the windows are closed.'

In the cloze translation task, 30.3% of the Czech respondents translated *zamknęte* with the formally more similar *zamkněte* 'lock [V, imperative]' or *zamykají* 'they lock' instead of the correct adjective *zavřená* 'closed'. Although the concepts of *close* and *lock* are related meaning-wise, they are not identical and, in this situation, establish a different meaning. Therefore, *zamkněte* and forms of *zamknout* 'to lock' were counted as wrong. As mentioned earlier, cases in which respondents entered a translation in EN are accepted.

16.4. Results: Target Words at Random Position

Table 65 provides an overview of the correct, wrong, and missing translations for the target words (in alphabetic order) in sentences at random position. A column for responses with divergent grammar is added (within the correct responses). The most frequent wrong answers are indicated in the right column. For the gaps which contained more than one word, the intelligibility scores are given for the whole phrase in the gap and, in addition, for its constituents separately.

Stimulus in gap	Correct	Grammar divergent	Wrong	No answer	Most frequent wrong answer
<i>będzie</i>	0.12	0.03	0.39	0.48	<i>běžně</i> 'regularly'
<i>biur</i>	0.06	0	0.50	0.44	<i>bytú</i> 'apartments [gen pl]'
<i>brzozy</i>	0.45	0	0.24	0.30	<i>bez</i> 'without', <i>bezu</i> '[elder gen/dat/loc]'
<i>bym</i>	0.30	0.18	0.15	0.55	n/a
<i>Ci</i>	0.09	0	0.12	0.79	<i>mi</i> 'me', <i>si</i> 'oneself', <i>jsi</i> '[you] are', <i>nebo</i> 'or'
<i>cytryna</i>	0.79	0.03	0.15	0.06	<i>kyselina citrónová</i> 'citric acid'
<i>czerstwy</i>	0	0	1.00	0	<i>čerstvý</i> 'fresh'
<i>czytał</i>	0.24	0.06	0.39	0.36	[various]
<i>czytałbym</i>	0.27	0.15	0.33	0.39	[various]
<i>czytałbym ci z oczu</i>	0.09	0	0.55	0.36	[various]
<i>do rektora</i>	0.06	0.03	0.58	0.36	<i>do kostela</i> 'to church', <i>učení</i> 'teaching/learning/apprenticeship', <i>učitele</i> 'teacher [accu]'
<i>doświadczenia</i>	0.03	0	0.85	0.12	<i>osvědčení</i> 'attestation'
<i>ekspozycja</i>	0.79	0	0.12	0.09	<i>znalost</i> 'knowledge', <i>výroba</i> 'production'
<i>głosowała</i>	0.27	0.21	0.52	0.21	<i>głosowała</i> 'she glossed'
<i>godzinę</i>	0.64	0.06	0.18	0.18	<i>peníze</i> 'money', <i>peněz</i> 'money [gen]', <i>plat</i> 'pay [n]'
<i>gorzej</i>	0.15	0	0.61	0.24	<i>horký</i> 'hot'
<i>gwiazdami</i>	0.64	0	0.09	0.27	<i>zdyh, vadili</i> '[they] harmed'
<i>lżaków</i>	0.25	0.09	0.41	0.34	<i>jazyky</i> 'languages', <i>lyžáky</i> 'skiing trip [pl]/ski boots'
<i>lód</i>	0.45	0	0.45	0.09	<i>jód</i> 'iodine'
<i>mięsa</i>	0.73	0	0.27	0	<i>města</i> 'town [gen]', <i>měsíce</i> 'moon [gen]', <i>měsíčně</i> 'monthly'
<i>możliwości (awansu)</i>	0.61	0	0.21	0.18	<i>množství</i> 'amount', <i>zaměstnanci</i> 'employees'
<i>możliwości (praktyk)</i>	0.36	0.03	0.36	0.27	<i>myslivci</i> 'hunters'
<i>nad jeziorem</i>	0.91	0.06	0.06	0.03	<i>pod dozorem</i> 'under surveillance', <i>samostatně</i> 'separately'
<i>nie widziałam</i>	0.12	0.09	0.73	0.15	<i>nevěděl jsem</i> 'I did not know'
<i>nowy zapach</i>	0.69	0.03	0.28	0.03	<i>nový zápach</i> 'new stench'

Stimulus in gap	Correct	Grammar divergent	Wrong	No answer	Most frequent wrong answer
<i>pełna</i>	0.70	0.06	0.18	0.12	<i>piłná</i> 'diligent', <i>pevná</i> 'stable [fem]'
<i>pięknie</i>	0.94	0.03	0.06	0	<i>hodně</i> 'much', <i>velmi</i> 'very'
<i>pokładami</i>	0.03	0	0.84	0.13	<i>pokladnami</i> 'cashdesks [instrumental]'
<i>powierzchnia</i>	0.06	0.06	0.59	0.34	<i>pověřčivá</i> 'superstitious [fem]', <i>povrchní</i> 'superficial'
<i>pracy zmianowej</i>	0.09	0	0.79	0.12	<i>zmiňně / zmiňňovaná práce</i> 'mentioned work'
<i>przekonana</i>	0	0	0.85	0.15	<i>překonaná / překonána</i> 'overwhelmed'
<i>ręka</i>	0.03	0	0.67	0.30	<i>řeka</i> 'river'
<i>rosna</i>	0.06	0.03	0.55	0.39	<i>rosa</i> 'dew'
<i>rusztowania</i>	0	0	0.56	0.44	<i>růst</i> 'growth'
<i>samochód</i>	0.21	0	0.42	0.36	<i>samochod</i> 'self-goer', <i>samoobchod</i> 'self-shop', <i>kolo</i> 'wheel / bike'
<i>samolotem</i>	0.39	0.03	0.30	0.30	<i>samopalem</i> 'machine gun [instr]'
<i>skąd</i>	0.09	0	0.55	0.36	<i>snad</i> 'hopefully'
<i>sklepu</i>	0.50	0	0.47	0.03	<i>sklepa</i> 'basement [gen]'
<i>sklepu</i>	0.24	0	0.76	0	<i>sklepu</i> 'basement [gen/dat]'
<i>smak</i>	0.44	0.09	0.38	0.19	<i>jídlo</i> 'food', <i>má</i> 'has [3rd pers sg] / mají 'have [3rd pers pl]'
<i>sok</i>	0.48	0	0.33	0.18	<i>suk</i> 'knot'
<i>sprzedaż</i>	0.16	0.03	0.28	0.56	<i>předáš</i> '[you] hand over'
<i>stary</i>	0.85	0.09	0	0.15	n/a
<i>stary żółty samochód</i>	0.21	0	0.64	0.15	<i>staré žluté kolo</i> 'old yellow bike', <i>starý žlutý samochod</i> 'old yellow self-goer'
<i>statek</i>	0	0	0.88	0.13	<i>statek</i> 'farm'
<i>sztucznych</i>	0.13	0	0.72	0.16	<i>tučných</i> 'fatty [gen pl]'
<i>towarami</i>	0.18	0.09	0.73	0.09	<i>továrnami</i> 'factories [instr]'
<i>tytoniowych</i>	0.09	0	0.78	0.13	<i>titanových</i> 'titan [adj, gen pl]'
<i>węzeł</i>	0.58	0	0.30	0.12	<i>věděl</i> 'he knew'
<i>wilka</i>	0.03	0.03	0.48	0.48	<i>mléka</i> 'milk [gen]', <i>chvilka</i> 'a while'
<i>wyjatkowe</i>	0.03	0	0.44	0.53	<i>majetkové</i> 'property [adj, pl]', <i>vyhlídkové</i> 'view [adj, pl]', <i>Vítkově</i> 'Vítkov [loc]'
<i>wynajęcia</i>	0	0	0.5	0.50	<i>vánoce</i> 'Christmas [gen pl]', <i> vynálezu</i> 'invention [gen]'
<i>z oczu</i>	0.21	0	0.12	0.67	<i>z octu</i> 'out of vinegar'
<i>zagrożających</i>	0.03	0	0.61	0.37	<i>zahraničních</i> 'foreign [gen pl]', <i>zakázaných</i> 'forbidden [gen pl]'
<i>zamknięte</i>	0.63	0.50	0.34	0.03	<i>zamkněte</i> 'lock [imp 3rd pers pl]'
<i>znaczek</i>	0	0	0.78	0.22	<i>značka</i> 'sign', <i>znáček</i> 'small sign'
<i>żółty</i>	0.45	0.03	0.24	0.30	<i>zlatý</i> 'golden'

Table 65: Results of the cloze translation task with target words at random position.

16.4.1. Comparison: Types of errors

In the following, I will attempt to provide explanations for the answers entered by the respondents. The results overall reveal a distinction between what I will call form-oriented responses and context-oriented responses. All wrong responses are classified into the following four categories:

- a) context-driven (wrong response still fitting the context),
- b) similarity-driven (neighbourhood or low LD to the stimulus, sometimes lower than the correct option, including false friends, rhymes, also errors caused by wrongly assumed pronunciation),
- c) association, bias, or priming,
- d) L_n interference.

Wrong answers are categorized under a) context-driven if they were among the words entered for this gap in the monolingual cloze experiments (section 0) by at least one person in at least one of the languages (CZ, DE, EN, or PL). Of course, in order to provide a translation that fits the context, respondents must have understood enough of the context.

If a response could not be categorised as context-driven, then it might be categorised in one of the other categories. When looking at the wrong answers in the right column in Table 65, there are some obviously similarity-driven responses, such as *řeka* ‘river’ for *ręka* ‘hand [instr]’. If such a response has lower LD to the stimulus than the actual correct response (in this case *rukou* ‘hand [instr]’ has an LD of 60%, while *řeka* has only 37.5%), it is categorised as similarity-driven. Also, if its LD is higher, but shares some common features with the stimulus, it is categorised as such. In these cases, I assume that either the context was not sufficiently understood or respondents simply focussed on the target word and did not pay attention to the context for reasons of effort.

If it cannot be categorised in either of the two categories, it is counted under c) association. Obvious interferences from languages other than the L1 of the reader are categorised under d). For instance, the word *doświadczenia* ‘experience [pl]’ was translated as *na shledanou* ‘goodbye’ by one of the Czech respondents. In this case, the reader obviously thought of the RU *do свидания* ‘goodbye’ and entered the CS translation of this accordingly. Wrong answers that could not be categorised in any of these four categories, such as mere repetitions of words from the sentence, re-types of the PL stimulus or responses consisting of only one or two letters that could not be identified as an existing word in any of the languages were not included in these statistics. Some responses can be categorised into more than only one category – these cases will be explained as follows, as they seem to cause especially strong misleading effects. An interesting example of such a double-category word is *myslivci* ‘hunters’ as a response to *możliwości* ‘opportunities’ in the sentence

Teraz **rosną** również **możliwości** odbycia interesujących praktyk w kraju.

‘Right now, **opportunities** to do interesting internships in your home country are increasing.’

We know from the transcripts of the cooperative translation experiments (CHAPTER II) that some of the Czech readers do not recognise the PL verb *rosną* ‘they increase, they grow’ as such, but that it is in 80% of all instances pronounced as /rosna/ which causes associations with *rosa* ‘dew’ instead, which, together with its similarity, apparently evokes an association of another word in the sentence – *możliwości* ‘possibilities’ – with *myslivci* ‘hunters’ (9 of 33 responses). Evidence for this association can be found in the transcripts of the cooperative translation experiment:

- *rosną* -> *jaro* ‘spring’, *růže* ‘rose’, *kytka* ‘flower’, *rosa* ‘dew’:

P14/7: B: Ted’... a co ta **rosna**?

A: **Tohoto jara?**

B: Nevím, mně **to zní jak...** jako, že to nemusí být ted’ něco, ale jakože **letošního** nebo **nynějšího**.

B: No, já nevím. To mi přijde jak **růže**, jak nějaká **kytka** nebo **rosa**.

A: **Květen?**

B: **Evokuje mi to jaro**, no.

A: **Tento květen** nebo **letošní** je lepší, vid’?

‘B: Now ... and what about that [reading *rosną*]?’

A: **This spring?**

B: I don’t know, to me **it sounds like** ... like, that it doesn’t have to be something now, but like **this year’s** or **present**.

B: Well, I don’t know. It seems to me like **rose** or some **flower** or **dew**.

A: **May?**

B: **It evokes spring** in me, yeah.

A: **This May** or **this year’s** is better, right?’

P3/7: A: No, to by mohlo být. Ono to totiž strašně **zní jako rosna** a je to zavádějící ...

‘A: Well, that could be. This in particular terribly **sounds like /rosna/** and it’s very confusing ...’

- **możliwości** -> *myslivost* ‘hunting’, *myslivci* ‘hunters’:

P2/7: B: [...] mně normálně **hrozně zní**, není to něco s **myslivostí**? Jako, prostě **myslivci**?

‘B: [...] For me this **terribly sounds** like, isn’t that something with **hunting**? Like, **hunters**?’

The word *możliwości* was also part of the stimulus sentence 10:

OFERTA: realne [możliwości] *awansu w firmie; 12,00 brutto*[godzinę] *+ premie miesięczne.*

‘OFFER: realistic promotion [opportunities] in the company, 12.00 gross [hour] + monthly boni.’

There, only one of the 33 respondents translated *możliwości* as *myslivci*, which suggests that we are dealing with semantic lexical priming through the word *rosną* in sentence 7, since both dew and hunter can, for instance, be associated with the concept of the forest.

The results of the categorisation of errors across the cloze translation experiments with target words in the different contexts and positions are compared in Table 66. Also, the same analyses of errors in the cloze translation experiments with target words in high- and low-constraint context were added and categorised accordingly in the overview in Table 66.

Stimuli set	Context-driven	Similarity-driven	Association	Ln interference
Cooperative translation task sentences	9.4%	27.8%	7.9%	0.4%
Other sentences	9.0%	29.0%	11.8%	0.2%
High-constraint sentences	9.1%	15.5%	4.1%	0.7%
Low-constraint sentences	7.9%	14.8%	8.2%	0.3%

Table 66: Comparison: types of wrong responses in different stimuli sets.

The results show that for all types of sentence stimuli, similarity-driven errors are the most frequent type of wrong responses. It could be identified that similarity-driven errors make up almost 30% of all errors made when translating the stimuli with target words in gaps at random position, while the ratio is only about half as high (around 15%) for the high- and low-constraint sentences in which the target words were placed at sentence final position. The fact that the difference in similarity-driven errors between high- and low-constraint sentences is minimal (0.7%) suggests that it might be relevant at which position of the sentence the target word is placed and that target words at sentence final position might be easier to comprehend in general. Respondents seem to rely more on similarity of the target word when it is at a random position in the sentence than when the word is at sentence final position. Nevertheless, it has to be kept in mind that the target words in the sentences with random context were selected

because they proved to be problematic in previous experiments, while the target words in the high-constraint sentences were chosen for their predictability in context and were not selected because of being especially problematic. As for the context-driven wrong responses, all values lie around 9%, except for the low-constraint sentences with around 8%. The greatest differences between the types of stimuli can be found among the errors due to associations or priming: The effect of associations and priming seems to play a rather small role in the sentences with highly predictable target words (only 4%). For all types of sentences, the share of errors due to Ln interference lies below the 1% level.

In the experiments with highly predictable target words, 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 a lower number of correct responses than in the condition without context, even if the target word is a frequent cognate. The following extract from a discussion by pair 5 is an interesting manifestation of how associations of the word *samochód* ‘car’ compete with context-driven decisions in sentence 6. This sentence contains the word *chleb* ‘bread’ that seems to be a semantically dominant concept:

- *samochód* -> *Trabant, chůdy* ‘stilts’:

P5/6: A: Mně tam ten **samochod**, to prostě, **naskakuje mi Trabant**, ale [...] Neříkalo se tak nějakýmu autu? **Chudy?** [...]

B: **Chleba**. Mně napadaj **chůdy** nebo nějaké takové ty... **takové ty pomocné chodítka** pro, **pro seniory**, ale to, to je blbost... **chleba**... něco takového. [...] Myslíš si, že to bude něco podobného na **chleba** jako... taky nějaké **pečivo** nebo?

A: **No ze zbytku věty by to něco s tím jídlem mohlo mít společnýho**. Taky mi to trochu připomíná takovou tu **pojízdnou prodejnu, co jezdila dřív po vesnicích**. [...] Tak, co si myslíš, že myslíš, že to je třeba nějakej dopravní prostředek?

B: No zní to tak, ale **nesedí mi to do kontextu. Nějaká trojkolka** nebo něco.

‘A: For me this *samochod*, that’s just, **reminds me of Trabant**, but [...] Wasn’t that the name of a car? **/xudi/**?’

B: **Bread**. I think of **stilts** or some sort of these ... **such auxiliary things for walking** for, **for seniors**, but that’s, that’s nonsense ... **bread** ... something like that. [...] Do you think this is something similar to **bread** like ... also some **pastry**, or ...?

A: **Well, by the rest of the sentence it might have something to do with food.** It reminds me a bit of these **mobile shops that were coming to villages in the past.** [...] So, what do you think, you think that it is maybe some vehicle?

B: Well, it sounds like that, but **it doesn't fit the context for me. Some tricycle** or something.'

16.4.2. Comparison: Target words with vs. without context

The target words from the stimuli sentences with random context were also tested in a free translation experiment without context in order to obtain data for the role of context in the specific sentences. This section compares the target words in terms of how frequently they were translated correctly by the Czech respondents in the conditions with vs. without context. It was hypothesised that target words are easier to understand when presented in context than without context, as was shown to be true for sentences with highly predictable target words in section 15.

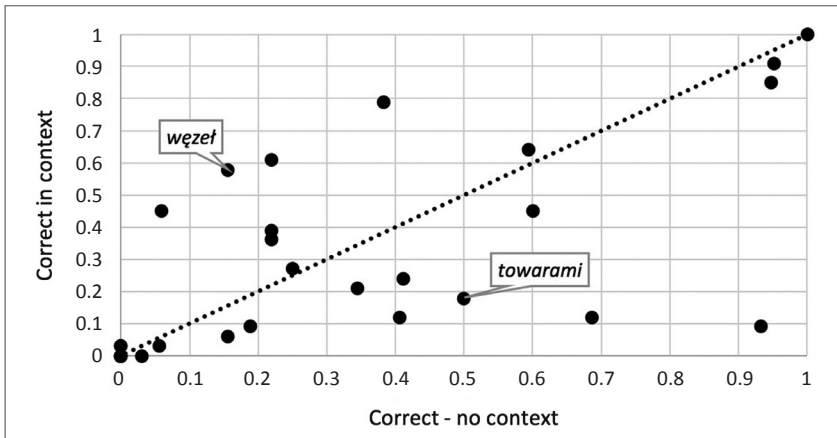


Figure 36: Comparison: target words in random context vs. without context.

Figure 36 shows the results with vs. without context. The target words were presented in their base forms in the condition without context. As a result, there is no clear overall improvement of intelligibility of the target words when compared to those in highly predictive context (compare with highly predictable target words in Figure 29). For some stimuli, the target word intelligibility improved indeed. For instance, in the sentence

Praga to ważny węzeł komunikacyjny.

'Prague is an important traffic hub.'

the intelligibility score of *węzeł* 'hub, knot' (CS *uzel*) is only 16% without context, but increases to 58% in the context presented. For other target words, intelligibility decreased when presented in context, for instance *towarami* 'goods [instr]' was translated correctly by only 18% in its inflected form in context, while without context and in its base form, it was translated correctly by 50%. In this case the lower intelligibility of the target word in context is likely to be due to the features of its inflected form. When viewing the responses provided, 55% were forms of *továrna* 'factory', which is likely to be related to the lower LD of PL *towarami* (instr pl of masc *towar* 'commodity') towards CS *továr-nami* (instr pl of fem *továrna* 'factory') than towards the correct cognate *tovary* 'goods [instr]'. The PL masculine instrumental ending *-ami* is mistaken for the feminine instrumental ending *-ami* in CS.

16.5. Summary

The error analysis of the responses provided in the cloze translation experiments (sentence stimuli from section 15 included) reveals that for all types of sentence stimuli, similarity-driven errors are the most frequent type of wrong responses even with target words in sentential context and make up almost 30% of all errors with target words in gaps at random position. For the high- and low-constraint sentences in which the target words were placed at sentence final position, the ratio is only about half as high (around 15%). The fact that the rates of similarity-driven errors do not differ significantly between high- and low-constraint sentences, but do differ between the sentences with target words at random positions and sentences with target words at sentence onset suggests that the position of the target word might be relevant. Target words at sentence final position might be easier to comprehend than at other positions in a sentence. Similarity of the target word seems to play a bigger role when the target word is at a random position in the sentence than at sentence final position. However, it has to be noted that the target words in the sentences with random context were selected because they were problematic in previous experiments. In contrast to them, the target words in the high-constraint sentences can be expected to be easier because of their predictability in context. Other wrong responses were identified as context-driven errors (around 8%-9%), errors due to associations or priming (4%-12%), which seems to play the smallest role in the sentences with highly predictable target words, and L_n interferences (all below 1%).

As for the comparison of target word intelligibility with vs. without context, there is no clear tendency in whether the random sentential context facilitates intercomprehension of the targets or not. This result differs from the result of the experiments with highly predictable target words, for which predictable context improved intelligibility by about 28% on average. Since the target words were presented in their base forms in the context-free condition, some of them might have been easier to comprehend than their inflected forms in the sentences. For instance, the PL ending *-ami* of a masculine noun in the instrumental plural was frequently mistaken for the CS ending *-ami* occurring in the instrumental plural forms of feminine nouns, leading the respondents to provide a feminine noun as a translation, which was wrong. In other cases, especially for those target words whose similarity towards their CS translations increased due to inflection in context, intelligibility in context improved.

CHAPTER VII: CONCLUSION AND OUTLOOK

This thesis is settled in the research domain on intercomprehension – the ability to understand a related but unknown foreign language without prior knowledge of this language. It examines the intelligibility of Polish stimuli presented to Czech readers in different types of translation experiments. The general focus lies on the question how stimuli-related linguistic predictors, specifically linguistic distance and surprisal as a measure of (un-)predictability in context, correlate with the intelligibility of stimuli from the different experiments. It attempts to find explanations for and patterns of failing intelligibility of certain stimuli.

After a summary of the findings from joint publications that emerged from the INCOMSLAV project in Chapter I (mainly methodology and measures of linguistic distance), the following chapters examine the experimental results. Chapter II of this thesis initially analyses the transcripts of audio recordings captured during a pairwise cooperative translation experiment in order to create a basis for a quantitative statistical analysis of the predictors for this language-reader scenario in the subsequent Chapters.

In the pairwise cooperative translation experiment, PL sentences were modified with regard to orthography, morphology, lexis, closed class words, and word order and were presented to pairs of Czech respondents who were asked to cooperatively translate these sentences. The hypothesis that the intelligibility of Polish to Czech readers can be increased by modifying a Polish sentence with certain Czech units was found true for modifications on all linguistic levels, but to different degrees. The respondent pairs were on average able to correctly translate about 74% of the words in the unmodified (original) versions of the stimuli sentences. A substitution of Polish orthographic units with their Czech correlates increased the intelligibility of the stimuli to 90%. This suggests that if Czech readers were aware of the regular orthographic correspondences and knew how to apply them to Polish text, they could understand about 90% of it (instead of only 74%). The orthographic modification led to the highest intelligibility score that could be obtained through modification on only one level, followed by the substitution of morphological affixes (88.41%), lexis (87.79%), and closed class words (85.21%). An optimisation of only word order led to the lowest increase in the share of correctly translated words (77.6%). This confirms previous findings about a limited, but existing effect of morpho-syntactic differences on mutual intelligibility of closely related languages. Results for the combined modifications on the different linguistic levels suggest that divergent lexis alone (closed class words excluded) accounts

for about 13.5% of the difficulties in Polish-Czech reading intercomprehension. This result is in line with the previously determined lexical distance of Polish to Czech readers (10%) in Jágrová, Stenger, Marti & Avgustinova (2017, p. 411) as explained in section 1.3.

The audio recordings of the pairwise translation experiment are evaluated in a qualitative manner to a great part, but it also contains quantitative analyses such as on how frequently unknown Polish characters and digraphs were pronounced in particular ways. The analysis delivered insights into many different aspects of Polish-Czech intercomprehension. Generally speaking, the difficulty of certain stimuli or words within stimuli sentences can manifest itself in:

- the use of a placeholder word for an unintelligible word in an otherwise understandable sentence,
- the order of disambiguation – difficult words are discussed only after the easier ones are understood,
- the repeated reading of the word, often with varying pronunciation.

Some of the most important outcomes of the qualitative analysis of the material recorded during the cooperative translation experiments were insights into how Czech readers handle unfamiliar characters, diacritics and digraphs. According to how the respondents read out the stimuli aloud, it could be shown when they knew the sound representation of certain Polish characters and their corresponding Czech characters, e.g. *w:v*. Respondents were able to identify and, be it consciously or not, apply some of the regular Polish-Czech correspondences for some cognates, but their application was not always consistent. Having successfully applied a correspondence once in a cognate pair did not mean that the same correspondence was recognised and applied in another cognate pair later. Common strategies how to handle unfamiliar diacritics in the Polish characters during the reading of the stimuli were to replace them by similar Czech diacritics if possible (*z* was correctly pronounced corresponding to the Czech letter *ž* in 82% of all read-out instances; *s* was pronounced corresponding to the Czech letter *š* in 65% of all instances, although this was correct in only 1 of 11 occurrences in the stimuli), to omit them (about 70% of instances for *q* and *ę*) or move them to another position or base letter in the word (about 11% for *q* and *ę*). This moving to another position sometimes resulted in a palatalisation of one of the consonants within a word (another 7% of all cases).

Respondents pronounced the letters *q* and *ę* as nasals in about 12% of all occurrences, although this pronunciation might have been not entirely correct. Some mistakes in translations turned out to be due to a wrong interpretation of the PL digraphs. Although there is a clear tendency that the digraphs *cz* and *sz* are pronounced correctly in about 80% of all cases, respondents pronounced *cz*

and *sz* as /ts/, /s/ or /s+/z/ in about 20% of all read-out instances, not recognising the regular correspondences *cz*:č and *sz*:š. This effect was even stronger for the digraph *rz* – respondents did not recognise the regular correspondence *rz*:ř in about 42% of all read-out cases and pronounced these stimuli with a syllable division between *r* and *z*, which led to incorrect translations in some cases.

Regarding lexis, the results show that respondents distrusted internationalisms within the stimuli sentences, although, on the one hand, these words (nearly) identically exist in Czech, but on the other hand, most of them are also very infrequent in Czech. This distrust seems to be specific for internationalisms only and was not observed for Slavic vocabulary with similarly low orthographic distance. The sentences contained a number of false friends and respondents explicitly mentioned being aware of some of these false friends and even of some non-cognates (e.g. *samochód*) owing to incidental learning.

Respondents were successfully able to draw lexical inferences from German, English, Russian, and Slovak for words within the stimuli for which no Czech cognates exist. They were also able to handle morphosyntactic differences that they could infer from other previously acquired languages. However, also cases of negative transfer that led to wrong translations were observed not only from Czech, but also from German, English, Croatian, and Slovak.

Building on the numerous insights from the pairwise cooperative translation experiments, hypotheses about the role of possible predictors are formulated in Chapter III. The first main hypothesis proposes that a pronunciation-based Czech and Slovak distance measure of word pairs (pron LD) that represents the actually **perceived** distance of the stimuli correlates better with intelligibility than Levenshtein distance does when the latter is calculated in the traditional way (trad LD). This measure takes into account that Czech readers are aware of the actual pronunciation of certain Polish characters, e.g. Polish *w* corresponding to Czech *v* which are treated as two different items in a traditional way of calculating Levenshtein distance, but in reality do not pose any obstacle for intercomprehension. It also assumes that Czech respondents have good receptive skills in Slovak, treating them as Czech and Slovak receptive bilinguals.

In order to gather a significant amount of data to test this hypothesis, web-based translation experiments with different types of stimuli were conducted. The stimuli presented in the web-based translation experiments were:

- individual words without context,
- noun phrases with two different linearisations, and
- target words in sentential context, presented as a cloze translation task.

There were different types of individual word stimuli in the free translation experiments without context:

- cognates containing regular Polish-Czech correspondences,
- the 100 most frequent Polish nouns, and
- individual target words from the stimuli sentences presented in the cloze translation task.

The results of the free translation experiments with individual words were analysed for a correlation with linguistic distance. It could be shown that pron LD can explain 39.8% of the variation in the data and correlates stronger with intelligibility than trad LD which can account for only 21.7% of the variation. The mean intelligibility of the cognates containing regular Polish-Czech correspondences proved to be 66.7%. As for the most frequently applicable Polish-Czech correspondences identified in Fischer et al. (2015), words containing correspondences that only required a tolerance of diacritical signs by the reader were highly intelligible, most of them reaching ceiling effect. The same applies to words whose pronunciation can be assumed clear to Czech readers. Among the most frequent Polish-Czech correspondences of characters (digraph correspondences not included), the *ć:t* correspondence turned out to be the most problematic. Polish *ć*, which regularly corresponds to the Czech *t* in infinitive verb endings, is relatively difficult to recognise because of its misleading orthographic and phonological similarity to Czech *č*. Czech respondents frequently mistook infinitive verb forms for masculine nouns ending in *-c* or *-č*, depending on the availability of such orthographic neighbours in Czech. This finding is in line with the strategy of replacing diacritics with a similar Czech diacritic or omitting diacritics, as observed in the cooperative translation experiments. Accordingly, the intelligibility of the verbs within the stimuli set ($n = 35$) was relatively low (38.8%). Monosyllabic infinitive verb forms with applicable correspondences in the stem proved to be especially difficult to comprehend. The application of the *a:e* correspondence appears to be difficult in stems of verbs as well as in other rather short words. However, *a:e* did not pose any problems in feminine noun endings of internationalisms. The *g:h* correspondence was largely applied successfully, again depending on the available orthographic neighbours.

Among the 100 most frequent Polish nouns, there were 16 nouns with identical Czech translation equivalents which were therefore not part of the stimuli set. The mean intelligibility of the remaining 84 nouns proved to be 55.03%. Under the assumption that the 16 identical nouns are entirely intelligible to Czech respondents, one can speak of an overall intelligibility of about 71% for the whole set of the 100 most frequent Polish nouns. The results

confirm the finding that pron LD correlates more strongly with intelligibility than trad LD: pron LD can explain 45% of the variation in the data, while *trad LD* can account for only 37%. Binary predictor variables for different grammatical gender of the words in the two languages and for stimuli that proved to be false friends were added in a regression analysis, since these turned out to be important factors impairing intelligibility in the subsequent cloze translation experiments (Chapter VI). A regression model with the variables pron LD and false friends can account for 58.2% of the variation in intelligibility of the 100 most frequent Polish nouns. Among all **wrong** responses, about 21% could be identified as due to interferences from English, German, Slovak, Serbo-Croatian, or Bulgarian.

A predictor on the context level was added to the analysis of the results for noun phrases in Chapter V and sentence stimuli in Chapter VI: surprisal as a measure of greater difficulty due to divergent word order in the noun phrases and (un)predictability of target words in the sentences.

Concerning the noun phrases, it was hypothesised that noun phrases with noun-adjective linearisation, which is not as typical in Czech as it is in Polish, should be more difficult to guess than the same noun phrases with adjective-noun linearisation. This should reflect in a lower number of correct translations and higher processing times in the noun-adjective condition. When viewing the whole data set of 1293 phrases in the adjective-noun condition and 1296 phrases in the noun-adjective condition, noun phrases with adjective-noun linearisation were translated slightly more often correctly than those with noun-adjective linearisation (49.5% > 41.63%). However, when viewing only the data for the most representative 30 noun phrases (at least 10 data points per phrase and condition), their difference in intelligibility of less than 3% is not statistically significant.

Also, the mean processing times of correctly translated noun phrases do not differ significantly between the two conditions. The correlations between processing time and the possible predictors are all very low. When viewing surprisal as a separate factor influencing processing time of correctly translated noun phrases, a weak but significant correlation could only be discovered for the noun-adjective condition. The highest correlation found for processing time was with a measure referred to as *overall difficulty* (unifying pron LD, lexical distance and surprisal) for both linearisations ($r = 0.259$, $p < 0.001$ for adjective-noun and $r = 0.194$, $p < 0.001$ for noun-adjective). In a regression analysis, total distance (unification of lexical distance and *pron LD*) and the sums of surprisal of the noun phrases account for 58% of the variation in the intelligibility of the noun phrases. This model has a slightly stronger correlation in the noun-adjective condition than in the adjective-noun condition.

The latter finding suggests that predictability effects become more relevant in the noun phrases with noun-adjective linearisation, i.e. in phrases with rather unusual word order, than in noun phrases with regular word order. Nevertheless, the difference between the two conditions is found to manifest itself most strongly in an analysis of the wrong answers: Respondents failed to correctly recognise the part of speech of the stimuli in noun-adjective linearisation about 2.6 times more often than in adjective-noun linearisation.

In a digression, a similarly designed experiment with German respondents who were also asked to translate noun phrases in adjective-noun and noun-adjective linearisation is touched upon. The phrases in this experiment consisted of internationalisms and Indo-European cognates. The results confirm the greater difficulty of the noun-adjective condition, although the difference is again rather small (intelligibility of adjective-noun 29.84% > noun-adjective 26.82%). Similar to the results of the experiment with Czech readers, the error analysis reveals that the greater difficulty of the noun-adjective linearisation is best reflected in the number of wrongly recognised part of speech in this condition. It was also found that the intelligibility scores correlate stronger with Germanic distance than with a purely German-to-Polish distance, assuming that respondents are DE and EN bilinguals at least on a receptive level. The fact that almost a third of all noun phrases were translated correctly without any respondents' prior knowledge of Polish can serve as an argument for a long-established principle in modern foreign language teaching: Lessons can be held in the target language from the beginning on, since learners can build upon their knowledge of previously acquired languages.

Both Czech and German respondents translated internationalisms about 3 times more often correctly than other cognates with the same orthographic distance. Although this might not seem surprising, results from the cooperative translation experiment revealed that Polish internationalisms that have infrequent Czech cognates caused problems with Czech readers when translating sentences (section 5).

In the cloze translation experiments in Chapter VI, two kinds of sentence stimuli are discussed: sentences with highly predictable target words in sentence final position and sentences with target words in random context – these were constructed for the cooperative translation experiment. The latter were additionally tested in cloze translation design for an additional and more representative sample than in the pairwise translation experiment.

The cloze translation experiments with highly predictable target words are the part that unifies all previous hypotheses and can, together with the cooperative translation experiments, be considered the most important chapter of this thesis, since all relevant factors come to play here. Context proved to help the

correct disambiguation of a great part of target words, although not all. The mean intelligibility of target words improved significantly from 49.71% without context to 67.99% in highly predictive context. Among the cases where context did not help were those target words that differ only in their inflected forms, while their base forms are identical in the two languages. This was especially the case with feminine nouns in accusative and instrumental case ending in *-ę* and *-ą*. Here, L1 interferences on the morphological level come into play. Feminine accusative forms ending in *-ę* were often translated with CS plural forms or other inflected forms ending in *-e* or *-ě*. Feminine instrumental forms ending in *-ą* were frequently translated with feminine nominative forms. Also, verb forms ending in *-ą* were mistaken for feminine nouns in the nominative case. In comparison to the condition without context, the predictive context also significantly helped the correct identification of POS, especially with infinitive verb forms. Infinitives were more frequently mistaken for nouns without context than in context. Furthermore, target nouns with divergent grammatical gender in the two languages proved to be more problematic than target nouns with identical grammatical gender. There proved to be the tendency that respondents maintained the grammatical gender of the stimulus in their response, which in these cases resulted in a wrong translation. Context, however, facilitated intelligibility of these target words significantly: The mean increase in intelligibility for words with divergent grammatical gender in the two languages is 28.3% as compared to the condition without context, which is about 10% more than the average increase in intelligibility for the whole dataset.

Predictability in context has a greater positive impact on non-cognates and false friends than on cognates. In accordance with this, surprisal turned out to be no good predictor for the intelligibility of cognates in context – during the disambiguation of these, respondents seem to rather rely on similarity than on context. However, if cognates are excluded from the regression analysis, surprisal has a decent correlation with the intelligibility of non-cognates (including false friends) and an even higher one with false friends only. Semantic associations with the target word itself or with another word in the stimulus sentence turned out to have a great potential to lead the respondent towards a correct understanding and to increase intelligibility in context. False friends that are also cognates and non-cognates that allow for associations with the correct translations of the stimuli were among those words for which intelligibility increased the most in context. The total number (more than the percentage) of non-cognates and false friends per sentence has a strong negative correlation with intelligibility ($r = -0.508$), as this crucially influences how much of the context readers actually understand.

No clear tendency could be observed in whether a random sentential context facilitates intercomprehension of target words at any position in sentences in comparison to no context. The error-analytical results suggest that similarity-driven errors are the most frequent type of wrong responses even with target words in sentential context and make up almost 30% of all errors with target words in gaps at random position. For sentences in which the target words were placed at sentence final position, the ratio is only about half as high (around 15%). The results suggest that also the position of the target word might be relevant for intelligibility. Target words at sentence final position might be easier to comprehend than at other positions in a sentence. This, however, should be subject to systematic future studies. Other wrong responses were identified as context-driven errors (around 8%-9%), errors due to associations or priming (4%-12%), and L_n interferences (below 1%). Errors due to association or priming seem to play the smallest role in the sentences with highly predictable target words (4% of all errors).

A logical follow-up of this thesis would be to test the same stimuli in Czech and present them to Polish readers in order to account for the asymmetries in intelligibility, also in relation to the predictors conditional entropy and word adaptation surprisal (WAS) discussed in section 1.4.2. The phenomenon of the inner speech during reading could also be observed from the Polish readers' perspective. It would be very interesting if the same procedures (frequent ignoring or moving of unknown diacritics to other letters) apply for the scenario Czech translated by Polish readers. In addition, it would be interesting how Polish readers in general pronounce the same Czech stimuli in order to establish a pronunciation-based orthographic distance measure for this direction of reading, accordingly. Further effort in the investigation on the assumed pronunciation of Czech (and also Polish, Croatian, and Serbian) stimuli was already invested by examining audio recordings of Russian and Serbian students reading individual Pan-Slavic cognates aloud (Jágrová & Stenger, 2019).

Instead of presenting readers with random constructed sentences in the cooperative translation experiments, it could have been more appropriate to present the sentences with highly predictable target words (those discussed in section 15). The same applies for the noun phrase stimuli – they could have been extracted from the same sentences. In an ideal world, data for all words occurring in these sentences could have been gathered in the free translation experiment of individual words so that false friends within the sentences (not only among the target words) could be more reliably identified.

There might be reason for criticism about the design of the stimuli with the combined modifications (section 4), especially for those sentences that did not already cause greater problems in the unmodified condition. For some stimuli, the ceiling effect (stimulus is too easy, informants' answers are close to 100% correct) could be observed. Some respondents noticed that there were Czech characters in the modified stimuli. These might be arguments for not including such modified stimuli into a cooperative translation experiment of this type.

Free translation experiments with the most frequent nouns in the language combinations Bulgarian, Czech, Polish, and Russian were already conducted over the experiment website and the results are being analysed now. The ultimate goal for the future is to establish an interactive Slavic intercomprehension matrix (Jágrová, Stenger & Avgustinova, 2019) consisting of the different predictors and intelligibility scores obtained from experiments in as many Slavic language-reader combinations as possible. Of course, the same methodology and the experiment website can be applied to any language combination outside of the Slavic language family.

With regard to the target words in highly predictable stimuli sentences, it would be very interesting how these words behave in varying, maybe even misleading contexts. Such misleading items could be, again, words directly preceding the target word or dominant concepts at other positions in the sentences that would lead to wrong semantic associations. Another option could also be to present target words not in a sentential context, but in a visual context (helpful or misleading) or in spoken modality (written and spoken stimuli separately and both at once).

As of March 2019, an e-learning functionality was added to the experiment website that allows participants to re-do their experiments several times. The software records the experiment statistics of the participants and displays a learning curve after each repeated experiment. This could be interesting for beginning learners of a language or for students of multilingual Slavic language courses, since the functionality offers to do an experiment at the beginning of a course and the same experiment again after successful completion of the course in order to track one's progress.

The presented experimental setting and the insights gained in this thesis can be beneficial in different areas. They are obviously relevant for all situations in which Czech native speakers are confronted with written Polish. A specifically relevant field here is foreign language acquisition – when Czech native speakers are learning Polish, whether it be in a formal or informal setting. Besides the holistic approach in foreign language acquisition, recent developments are giving rise to the question of focussing rather on partial competences than on an excellent command of an Ln. In the EU brochure *Studies on translation*

and multilingualism, it is explicitly mentioned that “with the development of new theories on foreign language learning, such as the concept of partial competences, intercomprehension gave hope that learners could develop at least some understanding of the languages belonging to the same family” (European Commission, 2012, p. 6). Hence, it is possible to reach satisfactory reading skills in a genetically so closely related language or also in other languages of the same language family with relatively low effort. From the results of this thesis, one could conclude the following two learning strategies: (i) Acquisition of the regular crosslingual correspondences on the orthographic (also on the phonetic and morphological) level in order to recognise cognates and word fragments and (ii) the mediation of frequent non-cognates and false friends in helpful sentential contexts. This could be one of the possible next steps and contributions to present-day intercomprehension research and didactics that can build up on this thesis.

Deutsche Zusammenfassung

Diese Arbeit ist auf dem Gebiet der Interkomprehensionsforschung angesiedelt. Interkomprehension wird definiert als die Fähigkeit, eine verwandte Sprache ohne Vorkenntnisse in dieser Sprache zu verstehen. Die Arbeit untersucht die Verständlichkeit verschiedener polnischer Stimuli, die tschechischen Versuchspersonen in unterschiedlichen Arten von Übersetzungsexperimenten präsentiert wurden. Der allgemeine Fokus dieser Arbeit liegt auf der Fragestellung, wie stimulusbezogene sprachliche Faktoren, insbesondere sprachliche Distanz und Surprisal als ein Maß für die (Un-)Vorhersehbarkeit im Kontext, mit der Verständlichkeit von Stimuli in den einzelnen Experimenten korrelieren. Die Arbeit versucht auch Erklärungen für geringe Verständlichkeit bestimmter Stimuli zu finden und Muster anhand von Fehleranalysen aufzudecken.

Nach einer Zusammenfassung von Erkenntnissen aus Publikationen in Kapitel I (vor allem Methoden und Maße der sprachlichen Distanz), die im Rahmen des Projekts INCOMSLAV entstanden sind, widmen sich die folgenden Kapitel den Ergebnissen der Experimente. In Kapitel II werden zunächst die Transkripte von Audioaufnahmen, die während eines paarweisen kooperativen Übersetzungsexperiments aufgezeichnet wurden, qualitativ analysiert, um eine Grundlage für eine quantitative Analyse der für dieses polnisch-tschechische Interkomprehensions-Szenario relevanten Faktoren in den nachfolgenden Kapiteln zu schaffen.

In den kooperativen Übersetzungsexperimenten in Paaren wurden polnische Sätze hinsichtlich ihrer Orthographie, Morphologie, Lexis, Wörter der geschlossenen Wortklassen sowie Wortfolge modifiziert und Paaren von Versuchspersonen präsentiert. Diese hatten die Aufgabe, diese Sätze zunächst laut vorzulesen und dann innerhalb von jeweils fünf Minuten eine schriftliche Übersetzung dieser Sätze einzugeben. Die Hypothese, dass die Verständlichkeit des Polnischen für tschechische Lesende erhöht werden kann, indem polnische Sätze durch bestimmte tschechische Einheiten modifiziert werden, konnte für Modifikationen auf allen sprachlichen Ebenen bestätigt werden, jedoch zu unterschiedlichen Graden. Die Versuchspersonenpaare waren in der Lage, durchschnittlich 74% der Wörter der unmodifizierten Sätze (im polnischen Original) zu übersetzen. Ein Ersetzen polnischer orthographischer Einheiten durch ihre tschechischen Entsprechungen führte zu einer Steigerung der Verständlichkeit der Stimuli auf 90%. Dies kann bedeuten, dass wenn tschechische Lesende sich der regelmäßigen orthographischen Entsprechungen zwischen den beiden Sprachen bewusst wären und diese auf polnischen Texten spontan anwenden könnten, sie etwa 90% davon auf Anhieb verstehen würden. Die Modifikationen auf orthographischer Ebene führten zu den

höchsten Verständlichkeitswerten, die durch Modifikationen auf nur einer sprachlichen Ebene erreicht werden konnten, gefolgt vom Ersetzen morphologischer Einheiten (88,41%), Lexis (87,79%) und Wörtern aus geschlossenen Wortklassen (85,21%). Eine Anpassung der Wortfolge allein führte zur geringsten Steigerung des Anteils an korrekt übersetzten Wörtern (77,60%). Dies bestätigt vorherige Erkenntnisse auf dem Gebiet der Interkomprehensionsforschung über einen eingeschränkten, aber doch existierenden Effekt von morpho-syntaktischen Unterschieden auf die Interkomprehension. Die Ergebnisse für Kombinationen von Modifikationen auf verschiedenen sprachlichen Ebenen zeigen, dass allein die Unterschiede in der Lexik (Wörter aus geschlossenen Klassen ausgenommen) etwa für 13,5% der Verständnisprobleme bei der polnisch-tschechischen Interkomprehension im Lesen verantwortlich sind. Dieses Ergebnis unterscheidet sich nur geringfügig von der in Jágrová, Stenger, Marti & Avgustinova (2017, S. 411) gemessenen lexikalischen Distanz des Polnischen für tschechische Lesende (10%), worauf in Abschnitt 1.3 eingegangen wird. Die Audioaufnahmen der kooperativen Übersetzungsexperimente bieten Einsicht in eine Reihe von Aspekten der polnisch-tschechischen Interkomprehension und werden größtenteils qualitativ ausgewertet, enthalten aber auch quantitative Analysen, z. B. darüber, wie häufig unbekannte polnische Zeichen und Digraphen auf eine bestimmte Art und Weise ausgesprochen wurden. Im Allgemeinen kann sich die Schwierigkeit bestimmter Stimuli oder einzelner Wörter innerhalb der Sätze manifestieren in:

- der Verwendung von Platzhalterwörtern, die für unverständliche Wörter in einem ansonsten verständlichen Kontext eingesetzt werden,
- der Reihenfolge der Disambiguierung: über schwierige Wörter wird diskutiert, nachdem die einfacheren verstanden worden sind,
- dem wiederholten Vorlesen der Wörter, oft mit variierender Aussprache.

Zu den wichtigsten Ergebnissen der qualitativen Analyse des Audiomaterials aus den kooperativen Übersetzungsexperimenten zählen Erkenntnisse darüber, wie tschechische Lesende mit unbekanntem Buchstaben, Diakritika und Digraphen umgehen. Je nachdem, wie die Versuchspersonen die Stimuli laut vorgelesen haben, konnte gezeigt werden, dass sie die Lautrepräsentation bestimmter polnischer Buchstaben und der ihnen im Tschechischen entsprechenden Buchstaben kennen, z. B. *w:v*. Versuchspersonen waren in der Lage, ob bewusst oder unbewusst, einige der regulären polnisch-tschechischen Korrespondenzen in einigen Kognaten anzuwenden, jedoch war die Anwendung dieser Korrespondenzen nicht immer konsistent. Eine erfolgreiche Anwendung einer Korrespondenz in einem Kognatenpaar bedeutete nicht, dass dieselbe Korrespondenz später in einem anderen Kognatenpaar angewandt wurde. Zu den

üblichen Strategien, wie mit unbekanntem Diakritika im Polnischen während des Vorlesens umgegangen wurde, zählte das Ersetzen der Diakritika durch ähnliche tschechische Diakritika, wenn möglich (das polnische *ź* wurde der Entsprechung nach korrekt wie das tschechische *ž* in 82% aller vorgelesenen Fälle ausgesprochen; *ś* wurde wie *š* in 65% aller Fälle ausgesprochen, obwohl das nur in 1 von 11 Fällen, in denen *ś* vorkam, korrekt war), das Ignorieren der Diakritika (etwa 70% der Fälle von *q* und *ę*) sowie ihre Verschiebung auf eine andere Buchstabenbasis im Wort (etwa 11% bei *q* und *ę*). Diese Verschiebung führte in einigen Fällen zu einer Palatalisierung von Konsonanten innerhalb eines Wortes (weitere 7% aller Fälle).

In etwa 12% aller vorgelesenen Fälle wurden *q* und *ę* als Nasale ausgesprochen, obwohl diese Aussprache nicht immer gänzlich korrekt war. Einige Übersetzungsfehler sind auf eine falsche Interpretation der polnischen Digraphen zurückzuführen. Obwohl eine klare Tendenz zur korrekten Aussprache der Digraphen *cz* und *sz* (80% korrekt ausgesprochen) zu erkennen ist, wurden diese in etwa 20% aller Fälle wie /ts/, /s/ oder /s+/z/ ausgesprochen, wobei die regulären Korrespondenzen *cz*:*č* oder *sz*:*š* nicht erkannt wurden. Dieser Effekt war noch stärker im Falle des Digraphen *rz*: In etwa 42% aller Fälle wurde die Korrespondenz *rz*:*ř* nicht erkannt, was sich darin zeigte, dass diese Stimuli mit einer Silbentrennung zwischen *r* und *z* ausgesprochen wurden, was in einigen Fällen zu falschen Übersetzungen führte.

Bezüglich der Lexik zeigen die Ergebnisse, dass die Versuchspersonen eine gewisse Skepsis gegenüber Internationalismen innerhalb der Stimuli pflegten, obwohl diese Wörter (nahezu) identisch im Tschechischen existieren. Dies scheint besonders für solche Internationalismen zu gelten, deren übersetzungsäquivalente Internationalismen im Tschechischen eher selten in den Sprachgebrauch einfließen, z. B. *rektor* ‚Rektor‘ oder *brutto* ‚Brutto‘. Diese Skepsis scheint spezifisch für Internationalismen zu gelten, denn sie konnte nicht für slavische Kognaten mit ähnlich geringer orthographischer Distanz beobachtet werden. Die Stimulusätze beinhalteten eine Reihe von falschen Freunden und überraschenderweise äußerten manche Versuchspersonen direkt, dass sie sich einiger falscher Freunde und sogar Nicht-Kognaten (z. B. *samochód* ‚Auto‘) bewusst sind, etwa als Folge von informellem oder inzidentellem Lernen.

Die Versuchspersonen waren in der Lage, lexikalische Kenntnisse des Deutschen, Englischen, Russischen und Slowakischen zu aktivieren und die Bedeutung solcher polnischer Stimuli zu inferieren, für die im Tschechischen keine Kognaten existieren. Sie waren auch in der Lage, mit morphosyntaktischen Phänomenen umzugehen, die sie aus anderen zuvor erworbenen Sprachen kannten. Jedoch konnten auch Fälle von negativem (falschem) Transfer, nicht nur aus dem Tschechischen, sondern auch aus dem Deutschen, Englischen, Kroatischen und Slowakischen, nachgewiesen werden.

Aufbauend auf den zahlreichen Ergebnissen aus den kooperativen Übersetzungsexperimenten in Paaren konnten Hypothesen für die Rolle möglicher Prädiktoren in Kapitel III formuliert werden. Die erste Haupthypothese ist, dass ein auf Aussprache basierendes, tschechisch-slovakisches Distanzmaß von Wortpaaren (*pron LD*), das die eigentlich **wahrgenommene** Distanz repräsentiert, besser mit der Verständlichkeit der Stimuli korreliert als die traditionell berechnete Levenshtein-Distanz (*trad LD*). Mit der Anwendung dieses Maßes soll auch beachtet werden, dass tschechische Lesende sich der eigentlichen Aussprache bestimmter polnischer Buchstaben bewusst sind (z. B. des polnischen *w*, das dem tschechischen *v* entspricht), die bei der traditionellen Art der Bestimmung der Levenshtein-Distanz als unterschiedliche Buchstaben behandelt werden, in der Realität aber keine Hürde für das Verständnis darstellen. Dieses Maß beachtet auch, dass tschechische Lesende gute rezeptive Fähigkeiten des Slovakischen besitzen und als tschechisch-slovakische Bilinguale, zumindest auf rezeptiver Ebene, betrachtet werden sollten. Es setzt voraus, dass solche polnisch-tschechischen Entsprechungen, die mit slovakisch-tschechischen identisch sind (z. B. *ie:e*), auch kein Hindernis darstellen.

Um eine signifikante Menge an Daten zu sammeln und diese Hypothese zu testen, wurden web-basierte Experimente mit unterschiedlichen Stimuli durchgeführt. Die in den web-basierten Experimenten präsentierten Stimuli waren:

- einzelne Wörter ohne Kontext,
- Nominalphrasen mit zwei unterschiedlichen Wortfolgen und
- Zielwörter im Satzkontext als Lückentext-Übersetzungsaufgabe.

Bei den Übersetzungsexperimenten mit einzelnen Wörtern ohne Kontext gab es drei unterschiedliche Arten von Stimuli:

- Kognaten mit anwendbaren polnisch-tschechischen Korrespondenzen,
- die 100 häufigsten polnischen Substantive und
- die Zielwörter aus den Satzstimuli der Lückentext-Übersetzungsaufgabe.

Die Ergebnisse der Übersetzungsexperimente mit einzelnen Wörtern ohne Kontext wurden auf eine Korrelation mit sprachlicher Distanz hin untersucht. Es konnte gezeigt werden, dass *pron LD* 39,8% der Varianz in den Daten erklären kann und stärker mit der Verständlichkeit korreliert als *trad LD*, welche nur 21,7% erklären kann. Die durchschnittliche Verständlichkeit der Kognaten mit anwendbaren polnisch-tschechischen Korrespondenzen lag bei 66,7%. Wenn man diejenigen Korrespondenzen genau betrachtet, die in Fischer et al. (2015) als die häufigsten polnisch-tschechischen identifiziert wurden, dann waren die Wörter, die nur Korrespondenzen enthielten, die eine Toleranz der Diakritika verlangen, sehr gut verständlich. Als problematischste Korrespondenz (Korrespondenzen mit Digraphen ausgenommen) hat sich *ć:t* erwiesen.

Die Korrespondenz des polnischen *ć*, welches regelmäßig dem tschechischen *t* in Endungen infinitiver Verbformen entspricht, ist relativ schwer als solche zu erkennen, nicht zuletzt wegen der orthographischen und phonologischen Ähnlichkeit des polnischen *ć* mit dem tschechischen *č*. Aus diesem Grund haben die tschechischen Versuchspersonen Infinitive häufig mit maskulinen Substantiven übersetzt, die auf *-c* oder *-č* enden, je nach Verfügbarkeit solcher orthographischer Nachbarn im Tschechischen. Diese Erkenntnis geht einher mit der in den kooperativen Übersetzungsexperimenten beobachteten Strategie, Diakritika in den polnischen Stimuli mit ähnlichen tschechischen Diakritika zu ersetzen oder sie zu ignorieren. Dementsprechend gering ist die durchschnittliche Verständlichkeit der Verben ($n = 35$) innerhalb der Stimuli ausgefallen (38,8%). Einsilbige Infinitive mit anzuwendenden Korrespondenzen im Wortstamm erwiesen sich als besonders schwer verständlich. Die Anwendung der Korrespondenz *a:e* in Wortstämmen von Verben und anderer eher kurzer Wörter hat sich als schwierig erwiesen, während dieselbe Korrespondenz in den Endungen femininer Internationalismen keine Probleme verursachte. Die Korrespondenz *g:h* wurde größtenteils erfolgreich angewandt, allerdings wiederum abhängig vom Vorhandensein möglicher orthographischer Nachbarn als konkurrierender Übersetzungsvarianten.

Zu den 100 häufigsten polnischen Substantiven zählen auch 16 Substantive, die mit ihren tschechischen Übersetzungsäquivalenten identisch sind. Diese wurden deshalb nicht im Experiment getestet. Die durchschnittliche Verständlichkeit der restlichen 84 Substantive beträgt 55,0%. Unter der Annahme, dass die 16 identischen Substantive zu 100% verständlich sind, kann man von einer Verständlichkeit der 100 häufigsten polnischen Substantive von ca. 71% sprechen. Die Ergebnisse bestätigen zudem die Hypothese, dass die aussprachebasierte Distanz *pron LD* stärker mit der Verständlichkeit korreliert als die auf traditionelle Weise gemessene orthographische Distanz *trad LD*: *pron LD* kann 45% der Varianz in den Daten erklären, während *trad LD* nur 37% erklären kann. Außer *pron LD* wurden die Variable für unterschiedliches grammatisches Geschlecht der Wörter in den beiden Sprachen sowie eine Variable für die Kategorie der falschen Freunde zur Regressionsanalyse hinzugezogen, denn diese Variablen stellten sich in den späteren Lückentext-Übersetzungsaufgaben als relevant heraus (Kapitel VI). Während für diese Stimuli keine Korrelation mit der Variable des unterschiedlichen grammatischen Geschlechts festgestellt werden konnte, war die Variable falsche Freunde hier relevant. Das Regressionsmodell mit den Variablen *pron LD* und falsche Freunde kann 58,2% der Varianz der Verständlichkeit der 100 häufigsten polnischen Substantive erklären. Unter allen falschen Antworten konnten außerdem etwa 21% als Interferenzen aus dem Englischen, Deutschen, Slowakischen, BKMS oder dem Bulgarischen identifiziert werden.

Eine Prädiktorvariable im Bereich des Kontexts wurde der Analyse der Experimente mit Nominalphrasen in Kapitel V und den Satzstimuli in Kapitel VI hinzugefügt: Surprisal als ein Maß für größere Schwierigkeit aufgrund von unterschiedlicher Wortfolge in den Nominalphrasen und (Un-)Vorhersehbarkeit von Zielwörtern in Sätzen.

Bezüglich der Nominalphrasen wurde die Hypothese aufgestellt, dass solche Nominalphrasen mit Substantiv-Adjektiv-Folge, welche im Tschechischen nicht so typisch ist wie im Polnischen, schwerer zu verstehen sein müssten als dieselben Phrasen mit Adjektiv-Substantiv-Folge. Dies sollte sich in der niedrigeren Anzahl korrekter Übersetzungen und höher Bearbeitungszeit in der Substantiv-Adjektiv-Kondition manifestieren. Beim Vergleich der 1293 Datenpunkte in der Adjektiv-Substantiv-Kondition und 1296 Datenpunkte in der Substantiv-Adjektiv-Kondition wurden die ersteren etwas häufiger korrekt übersetzt als die in der Substantiv-Adjektiv-Kondition (49,5% > 41,63%). Wenn man jedoch nur die Daten der 30 repräsentativsten Phrasen betrachtet (mit mindestens 10 Datenpunkten pro Phrase und Wortfolge), dann beträgt die Differenz ihrer Verständlichkeit weniger als 3% und ist statistisch nicht signifikant. Auch die durchschnittliche Bearbeitungszeit der korrekt übersetzten Phrasen unterscheidet sich in beiden Konditionen nicht signifikant. Die Korrelationen zwischen Bearbeitungszeit und möglichen Prädiktoren sind alle sehr gering. Wenn man Surprisal als separate Variable betrachtet, dann existiert eine schwache, aber signifikante Korrelation mit der Bearbeitungszeit der korrekt übersetzten Phrasen in der Substantiv-Adjektiv-Kondition. Die stärkste, aber dennoch sehr niedrige Korrelation mit Bearbeitungszeit wurde für die Variable *overall difficulty* (,Gesamtschwierigkeit', Vereinigung von *pron LD* und Surprisal) in beiden Konditionen gefunden ($r = 0.259$, $p < 0.001$ für Adjektiv-Substantiv und $r = 0.194$, $p < 0.001$ für Substantiv-Adjektiv). Die in einem Regressionsmodell zusammengeführten Variablen *total distance* (,Gesamtdistanz', Vereinigung von lexikalischer Distanz und *pron LD*) und die Summe des Surprisals pro Phrase können 58% der Varianz in der Verständlichkeit der Nominalphrasen erklären. Dieses Modell hat eine etwas stärkere Korrelation in der Substantiv-Adjektiv-Kondition als in der Adjektiv-Substantiv-Kondition. Dieses Ergebnis deutet darauf hin, dass Effekte der Vorhersehbarkeit bei Nominalphrasen mit Substantiv-Adjektiv-Folge, d.h. in Phrasen mit eher ungewohnter Wortfolge, relevanter sind als in Phrasen mit gewöhnlicher Wortfolge. Nichtsdestotrotz scheint sich der Unterschied in der Schwierigkeit zwischen beiden Konditionen am stärksten in der Fehleranalyse zu zeigen: Die Versuchspersonen haben in der Substantiv-Adjektiv-Kondition etwa 2,6-mal häufiger die Wortart der Stimuli falsch erkannt als in der Adjektiv-Substantiv-Kondition.

In einem Exkurs werden ähnlich gestaltete Experimente mit deutschen Versuchspersonen ausgewertet, die auch die Aufgabe hatten, polnische Nominalphrasen in Adjektiv-Substantiv und Substantiv-Adjektiv-Folge zu übersetzen. Die Phrasen in diesem Experiment bestanden aus polnischen Internationalismen und indoeuropäischen Kognaten. Die Ergebnisse bestätigen die größere Schwierigkeit der Substantiv-Adjektiv-Kondition, obwohl die Differenz zwischen beiden Konditionen wieder eher gering ausfällt (Verständlichkeit von Adjektiv-Substantiv 29,84% > 26,82% von Substantiv-Adjektiv). Ähnlich wie bei den Experimenten mit den tschechischen Versuchspersonen fällt bei der Fehleranalyse auf, dass sich die größere Schwierigkeit der Substantiv-Adjektiv-Folge am stärksten in der Anzahl falsch erkannter Wortarten in dieser Kondition aufzeigt. Außerdem korreliert die Verständlichkeit stärker mit einer *Germanic distance* (‘Germanischen Distanz’), die davon ausgeht, dass die Versuchspersonen zumindest auf rezeptiver Ebene als deutsch-englisch Bilinguale zu betrachten sind, als mit einer rein polnisch-deutschen Distanz. Die Tatsache, dass nahezu ein Drittel der Nominalphrasen korrekt übersetzt wurden, ohne dass die Versuchspersonen über Vorkenntnisse des Polnischen verfügten, kann als Argument für ein im Fremdsprachenunterricht lange geltendes Prinzip gelten: Der Unterricht kann von Anfang an in der Zielsprache abgehalten werden, denn Lernende können sich auf ihre Kenntnisse bereits erworbener Sprachen stützen.

Sowohl die tschechischen als auch die deutschen Versuchspersonen übersetzten Internationalismen etwa dreimal häufiger korrekt als andere Kognaten mit derselben orthographischen Distanz. Obwohl dies nicht überraschend zu sein scheint, zeigen die Ergebnisse der kooperativen Übersetzungsexperimente, dass die in den Satzstimuli vorkommenden polnischen Internationalismen, die mit im Tschechischen wenig frequenten Kognaten übersetzt werden können, zu Schwierigkeiten führten (Abschnitt 5).

In Kapitel VI werden zwei Arten von Satzstimuli ausgewertet, die in Lückentext-Übersetzungsexperimenten präsentiert wurden: Sätze mit sehr vorhersehbaren Zielwörtern am Satzende und Sätze mit Zielwörtern an unterschiedlichen Stellen im Satz und mit beliebigem Kontext, von denen ein Teil für die kooperativen Übersetzungsexperimente in Paaren konstruiert worden ist. Die letzteren wurden zusätzlich in diesen Lückentext-Übersetzungsexperimenten getestet, um ein repräsentativeres Sample an Daten zu erheben als im kooperativen Übersetzungsexperiment in Paaren.

Die Lückentext-Übersetzungsexperimente mit vorhersehbaren Zielwörtern sind der Teil der Arbeit, der alle Hypothesen vereinigt und zusammen mit den kooperativen Übersetzungsexperimenten in Paaren als wichtigstes Kapitel dieser Arbeit betrachtet werden kann, da darin alle relevanten Faktoren zum

Tragen kommen. Es konnte gezeigt werden, dass Kontext zur Verständlichkeit eines Großteils der Zielwörter, aber nicht aller, beiträgt. Die durchschnittliche Verständlichkeit von Zielwörtern erhöhte sich von 49,71% ohne Kontext auf 67,99% im vorhersehbaren Kontext. Zu den Fällen, in denen Kontext nicht zu einer besseren Verständlichkeit führte, gehörten solche Zielwörter, die in ihren Grundformen in beiden Sprachen identisch sind (z. B. PL/CS *ryba* ‚Fisch‘), sich jedoch in ihren flektierten Formen im Kontext unterscheiden (PL *rybę* vs. CS *rybu* ‚Fisch [Akkusativ]‘). Dies war insbesondere der Fall bei femininen Substantiven im Akkusativ und Instrumental, die auf *-ę* bzw. *-ą* enden. Hier scheinen L1-Interferenzen auf morphologischer Ebene eine Rolle zu spielen. Feminine Akkusativformen, die auf *-ę* enden, wurden häufig mit tschechischen Pluralformen oder anderen auf *-e* or *-ě* endenden Formen übersetzt. Polnische feminine Instrumentalformen mit der Endung *-ą* wurden häufig für feminine Nominativformen gehalten.

Im Vergleich zur Kondition ohne Kontext war der Kontext außerdem bei der korrekten Identifizierung von Wortarten hilfreich, besonders der Verben im Infinitiv. Infinitive wurden in der Kondition ohne Kontext häufiger für Substantive gehalten als in der Kondition mit Kontext. Außerdem stellten sich solche Substantive als problematisch heraus, die im Polnischen ein anderes grammatisches Geschlecht besitzen als ihre tschechischen Übersetzungen. Es konnte nachgewiesen werden, dass bei den Versuchspersonen eine Tendenz vorherrscht, das grammatische Geschlecht des Stimulus in ihrer Übersetzung beizubehalten, was zu falschen Übersetzungen führte. Der Kontext war bei der Disambiguierung solcher Zielwörter hilfreich: Die Verständlichkeit der Zielwörter mit abweichendem grammatischen Geschlecht in beiden Sprachen konnte im Kontext im Schnitt um 28,3% im Vergleich zur Kondition ohne Kontext gesteigert werden. Dies sind etwa 10% mehr als der durchschnittliche Unterschied zwischen allen Zielwörtern in den beiden Konditionen.

Die Vorhersehbarkeit im Kontext hat einen größeren positiven Einfluss auf Nichtkognaten und falsche Freunde als auf Kognaten. Dementsprechend konnte kein signifikanter Zusammenhang zwischen Surprisal und der Verständlichkeit von Kognaten beobachtet werden. Bei der Disambiguierung dieser scheinen sich Versuchspersonen eher auf die Ähnlichkeit der Zielwörter als auf den Kontext zu verlassen. Wenn folglich die Daten der Kognaten aus dem Regressionsmodell ausgeschlossen werden, hat Surprisal eine mäßige negative Korrelation mit der Verständlichkeit der Nicht-Kognaten (inklusive falscher Freunde; $r = -0.411$) und eine noch stärkere Korrelation, wenn nur diejenigen Zielwörter betrachtet werden, die als falsche Freunde identifiziert wurden ($r = -0.443$). In solchen Sätzen, wo Zielwörter oder andere Wörter innerhalb der Stimulissätze Spielraum für semantische Assoziationen mit der korrekten

Übersetzung bieten und die Versuchspersonen zum korrekten Verständnis hinführen, hat der Satzkontext ein großes Potential für eine Erhöhung der Verständlichkeit. Falsche Freunde, die auch Kognaten sind, sowie Nicht-Kognaten, die Assoziationen mit korrekten Übersetzungen der Zielwörter erlauben, waren diejenigen Zielwörter, für die sich die Verständlichkeit im Kontext am stärksten erhöht hat. Die Anzahl (mehr noch als der Anteil) der Nicht-Kognaten und falschen Freunde pro Satz hat eine starke negative Korrelation mit ihrer Verständlichkeit ($r = -0.508$), denn ihr Vorhandensein beeinflusst, wieviel Kontext die Versuchspersonen eigentlich verstehen.

Keine klare Tendenz konnte dahingehend festgestellt werden, ob ein beliebiger Satzkontext die Verständlichkeit von Zielwörtern an beliebiger Stelle im Satz im Vergleich zur Kondition ohne Kontext erhöht. Die fehleranalytischen Ergebnisse zeigen, dass die häufigste Fehlerursache auf die Ähnlichkeit der Zielwörter mit einer anderen als der korrekten Übersetzungsvariante zurückzuführen ist und nahezu bei 30% aller falschen Übersetzungen der Zielwörter in Sätzen mit beliebigem Kontext nachgewiesen werden kann. Im Vergleich dazu ist der Anteil dieser Art von Fehler bei den Zielwörtern in vorhersehbarrem Satzkontext (Abschnitt 15) nur etwa halb so hoch (etwa 15%). Dies deutet darauf hin, dass auch die Position der Zielwörter relevant für die Verständlichkeit sein könnte. Zielwörter am Satzende könnten einfacher zu verstehen sein als an anderer Stelle im Satz. Dies sollte jedoch systematisch in künftigen Studien untersucht werden. Andere falsche Antworten konnten als kontextbedingte Fehler (etwa 8%-9%), Fehler aufgrund von falschen Assoziationen oder Priming (4%-12%) sowie Ln-Interferenzen (unter 1%) identifiziert werden. Fehler aufgrund von falschen Assoziationen und Priming scheinen unter allen Arten von Stimuli bei den Zielwörtern im vorhersehbaren Kontext die geringste Rolle zu spielen (4% aller falschen Antworten).

Eine logische Fortsetzung dieser Arbeit wäre es, tschechische Übersetzungen derselben Stimuli polnischen Versuchspersonen zu präsentieren, um Asymmetrien in der Interkomprehension dieses Sprachenpaars zu untersuchen – dies könnte neue Erkenntnisse in Bezug auf die Prädiktoren bedingte Entropie und Wortadaptationssurprisal (Abschnitt 2.6) liefern. Das Phänomen der inneren Stimme während des Lesens könnte auch aus der Perspektive der polnischen Lesenden untersucht werden. Es wäre hier besonders interessant zu sehen, ob dieselben Strategien und Prozesse (Ignorieren oder Verschieben von Diakritika auf andere Buchstaben) stattfinden, wenn polnische Versuchspersonen Tschechisch übersetzen. Zusätzlich könnte erfasst werden, wie polnische Versuchspersonen tschechische Stimuli allgemein aussprechen, um ein aussprachebasiertes Distanzmaß entsprechend auch für diese Leserichtung aufzustellen. Weitere Untersuchungen bezüglich der angenommenen Aussprache des

Tschechischen (und auch des Polnischen, Kroatischen und Serbischen) wurden bereits in Form von Audioaufnahmen mit russischen und serbischen Studierenden angestellt, die einzelne panslavische Kognaten laut vorgelesen haben (Jágrová & Stenger, 2019).

Anstatt den Versuchspersonen in den kooperativen Übersetzungsexperimenten in Paaren konstruierte Sätze zu präsentieren, wäre es möglicherweise angebrachter gewesen, ihnen die Sätze mit vorhersehbaren Zielwörtern zu präsentieren (aus Abschnitt 15). Dasselbe gilt für die Stimuli in den Experimenten mit Nominalphrasen – diese hätten aus denselben Sätzen extrahiert werden können. Idealerweise hätten Daten über die Verständlichkeit aller in diesen Sätzen vorkommenden Wörter in den freien Übersetzungsexperimenten ohne Kontext erhoben werden können, sodass falsche Freunde innerhalb der Sätze (und nicht nur unter den Zielwörtern) hätten auf zuverlässige Art und Weise experimentell identifiziert werden können.

Einen Grund zur Kritik könnte auch das Design der Stimuli in den Experimenten mit kombinierten Modifikationen von Sätzen (Abschnitt 10) bieten – besonders solcher Stimuli, die schon in ihrer unmodifizierten Variante (Original) den Versuchspersonen keine großen Probleme bereiteten. Bei diesen Stimuli führten die Modifikationen zum sogenannten Deckeneffekt: der Stimulus ist zu einfach zu verstehen und die Antworten der Versuchspersonen waren zu nahezu 100% korrekt. Einige der Versuchspersonen bemerkten, dass sich in den modifizierten Stimuli tschechische Buchstaben befanden. Dies können Argumente gegen das Anwenden solcher modifizierten Sätze in Übersetzungsexperimenten sein.

Freie Übersetzungsexperimente mit den häufigsten Substantiven in den Sprachkombinationen Bulgarisch, Polnisch, Russisch und Tschechisch wurden bereits über die Experiment-Website durchgeführt und die Ergebnisse daraus werden analysiert. Das nächste Ziel ist es, eine interaktive slavische Interkomprehensionsmatrix (Jágrová, Stenger & Avgustinova, 2019) aus den unterschiedlichen Prädiktoren und experimentellen Ergebnissen in so vielen slavischen Sprache-Lesenden-Kombinationen wie möglich aufzustellen. Natürlich kann dieselbe Methodik und die Experiment-Website für jede andere Sprachenkombination auch außerhalb der slavischen Sprachfamilie genutzt werden.

In Bezug auf die vorhersehbaren Zielwörter im Kontext wäre es interessant zu erforschen, wie sich dieselben Zielwörter in variablen, möglicherweise sogar irreführenden Kontexten verhalten. Solche irreführenden Bestandteile der Stimuli könnten wiederum Wörter sein, die im Satz direkt vor den Zielwörtern stehen, oder es könnten semantisch dominante Konzepte an einer anderen

Stelle im Satz sein, die bei den Versuchspersonen zu falschen semantischen Assoziationen führen könnten. Eine weitere Möglichkeit wäre es, die Zielwörter nicht in einem Satzkontext, sondern in einem visuellen Kontext (hilfreich oder irreführend) oder in gesprochener Form (schriftliche und gesprochene Stimuli separat oder beide zusammen) zu präsentieren.

Zur Experiment-Website wurde eine E-Learning-Funktionalität hinzugefügt, die es erlaubt, dass Versuchspersonen ihre Experimente beliebig oft wiederholen. Die Software speichert die Ergebnisse der einzelnen Experimente und zeichnet eine Lernkurve nach jedem wiederholten Experiment. Dies könnte für Sprachlernanfänger oder für Studierende slavischer Mehrsprachenkurse interessant sein. So kann ein solches Experiment z. B. zu Beginn eines Mehrsprachenkurses absolviert und am Ende des Semesters wiederholt werden, um den eigenen Fortschritt in der Interkomprehension zu überprüfen.

Das in dieser Arbeit präsentierte experimentelle Setting und die daraus resultierenden Erkenntnisse können in verschiedenen Bereichen relevant sein. Naheliegend sind alle Situationen, in denen Personen mit Tschechisch als L1 mit geschriebenem Polnisch konfrontiert sind. Andererseits können die Erkenntnisse auch konkret für den Bereich des Fremdsprachenerwerbs relevant sein, wenn Personen mit Tschechisch als Muttersprache Polnisch lernen, sei es in einem formellen oder informellen Rahmen. Neben dem ganzheitlichen Ansatz des Fremdsprachenerwerbs geben aktuelle Entwicklungen der Frage nach dem Erwerb von Teilkompetenzen im Gegensatz zum Anstreben der perfekten Beherrschung einer Fremdsprache Raum. Relativ gute Lernergebnisse in der Teilkompetenz Lesen könnten gerade bei genetisch so nah verwandten Sprachen, aber auch anderen Sprachen innerhalb derselben Sprachfamilie, mit relativ wenig Aufwand zu erreichen sein. Aus den hier dargebotenen Studien könnte man konkret auf zwei Lernstrategien schließen: (i) die Erschließung und Aneignung regelmäßiger zwischensprachlicher Korrespondenzen auf der orthographischen (auch phonetischen und morphologischen) Ebene zur Erkennung von Kognaten oder Wortbausteinen und (ii) das Vermitteln häufig auftretender Nicht-Kognaten und falscher Freunde in entsprechend hilfreichen Satzkontexten. Dies wäre einer der denkbaren nächsten Schritte und Beiträge zur gegenwärtigen Interkomprehensionsforschung und -didaktik, die auf die vorliegende Arbeit aufbauen können.

Trotz sorgfältiger inhaltlicher Kontrolle aller zitierter Links zum Zeitpunkt der Veröffentlichung kann die Autorin keine Garantie für die Fortexistenz von Links übernehmen, die als Quelle genutzt wurden. Um den verlinkten Inhalt zum Zeitpunkt der Verteidigung der Arbeit einsehen zu können, ist die Nutzung der Wayback Machine des Internet Archive nützlich, <https://archive.org/web/>.

Zudem wird keine Haftung für die Inhalte externer Links übernommen. Für den Inhalt und das Hosting der verlinkten Seiten sind ausschließlich deren Betreiber verantwortlich.

CS/PL	a	á	ä	e	ę	i	y	o	ó	u	b	c	ć	d	f	g	h	j	k	l	ł	m	n	ń	p	r	s	ś	t	w	z	ź	ż		
m	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1		
n	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0.5	1	1	1	1	1	1	1	1	1	
ñ	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	0.5	0.5	1	1	1	1	1	1	1	1	1	1	
p	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	
q	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
r	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	
ř	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.5	1	1	1	1	1	1	1	1	1	
s	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0.5	1	1	1	1	1	1	
š	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.5	0.5	1	1	1	1	1	1	
t	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	
ť	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.5	1	1	1	1	
v	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
w	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
x	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
z	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
ż	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.5
ź	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.5
ż	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.5

Table A 1: Alignment matrix used for the calculation of trad LD.

CS/PL	a	e	ę	i	y	o	ó	u	b	c	ć	d	f	g	h	j	k	l	ł	m	n	ń	p	r	s	ś	t	w	z	ż	z
n	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	0	0.5	1	1	1	1	1	1	1	1	
ñ	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	0.5	0.5	1	1	1	1	1	1	1	1	
p	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	
q	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
r	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	
ř	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.5	1	1	1	1	1	1	1	
s	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0.5	1	1	1	1	1	
ś	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.5	0.5	1	1	1	1	1	
t	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	
ť	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.5	1	1	1	
v	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	
w	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	
x	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
z	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0.5	0.5
ż	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.5	0.5	0

Table A 2: Alignment matrix used for the calculation of pron LD.

2. Questionnaire on Sociodemographic Data

This questionnaire had to be filled in by all respondents after they had registered with the website and before they were automatically assigned an experiment.

Basic Information

How old are you?
21

What is your gender?
female

In which country do you live?
Czech Republic

Which language is spoken where you live?
----- Add another
Czech

How long have you lived in this area?
21

Have you lived in an area where another language is spoken?
 No Yes
Which language is spoken there? How long have you lived there (years)? Add another
Arabic 2
Portuguese 2

Where did you go to school?
----- For how long? Add another
Czech Republic 10

What is your highest educational attainment?

Do you hold a university degree in linguistics?
 No Yes

Which language(s) do you speak at home?
----- Add another
Czech

List all the languages that you have ever learned!
----- For how long? Add another
English 10
Portuguese 1

Continue

Figure A 1: Questionnaire on sociodemographic data on the experiment website (EN version).

3. Instruction for the Participants in the Cooperative Translation Experiments

Vítejte u experimentu porozumění polštiny!

Uvidíte několik vět v polštině. Některé věty pro Vás připravil náš polský přítel, který se učí česky. Tyto věty jsou poněkud „počeštěné“. Takže to pro vás vlastně bude hračka.

*Nejdříve si celou větu přečtěte a **nahlas** si řekněte, o co asi v této větě jde nebo co je jejím tématem.*

A pak můžete začít s překladem do češtiny. Podívejte se nejdříve na známá slova, u kterých jste si jisti, co znamenají. Zkuste pak každé slovo ve větě přeložit tak, jak si myslíte, že by bylo v češtině správně. I když některé slovo neznáte, vyvoďte si ho z kontextu nebo hádejte. Přitom se dorozumívejte s vaším partnerem a řeknete si nahlas, co si myslíte, jaké máte myšlenky. (O všem, co si myslíte, mluvte. Zkuste vyslovit každou myšlenku, každý nápad nahlas a konzultujte to s vaším partnerem.)

U „počeštěných“ vět našeho polského přítele opravte věty tak, aby byly správně česky. Cílem je vždy mít dobře znějící českou větu jako překlad. Napište váš překlad vždy do políčka pod větou.

Čas běží! na každou větu máte asi 5 minut, celkem budete mít 12 vět.

Figure A 2: Task as displayed to the respondents in the cooperative translation experiment.

4. Intelligibility of Stimuli in the Different Experiments

4.1. Stimuli with Regular PL-CS Correspondences

Table A3 lists the PL stimuli with applicable regular PL-CS correspondences that were presented to the Czech respondents, their intelligibility scores as well as the ratios of wrong and missing responses. Mean processing time is provided in the column *Total time* (ms).

Stimulus PL	Cognate CS	Intelligibility	Wrong	No answer	Total time (ms)
<i>akademia</i>	<i>akademie</i>	100.00%	0.00%	0.00%	3354.4
<i>aligator</i>	<i>aligátor</i>	100.00%	0.00%	0.00%	4544.4
<i>aparát</i>	<i>aparát</i>	88.57%	11.43%	0.00%	3603.0
<i>apetyt</i>	<i>apetýt</i>	97.14%	0.00%	2.86%	3533.2
<i>autostrada</i>	<i>autostráda</i>	40.00%	40.00%	20.00%	6533.4
<i>bal</i>	<i>bál</i>	80.00%	14.29%	5.71%	3641.5
<i>bardo</i>	<i>brdo</i>	5.71%	57.14%	37.14%	5500.8
<i>biada</i>	<i>běda</i>	2.86%	88.57%	8.57%	3855.9
<i>biały</i>	<i>bílý</i>	100.00%	0.00%	0.00%	4955.2
<i>bić</i>	<i>bít</i>	11.43%	88.57%	0.00%	3726.7
<i>bieda</i>	<i>bída</i>	88.57%	11.43%	0.00%	3121.8
<i>biegać</i>	<i>běhat</i>	42.86%	31.43%	25.71%	5956.4
<i>boleć</i>	<i>bolet</i>	37.14%	60.00%	2.86%	3729.1
<i>broníć</i>	<i>bránit</i>	40.00%	20.00%	40.00%	6430.4
<i>brzoza</i>	<i>bříza</i>	0.00%	80.00%	20.00%	5145.0
<i>bujny</i>	<i>bujný</i>	74.29%	17.14%	8.57%	4485.7

Stimulus PL	Cognate CS	Intelligibility	Wrong	No answer	Total time (ms)
<i>bukiet</i>	<i>buket</i>	68.57%	22.86%	8.57%	4501.7
<i>burza</i>	<i>bouře</i>	2.86%	97.14%	0.00%	3251.8
<i>byk</i>	<i>býk</i>	100.00%	0.00%	0.00%	2971.2
<i>cel</i>	<i>cíl</i>	33.33%	52.38%	14.29%	4940.5
<i>cenny</i>	<i>cenný</i>	100.00%	0.00%	0.00%	4199.8
<i>chart</i>	<i>chrt</i>	11.43%	74.29%	14.29%	6014.1
<i>chłodny</i>	<i>chladný</i>	88.57%	11.43%	0.00%	4575.5
<i>chmiel</i>	<i>chmel</i>	100.00%	0.00%	0.00%	3088.4
<i>chodzić</i>	<i>chodit</i>	65.71%	34.29%	0.00%	4438.7
<i>ciało</i>	<i>tělo</i>	9.52%	85.71%	4.76%	4940.9
<i>cichy</i>	<i>tichý</i>	20.00%	65.71%	14.29%	5461.1
<i>cielątko</i>	<i>telátko</i>	40.00%	40.00%	20.00%	5101.8
<i>ciemny</i>	<i>temný</i>	40.00%	60.00%	0.00%	3977.6
<i>ciepły</i>	<i>teplý</i>	80.00%	0.00%	20.00%	5922.4
<i>cierní</i>	<i>trn</i>	0.00%	88.57%	11.43%	5560.9
<i>cis</i>	<i>tis</i>	2.86%	60.00%	37.14%	6481.2
<i>cukier</i>	<i>cukr</i>	94.29%	2.86%	2.86%	2554.5
<i>czajka</i>	<i>čejka</i>	11.43%	77.14%	11.43%	5459.3
<i>czajnik</i>	<i>čajník</i>	54.29%	37.14%	8.57%	5338.5
<i>czarny</i>	<i>černý</i>	80.00%	20.00%	0.00%	4530.0
<i>czerw</i>	<i>červ</i>	80.00%	20.00%	0.00%	4425.8
<i>czerwony</i>	<i>červený</i>	91.43%	5.71%	2.86%	4864.7

Stimulus PL	Cognate CS	Intelligibility	Wrong	No answer	Total time (ms)
<i>człowiek</i>	<i>člověk</i>	100.00%	0.00%	0.00%	4914.0
<i>czoło</i>	<i>čelo</i>	45.71%	48.57%	5.71%	4650.1
<i>czasnek</i>	<i>česnek</i>	97.14%	2.86%	0.00%	4692.9
<i>dąb</i>	<i>dub</i>	38.10%	42.86%	19.05%	5428.4
<i>dąć</i>	<i>dout</i>	0.00%	94.29%	5.71%	4116.3
<i>dama</i>	<i>dáma</i>	87.50%	12.50%	0.00%	2838.1
<i>dawać</i>	<i>dávat</i>	71.05%	21.05%	7.89%	4324.0
<i>dłoń</i>	<i>dlaň</i>	75.00%	25.00%	0.00%	4247.4
<i>dobry</i>	<i>dobrý</i>	100.00%	0.00%	0.00%	2840.0
<i>droga</i>	<i>dráha</i>	5.71%	91.43%	2.86%	3016.8
<i>drzewo</i>	<i>dřevo</i>	95.95%	4.05%	0.00%	4768.6
<i>dumać</i>	<i>dumat</i>	60.00%	31.43%	8.57%	5141.2
<i>dusza</i>	<i>duše</i>	100.00%	0.00%	0.00%	3799.9
<i>dwa</i>	<i>dva</i>	94.29%	0.00%	5.71%	2580.3
<i>dychać</i>	<i>dýchat</i>	75.00%	25.00%	0.00%	5562.8
<i>dym</i>	<i>dým</i>	100.00%	0.00%	0.00%	2656.0
<i>dynia</i>	<i>dýně</i>	90.48%	4.76%	4.76%	4261.1
<i>dżem</i>	<i>džem</i>	100.00%	0.00%	0.00%	3054.5
<i>dzień</i>	<i>den</i>	52.38%	47.62%	0.00%	5109.3
<i>dżungla</i>	<i>džungle</i>	87.50%	0.00%	12.50%	3676.3
<i>ekonomia</i>	<i>ekonomie</i>	97.14%	0.00%	2.86%	3876.1
<i>energia</i>	<i>energie</i>	100.00%	0.00%	0.00%	3232.8

Stimulus PL	Cognate CS	Intelligibility	Wrong	No answer	Total time (ms)
<i>galeria</i>	<i>galerie</i>	100.00%	0.00%	0.00%	3300.4
<i>garb</i>	<i>hrb</i>	50.00%	50.00%	0.00%	4990.0
<i>gardło</i>	<i>hrdlo</i>	38.10%	33.33%	28.57%	5246.5
<i>gęba</i>	<i>huba</i>	0.00%	40.00%	60.00%	6366.0
<i>gęsty</i>	<i>hustý</i>	0.00%	87.50%	12.50%	4027.0
<i>glina</i>	<i>hlína</i>	62.50%	37.50%	0.00%	3781.3
<i>głowa</i>	<i>hlava</i>	57.14%	38.10%	4.76%	4807.3
<i>głuchy</i>	<i>hluchý</i>	86.76%	10.29%	2.94%	3908.8
<i>gnida</i>	<i>hnida</i>	62.50%	37.50%	0.00%	4191.3
<i>gniew</i>	<i>hněv</i>	66.67%	19.05%	14.29%	4924.7
<i>gołąb</i>	<i>holub</i>	25.00%	62.50%	12.50%	6352.5
<i>goły</i>	<i>holý</i>	12.50%	87.50%	0.00%	4963.5
<i>gonić</i>	<i>honit</i>	25.00%	75.00%	0.00%	6450.1
<i>góra</i>	<i>hora</i>	80.00%	20.00%	0.00%	3915.4
<i>grać</i>	<i>hrát</i>	34.29%	65.71%	0.00%	3732.8
<i>groch</i>	<i>hrách</i>	0.00%	95.24%	4.76%	3593.9
<i>grzbiet</i>	<i>hřbet</i>	25.71%	51.43%	22.86%	5944.1
<i>grzech</i>	<i>hřích</i>	19.05%	57.14%	23.81%	6598.4
<i>historia</i>	<i>historie</i>	100.00%	0.00%	0.00%	4645.3
<i>ja</i>	<i>já</i>	80.00%	20.00%	0.00%	2770.8
<i>jabłko</i>	<i>jablko</i>	100.00%	0.00%	0.00%	3212.7
<i>jagnię</i>	<i>jehně</i>	23.81%	52.38%	23.81%	4950.5

Stimulus PL	Cognate CS	Intelligibility	Wrong	No answer	Total time (ms)
<i>jaguar</i>	<i>jaguár</i>	100.00%	0.00%	0.00%	3486.1
<i>jama</i>	<i>jáma</i>	97.14%	2.86%	0.00%	3202.8
<i>jarzq̄b</i>	<i>jeřáb</i>	100.00%	0.00%	0.00%	3840.2
<i>jasny</i>	<i>jasný</i>	94.29%	5.71%	0.00%	3653.5
<i>jastrzq̄b</i>	<i>jestřáb</i>	88.57%	2.86%	8.57%	4503.1
<i>jawor</i>	<i>javor</i>	97.14%	2.86%	0.00%	3619.1
<i>jeleń</i>	<i>jelen</i>	95.24%	0.00%	4.76%	3037.7
<i>jesień</i>	<i>jeseň</i>	0.00%	80.00%	20.00%	5603.0
<i>jezioro</i>	<i>jezero</i>	95.24%	4.76%	0.00%	3485.8
<i>język</i>	<i>jazyk</i>	100.00%	0.00%	0.00%	2982.0
<i>kalendarz</i>	<i>kalendář</i>	100.00%	0.00%	0.00%	3899.3
<i>kamień</i>	<i>kámen</i>	100.00%	0.00%	0.00%	3634.0
<i>kanał</i>	<i>kanál</i>	100.00%	0.00%	0.00%	3602.8
<i>karawana</i>	<i>karavana</i>	62.50%	37.50%	0.00%	4798.1
<i>kark</i>	<i>krk</i>	42.86%	34.29%	22.86%	5309.5
<i>karnawał</i>	<i>karneval</i>	100.00%	0.00%	0.00%	4151.2
<i>kq̄sac̄</i>	<i>kousat</i>	8.57%	74.29%	17.14%	6049.7
<i>kasztan</i>	<i>kaštan</i>	100.00%	0.00%	0.00%	3803.4
<i>kawa</i>	<i>káva</i>	100.00%	0.00%	0.00%	2645.6
<i>kawaler</i>	<i>kavalír</i>	75.00%	25.00%	0.00%	4932.9
<i>kiwi</i>	<i>kivi</i>	100.00%	0.00%	0.00%	2036.8
<i>kłos</i>	<i>klas</i>	28.57%	57.14%	14.29%	5119.8

Stimulus PL	Cognate CS	Intelligibility	Wrong	No answer	Total time (ms)
<i>kobyła</i>	<i>kobyła</i>	100.00%	0.00%	0.00%	3491.4
<i>koktajl</i>	<i>koktejl</i>	95.24%	0.00%	4.76%	4595.3
<i>kolano</i>	<i>koleno</i>	95.24%	0.00%	4.76%	3529.6
<i>komar</i>	<i>komár</i>	100.00%	0.00%	0.00%	3285.4
<i>konopie</i>	<i>konopí</i>	100.00%	0.00%	0.00%	3492.9
<i>kopać</i>	<i>kopat</i>	25.00%	75.00%	0.00%	4278.1
<i>kora</i>	<i>kůra</i>	50.00%	25.00%	25.00%	3015.4
<i>korzeń</i>	<i>kořen</i>	47.62%	38.10%	14.29%	5270.4
<i>kosić</i>	<i>kosit</i>	12.50%	87.50%	0.00%	4681.1
<i>koziół</i>	<i>kozel</i>	100.00%	0.00%	0.00%	3233.8
<i>krew</i>	<i>krev</i>	85.71%	11.43%	2.86%	3052.5
<i>krokodyl</i>	<i>krokodýl</i>	100.00%	0.00%	0.00%	4397.1
<i>krowa</i>	<i>kráva</i>	38.10%	61.90%	0.00%	4237.5
<i>krzemień</i>	<i>křemen</i>	12.50%	62.50%	25.00%	6468.0
<i>kupić</i>	<i>koupit</i>	57.35%	38.24%	4.41%	3719.2
<i>kwiat</i>	<i>květ</i>	80.00%	20.00%	0.00%	5817.2
<i>łąka</i>	<i>louka</i>	14.29%	65.71%	20.00%	5815.3
<i>łapać</i>	<i>lapat</i>	12.50%	87.50%	0.00%	4775.3
<i>las</i>	<i>les</i>	51.43%	45.71%	2.86%	4488.5
<i>laska</i>	<i>líška</i>	0.00%	100.00%	0.00%	2740.2
<i>lato</i>	<i>léto</i>	20.00%	80.00%	0.00%	5320.6
<i>lecieć</i>	<i>letět</i>	0.00%	100.00%	0.00%	6080.4

Stimulus PL	Cognate CS	Intelligibility	Wrong	No answer	Total time (ms)
<i>lepić</i>	<i>lepit</i>	50.00%	50.00%	0.00%	5501.1
<i>lew</i>	<i>lev</i>	100.00%	0.00%	0.00%	2619.4
<i>lewy</i>	<i>levý</i>	100.00%	0.00%	0.00%	4356.1
<i>leżeć</i>	<i>ležet</i>	50.00%	50.00%	0.00%	4828.0
<i>linia</i>	<i>linie</i>	80.00%	20.00%	0.00%	3134.6
<i>lipa</i>	<i>lípa</i>	100.00%	0.00%	0.00%	2763.5
<i>łokieć</i>	<i>loket</i>	80.00%	0.00%	20.00%	5560.6
<i>macocha</i>	<i>macecha</i>	57.14%	42.86%	0.00%	5430.0
<i>mądry</i>	<i>moudrý</i>	28.57%	57.14%	14.29%	3875.9
<i>maj</i>	<i>máj</i>	100.00%	0.00%	0.00%	2798.6
<i>mak</i>	<i>mák</i>	100.00%	0.00%	0.00%	3154.5
<i>mało</i>	<i>málo</i>	97.14%	2.86%	0.00%	3030.5
<i>mały</i>	<i>malý</i>	100.00%	0.00%	0.00%	2918.0
<i>martwy</i>	<i>mrtvý</i>	87.50%	0.00%	12.50%	3464.3
<i>materiał</i>	<i>materiál</i>	100.00%	0.00%	0.00%	3199.0
<i>mąż</i>	<i>muž</i>	23.81%	76.19%	0.00%	5565.0
<i>mazać</i>	<i>mazat</i>	37.50%	37.50%	25.00%	5471.5
<i>męka</i>	<i>muka</i>	9.52%	80.95%	9.52%	6262.9
<i>miesiąc</i>	<i>měsíc</i>	100.00%	0.00%	0.00%	4866.4
<i>mięso</i>	<i>maso</i>	80.00%	20.00%	0.00%	5730.4
<i>milczeń</i>	<i>mlčet</i>	12.50%	87.50%	0.00%	6621.3
<i>miły</i>	<i>milý</i>	100.00%	0.00%	0.00%	3077.5

Stimulus PL	Cognate CS	Intelligibility	Wrong	No answer	Total time (ms)
<i>ministerstwo</i>	<i>ministerstvo</i>	100.00%	0.00%	0.00%	4418.0
<i>młody</i>	<i>mladý</i>	100.00%	0.00%	0.00%	4107.8
<i>mokry</i>	<i>mokrý</i>	95.12%	4.88%	0.00%	3821.5
<i>morze</i>	<i>moře</i>	75.00%	25.00%	0.00%	3611.6
<i>mucha</i>	<i>moucha</i>	97.14%	2.86%	0.00%	3206.5
<i>myć</i>	<i>mýt</i>	25.00%	75.00%	0.00%	4611.5
<i>mysz</i>	<i>myš</i>	80.00%	20.00%	0.00%	2517.2
<i>nadzieja</i>	<i>naděje</i>	60.00%	40.00%	0.00%	5200.6
<i>niebo</i>	<i>nebe</i>	25.49%	72.55%	1.96%	2703.8
<i>niemy</i>	<i>němý</i>	100.00%	0.00%	0.00%	2599.8
<i>niewiasta</i>	<i>nevěsta</i>	100.00%	0.00%	0.00%	3928.6
<i>noga</i>	<i>noha</i>	80.00%	20.00%	0.00%	3350.6
<i>nowy</i>	<i>nový</i>	100.00%	0.00%	0.00%	2064.4
<i>obawa</i>	<i>obava</i>	80.00%	20.00%	0.00%	2671.6
<i>obłok</i>	<i>oblak</i>	11.43%	88.57%	0.00%	4100.8
<i>obuć</i>	<i>about</i>	20.00%	80.00%	0.00%	3486.2
<i>oficer</i>	<i>oficír</i>	80.00%	5.71%	14.29%	7139.5
<i>ogień</i>	<i>ohěň</i>	71.43%	23.81%	4.76%	4733.8
<i>ogon</i>	<i>ohon</i>	42.86%	42.86%	14.29%	5821.5
<i>okrągły</i>	<i>okrouhlý</i>	0.00%	87.50%	12.50%	6654.5
<i>okulary</i>	<i>okuláry</i>	97.14%	2.86%	0.00%	4241.7
<i>olsza</i>	<i>olše</i>	57.14%	38.10%	4.76%	4682.5

Stimulus PL	Cognate CS	Intelligibility	Wrong	No answer	Total time (ms)
<i>orzech</i>	<i>ořech</i>	100.00%	0.00%	0.00%	2855.8
<i>osioł</i>	<i>osel</i>	95.24%	4.76%	0.00%	3286.7
<i>ostrzy</i>	<i>ostrý</i>	100.00%	0.00%	0.00%	2757.8
<i>otawa</i>	<i>otava</i>	40.00%	60.00%	0.00%	6318.8
<i>owies</i>	<i>oves</i>	100.00%	0.00%	0.00%	3008.2
<i>owoce</i>	<i>ovoce</i>	100.00%	0.00%	0.00%	2898.6
<i>padać</i>	<i>padat</i>	62.86%	34.29%	2.86%	3137.9
<i>palić</i>	<i>pálit</i>	40.00%	60.00%	0.00%	4436.0
<i>pałka</i>	<i>pálka</i>	97.14%	2.86%	0.00%	4390.9
<i>papier</i>	<i>papír</i>	97.14%	2.86%	0.00%	3161.7
<i>para</i>	<i>pára</i>	100.00%	0.00%	0.00%	2016.8
<i>partia</i>	<i>partie</i>	28.57%	71.43%	0.00%	4052.4
<i>paw</i>	<i>páv</i>	80.00%	0.00%	20.00%	2613.2
<i>pełny</i>	<i>plný</i>	20.00%	80.00%	0.00%	6044.4
<i>pepek</i>	<i>pupek</i>	28.57%	52.38%	19.05%	6579.3
<i>piana</i>	<i>pěna</i>	9.52%	90.48%	0.00%	4394.9
<i>piasek</i>	<i>písek</i>	52.38%	47.62%	0.00%	3384.8
<i>pić</i>	<i>pít</i>	20.00%	80.00%	0.00%	4474.8
<i>pięć</i>	<i>pět</i>	0.00%	100.00%	0.00%	6092.6
<i>piekło</i>	<i>peklo</i>	100.00%	0.00%	0.00%	2371.6
<i>pień</i>	<i>peň</i>	20.00%	60.00%	20.00%	4936.2
<i>pies</i>	<i>pes</i>	80.95%	4.76%	14.29%	3986.1

Stimulus PL	Cognate CS	Intelligibility	Wrong	No answer	Total time (ms)
<i>pięta</i>	<i>pata</i>	11.43%	80.00%	8.57%	4926.3
<i>pióro</i>	<i>pero</i>	51.43%	34.29%	14.29%	5742.4
<i>plakać</i>	<i>plakat</i>	71.43%	28.57%	0.00%	3505.9
<i>plan</i>	<i>plán</i>	100.00%	0.00%	0.00%	3097.3
<i>podły</i>	<i>podlý</i>	100.00%	0.00%	0.00%	2849.8
<i>popiół</i>	<i>popel</i>	37.14%	48.57%	14.29%	5788.7
<i>prac</i>	<i>prát</i>	45.71%	54.29%	0.00%	4510.0
<i>prawda</i>	<i>pravda</i>	100.00%	0.00%	0.00%	2551.8
<i>prawy</i>	<i>pravý</i>	88.57%	11.43%	0.00%	3469.3
<i>pręt</i>	<i>prut</i>	4.76%	76.19%	19.05%	5012.9
<i>przedni</i>	<i>přední</i>	100.00%	0.00%	0.00%	3965.2
<i>pszenica</i>	<i>pšenice</i>	80.00%	20.00%	0.00%	4980.8
<i>ptak</i>	<i>pták</i>	100.00%	0.00%	0.00%	2274.6
<i>pusty</i>	<i>pustý</i>	100.00%	0.00%	0.00%	2853.2
<i>radio</i>	<i>rádio</i>	100.00%	0.00%	0.00%	3227.2
<i>raj</i>	<i>ráj</i>	94.29%	5.71%	0.00%	2523.3
<i>rakieta</i>	<i>raketa</i>	97.14%	0.00%	2.86%	3523.5
<i>referat</i>	<i>referát</i>	95.38%	4.62%	0.00%	4023.8
<i>ręka</i>	<i>ruka</i>	5.56%	94.44%	0.00%	3229.1
<i>religia</i>	<i>religie</i>	82.86%	11.43%	5.71%	6804.4
<i>rezultat</i>	<i>rezultát</i>	52.38%	33.33%	14.29%	6272.4
<i>róg</i>	<i>roh</i>	48.57%	45.71%	5.71%	3613.4

Stimulus PL	Cognate CS	Intelligibility	Wrong	No answer	Total time (ms)
<i>równy</i>	<i>rovný</i>	97.14%	0.00%	2.86%	3296.8
<i>rozdzielić</i>	<i>rozdělit</i>	71.43%	25.71%	2.86%	5791.7
<i>rozumny</i>	<i>rozumný</i>	100.00%	0.00%	0.00%	4559.4
<i>rzqd</i>	<i>řád</i>	52.38%	33.33%	14.29%	5165.0
<i>rzeka</i>	<i>řeka</i>	100.00%	0.00%	0.00%	3471.8
<i>rzepa</i>	<i>řepa</i>	100.00%	0.00%	0.00%	2984.9
<i>rzezać</i>	<i>řezat</i>	40.00%	60.00%	0.00%	4446.2
<i>sadło</i>	<i>sádlo</i>	82.86%	17.14%	0.00%	3792.3
<i>siać</i>	<i>sít</i>	0.00%	60.00%	40.00%	5671.4
<i>siekać</i>	<i>sekat</i>	60.00%	20.00%	20.00%	5132.2
<i>siemię</i>	<i>sémě</i>	61.90%	23.81%	14.29%	6199.5
<i>sita</i>	<i>síla</i>	100.00%	0.00%	0.00%	2089.4
<i>siostra</i>	<i>sestra</i>	80.00%	20.00%	0.00%	3933.6
<i>skakać</i>	<i>skákat</i>	100.00%	0.00%	0.00%	3984.2
<i>sława</i>	<i>sláva</i>	100.00%	0.00%	0.00%	2219.6
<i>słoń</i>	<i>slon</i>	100.00%	0.00%	0.00%	2275.6
<i>słowik</i>	<i>slavík</i>	9.52%	80.95%	9.52%	4968.5
<i>słyszeć</i>	<i>slyšet</i>	80.00%	0.00%	20.00%	4512.0
<i>smutny</i>	<i>smutný</i>	100.00%	0.00%	0.00%	3940.4
<i>sójka</i>	<i>sojka</i>	100.00%	0.00%	0.00%	2263.0
<i>sokół</i>	<i>sokol</i>	100.00%	0.00%	0.00%	2751.5
<i>solić</i>	<i>solit</i>	40.00%	60.00%	0.00%	4554.0

Stimulus PL	Cognate CS	Intelligibility	Wrong	No answer	Total time (ms)
<i>sowa</i>	<i>sova</i>	100.00%	0.00%	0.00%	3490.0
<i>spać</i>	<i>spát</i>	60.00%	40.00%	0.00%	3038.8
<i>spacer</i>	<i>špacír</i>	12.31%	70.77%	16.92%	6046.9
<i>stać</i>	<i>stát</i>	40.00%	40.00%	20.00%	4179.8
<i>stary</i>	<i>starý</i>	94.74%	2.63%	2.63%	3483.4
<i>struga</i>	<i>strouha</i>	48.57%	37.14%	14.29%	4962.5
<i>suchy</i>	<i>suchý</i>	100.00%	0.00%	0.00%	2297.0
<i>surowy</i>	<i>surový</i>	80.00%	20.00%	0.00%	3526.2
<i>sypać</i>	<i>sypat</i>	60.00%	40.00%	0.00%	2961.0
<i>szkoła</i>	<i>škola</i>	91.43%	5.71%	2.86%	2835.8
<i>szlachta</i>	<i>šlechta</i>	45.71%	45.71%	8.57%	5320.7
<i>szpital</i>	<i>špitál</i>	50.00%	50.00%	0.00%	4402.8
<i>talerz</i>	<i>talíř</i>	60.00%	40.00%	0.00%	3931.2
<i>teczka</i>	<i>tečka</i>	97.14%	2.86%	0.00%	3484.3
<i>telewizor</i>	<i>televizor</i>	97.14%	0.00%	2.86%	3610.3
<i>teoria</i>	<i>teorie</i>	88.57%	2.86%	8.57%	2869.2
<i>tępy</i>	<i>tupý</i>	2.86%	80.00%	17.14%	5006.4
<i>tlusty</i>	<i>tlustý</i>	97.14%	2.86%	0.00%	3459.2
<i>tramwaj</i>	<i>tramvaj</i>	100.00%	0.00%	0.00%	3315.8
<i>trawa</i>	<i>tráva</i>	97.14%	2.86%	0.00%	3494.9
<i>trzeć</i>	<i>třít</i>	5.71%	80.00%	14.29%	6028.5
<i>twardy</i>	<i>tvrdý</i>	100.00%	0.00%	0.00%	3094.8

Stimulus PL	Cognate CS	Intelligibility	Wrong	No answer	Total time (ms)
<i>tył</i>	<i>týl</i>	66.67%	23.81%	9.52%	3858.4
<i>tytuł</i>	<i>tytul</i>	100.00%	0.00%	0.00%	3381.7
<i>umierać</i>	<i>umírat</i>	74.29%	25.71%	0.00%	5247.3
<i>umrzeć</i>	<i>umřít</i>	51.43%	48.57%	0.00%	5388.3
<i>usta</i>	<i>ústa</i>	80.00%	20.00%	0.00%	3721.8
<i>ważny</i>	<i>važný</i>	97.06%	0.00%	2.94%	3849.7
<i>wdowa</i>	<i>vdova</i>	91.43%	2.86%	5.71%	2650.9
<i>we</i>	<i>ve</i>	57.14%	37.14%	5.71%	4558.1
<i>wesoły</i>	<i>veselý</i>	82.86%	14.29%	2.86%	4306.4
<i>wesz</i>	<i>veš</i>	37.50%	62.50%	0.00%	4182.3
<i>wiara</i>	<i>víra</i>	66.67%	23.81%	9.52%	4262.0
<i>wiatr</i>	<i>vítr</i>	100.00%	0.00%	0.00%	3157.2
<i>wieczór</i>	<i>večer</i>	22.86%	57.14%	20.00%	6002.8
<i>wiedzieć</i>	<i>vědět</i>	48.57%	48.57%	2.86%	5402.3
<i>wierny</i>	<i>věrný</i>	100.00%	0.00%	0.00%	3687.7
<i>wierzba</i>	<i>vrba</i>	38.10%	47.62%	14.29%	5969.1
<i>wierzch</i>	<i>vrch</i>	20.00%	45.71%	34.29%	5733.3
<i>wilk</i>	<i>vlk</i>	40.00%	40.00%	20.00%	4584.4
<i>włosy</i>	<i>vlasy</i>	94.29%	5.71%	0.00%	3534.2
<i>wnuk</i>	<i>vnuk</i>	100.00%	0.00%	0.00%	3114.4
<i>woda</i>	<i>voda</i>	100.00%	0.00%	0.00%	2743.4
<i>wola</i>	<i>vůle</i>	31.43%	62.86%	5.71%	3912.8

Stimulus PL	Cognate CS	Intelligibility	Wrong	No answer	Total time (ms)
wrona	vrána	100.00%	0.00%	0.00%	5527.6
wstać	vstát	50.00%	37.50%	12.50%	5434.1
wy	vy	94.29%	2.86%	2.86%	2624.1
wydra	vydra	100.00%	0.00%	0.00%	3519.3
wziąć	vzít	37.50%	62.50%	0.00%	5555.1
zqb	zub	28.57%	52.38%	19.05%	5141.9
żaba	žába	100.00%	0.00%	0.00%	3174.8
zabić	zabít	77.14%	17.14%	5.71%	4402.2
żał	žal	61.90%	33.33%	4.76%	4180.9
żebro	žebro	100.00%	0.00%	0.00%	2593.8
żelazo	železo	80.00%	20.00%	0.00%	4358.8
zielony	zelený	80.00%	20.00%	0.00%	6657.6
ziemia	země	60.00%	40.00%	0.00%	4853.2
złoto	zlato	100.00%	0.00%	0.00%	3834.8
zły	zlý	100.00%	0.00%	0.00%	2873.4
żółty	žlutý	60.00%	40.00%	0.00%	6565.8
Mean		66.73%	28.12%	5.15%	4211.5
SD		33.92%	30.33%	8.84%	1164.2

Table A 3: Stimuli with regular PL-CS correspondences, their intelligibility and processing times.

4.2. Most Frequent PL Nouns

Table A 4 lists the stimuli among the 100 most frequent PL nouns that were presented to Czech readers in a free translation experiment, their intelligibility to Czech readers and the different predictor variables for each stimulus. The column labelled CS contains the orthographically closest CS cognates of the stimuli that can be mutual translations in a particular context and can serve as a transfer base. They are not meant to be the optimal translations, but they have at least one meaning in common. If there was more than one possible translation, the orthographically closest was chosen. All cognate translations are marked green. If a field is not marked green, there is no cognate translation available. In the column labelled Gender, a value of 1 is indicated if the grammatical gender of the CS cognate is different from the PL stimulus.

PL	CS	Intelligibility	Pron LD	FF	NC	WAS	WAS norm	Gender
<i>akcja</i>	<i>akce</i>	100.00%	40.00%	0	0	6.544	1.309	0
<i>bank</i>	<i>banka</i>	86.67%	20.00%	0	0	6.675	1.335	1
<i>bóg</i>	<i>bůh</i>	26.67%	66.67%	0	0	2.737	0.912	0
<i>cel</i>	<i>cíl</i>	96.67%	33.33%	0	0	6.248	2.083	0
<i>chwila</i>	<i>chvilé</i>	96.67%	16.67%	0	0	7.272	1.212	0
<i>ciało</i>	<i>tělo</i>	9.52%	60.00%	1	0	7.432	1.486	0
<i>czas</i>	<i>čas</i>	96.67%	37.50%	0	0	4.533	1.133	0
<i>część</i>	<i>část</i>	0.00%	80.00%	0	0	5.067	1.013	0
<i>człowiek</i>	<i>člověk</i>	100.00%	31.25%	0	0	8.522	1.065	0
<i>decyzja</i>	<i>rozhodnutí</i>	14.70%	100.00%	0	1	40.614	4.061	1
<i>dom</i>	<i>dům</i>	91.18%	33.33%	0	0	4.354	1.451	0
<i>droga</i>	<i>dráha</i>	5.71%	40.00%	1	0	6.086	1.217	0
<i>drzwi</i>	<i>dveře</i>	6.67%	83.33%	0	0	19.319	3.220	0
<i>działanie</i>	<i>děj</i>	16.67%	88.89%	0	0	27.041	3.005	1
<i>dziecko</i>	<i>děcko</i>	100.00%	28.57%	0	0	6.594	0.942	0
<i>dzień</i>	<i>den</i>	52.38%	50.00%	0	0	3.408	0.682	0
<i>głos</i>	<i>hlas</i>	26.67%	50.00%	0	0	4.208	1.052	0
<i>głowa</i>	<i>hlava</i>	57.14%	40.00%	0	0	4.955	0.991	0
<i>gmina</i>	<i>komuna</i>	0.00%	50.00%	0	0	13.292	2.215	0
<i>godzina</i>	<i>hodina</i>	46.67%	28.57%	0	0	5.577	0.797	0
<i>góra</i>	<i>hora</i>	80.00%	25.00%	0	0	2.235	0.559	0
<i>grupa</i>	<i>grupa</i>	96.67%	0.00%	0	0	7.161	1.432	0
<i>informacja</i>	<i>informace</i>	94.12%	20.00%	0	0	10.157	1.016	0
<i>kobieta</i>	<i>žena</i>	6.67%	71.43%	0	1	15.327	2.190	0
<i>komisja</i>	<i>komise</i>	93.33%	28.57%	0	0	8.171	1.167	0
<i>koniec</i>	<i>konec</i>	96.67%	16.67%	0	0	4.930	0.822	0
<i>mężczyzna</i>	<i>muž</i>	3.33%	77.78%	0	0	15.642	1.738	1
<i>miasto</i>	<i>město</i>	20.00%	33.33%	1	0	5.330	0.888	0
<i>miejsce</i>	<i>místo</i>	10.00%	57.14%	0	0	11.325	1.618	0

PL	CS	Intelligibility	Pron LD	FF	NC	WAS	WAS norm	Gender
<i>miesiąc</i>	<i>měsíc</i>	100.00%	28.57%	0	0	8.829	1.261	0
<i>minister</i>	<i>ministr</i>	97.67%	12.50%	0	0	8.907	1.113	0
<i>możliwość</i>	<i>možnost</i>	21.88%	50.00%	0	0	8.678	0.964	0
<i>myśl</i>	<i>mysl</i>	96.67%	12.50%	0	0	3.409	0.852	0
<i>numer</i>	<i>číslo</i>	90.00%	100.00%	0	1	21.375	3.563	1
<i>ojciec</i>	<i>otec</i>	58.82%	50.00%	0	0	7.000	1.167	0
<i>okres</i>	<i>období</i>	3.33%	83.33%	1	1	21.133	3.522	1
<i>państwo</i>	<i>panstvo</i>	70.00%	7.14%	0	0	2.448	0.350	0
<i>pieniądze</i>	<i>peníze</i>	96.67%	33.33%	0	0	15.084	1.676	0
<i>pokój</i>	<i>pokoj</i>	100.00%	0.00%	0	0	4.048	0.810	0
<i>poprawka</i>	<i>oprava</i>	3.33%	25.00%	0	0	9.831	1.229	0
<i>poseł</i>	<i>posel</i>	90.00%	0.00%	0	0	2.768	0.554	0
<i>powód</i>	<i>důvod</i>	6.67%	40.00%	1	0	9.205	1.841	0
<i>praca</i>	<i>práce</i>	93.33%	20.00%	0	0	8.736	1.747	0
<i>prawda</i>	<i>pravda</i>	100.00%	0.00%	0	0	3.098	0.516	0
<i>prawo</i>	<i>právo</i>	90.00%	0.00%	0	0	4.936	0.987	0
<i>problem</i>	<i>problém</i>	97.06%	0.00%	0	0	6.362	0.909	0
<i>przepis</i>	<i>předpis</i>	26.67%	31.25%	0	0	11.775	1.472	0
<i>przypadek</i>	<i>případ</i>	40.00%	38.89%	0	0	12.376	1.375	0
<i>punkt</i>	<i>bod</i>	63.33%	100.00%	0	1	16.140	3.228	0
<i>pytanie</i>	<i>ptaní</i>	80.00%	28.57%	0	0	8.639	1.234	0
<i>raz</i>	<i>ráz</i>	73.33%	0.00%	0	0	5.767	1.922	0
<i>ręka</i>	<i>ruka</i>	5.56%	25.00%	1	0	2.746	0.686	0
<i>rodzina</i>	<i>rodina</i>	93.33%	14.29%	0	0	5.992	0.856	0
<i>rząd</i>	<i>řád</i>	52.38%	50.00%	0	0	6.008	1.502	0
<i>rzecz</i>	<i>věc</i>	2.94%	60.00%	1	1	11.974	2.395	0
<i>sila</i>	<i>síla</i>	100.00%	0.00%	0	0	3.888	0.972	0
<i>słowo</i>	<i>slovo</i>	100.00%	0.00%	0	0	1.622	0.324	0
<i>śmierć</i>	<i>smrt</i>	20.00%	58.33%	0	0	3.650	0.608	0
<i>sprawa</i>	<i>záležitost</i>	0.00%	100.00%	1	1	43.439	4.344	0
<i>środek</i>	<i>prostředek</i>	3.33%	60.00%	0	0	18.869	3.145	0
<i>stan</i>	<i>stav</i>	0.00%	25.00%	1	0	5.694	1.424	0
<i>strona</i>	<i>strana</i>	16.67%	16.67%	1	0	6.083	1.014	0
<i>świat</i>	<i>svět</i>	56.67%	50.00%	0	0	4.769	0.954	0
<i>sytuacja</i>	<i>situace</i>	94.12%	25.00%	0	0	10.492	1.312	0
<i>szkoła</i>	<i>škola</i>	91.43%	33.33%	0	0	11.387	1.898	0
<i>temat</i>	<i>téma</i>	86.67%	20.00%	0	0	9.077	1.815	1
<i>twarz</i>	<i>tvář</i>	26.67%	30.00%	0	0	7.184	1.437	0
<i>udział</i>	<i>úděl</i>	6.67%	50.00%	0	0	6.766	1.128	0
<i>usta</i>	<i>ústa</i>	80.00%	0.00%	0	0	2.787	0.697	0
<i>ustawa</i>	<i>stanovy</i>	5.88%	50.00%	1	0	18.222	2.278	1
<i>uwaga</i>	<i>úvaha</i>	36.67%	20.00%	0	0	3.534	0.707	0
<i>wiek</i>	<i>věk</i>	80.00%	25.00%	0	0	4.443	1.111	0
<i>władza</i>	<i>vláda</i>	93.33%	16.67%	0	0	5.339	0.890	0
<i>wniosek</i>	<i>návrh</i>	0.00%	87.50%	0	1	24.938	3.117	0

PL	CS	Intelligibility	Pron LD	FF	NC	WAS	WAS norm	Gender
woda	voda	100.00%	0.00%	0	0	1.718	0.430	0
wojna	vojna	90.00%	0.00%	0	0	4.180	0.836	0
wynik	výsledek	0.00%	62.50%	1	1	21.927	2.741	0
wzgląd	vzhled	36.67%	33.33%	0	0	5.371	0.895	0
zasada	zásada	96.67%	0.00%	0	0	7.064	1.177	0
zdanie	zdání	3.33%	16.67%	0	0	10.468	1.745	0
ziemia	země	60.00%	50.00%	0	0	8.533	1.422	0
zmiana	změna	63.33%	33.33%	0	0	7.441	1.240	0
związek	svazek	56.67%	35.71%	0	0	12.441	1.777	0
życie	život	33.33%	66.67%	0	0	13.396	2.233	1
Mean		55.03%	36.72%	14.29%	10.71%			10.71%

Table A 4: Intelligibility of the 100 most frequent PL nouns and their predictor variables.

The mean distance measures indicated in the last line of Table A 4 can differ from the lexical and orthographic distance measures from section 1.3 in this thesis and in Jágrová, Stenger, Marti & Avgustinova (2017), because 16 nouns that are identical in the two languages were not presented in the experiments and thus are not part of the analysis in Table A 4.

4.3. Free Translation of NPs

4.3.1. PL NPs for CS readers with the most representative data

Table A 5 lists the intelligibility of the NPs for which the most representative data could be collected (≥ 10 data points per NP and condition), their total distance and their lexical distance with regard to the category non-cognates or false friends.

	N	Intelligibility		Total dist	NC	FF
		AN	NA			
<i>publiczna</i>	<i>droga</i>	35.29%	33.33%	64.29%	0	1
<i>jedyne</i>	<i>dziecko</i>	94.12%	86.67%	15.38%	0	0
<i>ogromna</i>	<i>firma</i>	100.00%	93.33%	8.33%	0	0
<i>pełna</i>	<i>godzina</i>	31.25%	29.41%	25.00%	0	0
<i>europejska</i>	<i>komisja</i>	100.00%	93.75%	33.33%	0	0
<i>daleki</i>	<i>kraj</i>	100.00%	93.75%	0.00%	0	0
<i>chory</i>	<i>mężczyzna</i>	12.50%	33.33%	50.00%	0	0
<i>zielony</i>	<i>miesiąc</i>	75.00%	82.35%	28.57%	0	0
<i>poprzednia</i>	<i>możliwość</i>	25.00%	20.00%	50.00%	0	0
<i>dziwna</i>	<i>myśl</i>	75.00%	88.24%	15.00%	0	0
<i>małe</i>	<i>oko</i>	100.00%	86.67%	0.00%	0	0
<i>wielki</i>	<i>pan</i>	94.12%	81.25%	11.11%	0	0
<i>właściwa</i>	<i>pomoc</i>	25.00%	29.41%	34.62%	0	0
<i>zmianowa</i>	<i>praca</i>	0.00%	23.53%	30.77%	0	0
<i>istotna</i>	<i>prawda</i>	0.00%	0.00%	50.00%	1	1
<i>miejskie</i>	<i>prawo</i>	82.35%	53.33%	28.57%	0	0
<i>określony</i>	<i>procent</i>	0.00%	0.00%	56.25%	1	1
<i>potrzebny</i>	<i>przepis</i>	37.50%	29.41%	23.53%	0	0
<i>ciekawe</i>	<i>pytanie</i>	20.00%	13.33%	64.29%	1	1
<i>obca</i>	<i>rodzina</i>	6.25%	5.88%	57.14%	1	1
<i>polska</i>	<i>sprawa</i>	17.65%	13.33%	50.00%	1	1
<i>piękna</i>	<i>twarz</i>	88.24%	80.00%	27.27%	0	0
<i>dawny</i>	<i>udział</i>	6.25%	5.88%	27.27%	0	0
<i>komunikacyjny</i>	<i>węzeł</i>	81.25%	47.06%	25.00%	0	0
<i>krajowa</i>	<i>władza</i>	25.00%	17.65%	7.69%	0	0
<i>specjalny</i>	<i>wniosek</i>	0.00%	11.76%	60.00%	1	1
<i>szczególny</i>	<i>wzgląd</i>	0.00%	0.00%	66.67%	1	0
<i>duże</i>	<i>zdanie</i>	0.00%	11.76%	75.00%	1	1
<i>poszczególne</i>	<i>ziemie</i>	5.88%	6.25%	66.67%	1	0
<i>gospodarczy</i>	<i>związek</i>	82.35%	68.75%	33.33%	0	0
Mean		44.00%	41.31%	36.17%	30.00%	26.67%
SD		39.53%	34.38%	21.63%	46.61%	44.98%

Table A 5: Intelligibility of the 30 NPs with the most representative data and predictors.

4.3.2. PL NP Stimuli for German readers

Table A 6 lists the stimuli NPs and their intelligibility in the AN and NA condition together with the closest GER translations as transfer bases towards which the distance was calculated. Therefore, the words in the column Closest GER translation are given in lower case.

Stimuli PL		Intelligibility		Closest GER transfer base	
A	N	AN	NA	A	N
<i>amerykańska</i>	<i>firma</i>	88.89%	88.57%	<i>american</i>	<i>firma</i>
<i>amerykański</i>	<i>ojciec</i>	2.94%	6.06%	<i>american</i>	<i>father</i>
<i>amerykański</i>	<i>cel</i>	0.00%	3.03%	<i>american</i>	<i>ziel</i>
<i>euuropejska</i>	<i>szkoła</i>	78.13%	62.16%	<i>european</i>	<i>school</i>
<i>euuropejski</i>	<i>teren</i>	48.48%	44.44%	<i>european</i>	<i>terrain</i>
<i>euuropejski</i>	<i>szeף</i>	23.53%	15.15%	<i>european</i>	<i>chef</i>
<i>finansowa</i>	<i>grupa</i>	81.25%	64.86%	<i>financial</i>	<i>group</i>
<i>fizyczny</i>	<i>temat</i>	15.63%	13.51%	<i>physical</i>	<i>thema</i>
<i>fizyczny</i>	<i>problem</i>	20.00%	12.50%	<i>physical</i>	<i>problem</i>
<i>francuska</i>	<i>komisja</i>	9.38%	2.70%	<i>french</i>	<i>kommission</i>
<i>francuska</i>	<i>gmina</i>	50.00%	32.43%	<i>french</i>	<i>gemeinde</i>
<i>francuski</i>	<i>mężczyzna</i>	5.41%	0.00%	<i>französischer</i>	<i>man</i>
<i>francuskie</i>	<i>drzwi</i>	3.03%	0.00%	<i>französische</i>	<i>door</i>
<i>możliwa</i>	<i>decyzja</i>	0.00%	0.00%	<i>mögliche</i>	<i>decision</i>
<i>możliwe</i>	<i>procenty</i>	0.00%	3.03%	<i>mögliche</i>	<i>prozente</i>
<i>możliwy</i>	<i>stan</i>	2.70%	0.00%	<i>möglicher</i>	<i>stand</i>
<i>możliwy</i>	<i>punkt</i>	0.00%	0.00%	<i>möglicher</i>	<i>punkt</i>
<i>nowa</i>	<i>noc</i>	40.54%	41.18%	<i>new</i>	<i>nacht</i>
<i>nowa</i>	<i>akcja</i>	36.11%	11.43%	<i>new</i>	<i>aktion</i>
<i>nowe</i>	<i>oko</i>	5.41%	5.88%	<i>new</i>	<i>eye</i>
<i>nowy</i>	<i>projekt</i>	22.86%	28.13%	<i>new</i>	<i>projekt</i>
<i>nowy</i>	<i>prezes</i>	21.21%	22.22%	<i>new</i>	<i>president</i>
<i>nowy</i>	<i>tysiąc</i>	0.00%	3.13%	<i>new</i>	<i>thousand</i>
<i>nowy</i>	<i>dzień</i>	3.13%	2.70%	<i>new</i>	<i>day</i>
<i>nowy</i>	<i>miesiąc</i>	91.43%	75.00%	<i>new</i>	<i>monat</i>
<i>nowy</i>	<i>dom</i>	0.00%	0.00%	<i>new</i>	<i>house</i>
<i>polityczna</i>	<i>sytuacja</i>	57.14%	43.75%	<i>political</i>	<i>situation</i>
<i>polska</i>	<i>matka</i>	22.22%	25.71%	<i>polish</i>	<i>mother</i>
<i>polski</i>	<i>rzqd</i>	86.11%	91.43%	<i>polish</i>	<i>order</i>
<i>polski</i>	<i>minister</i>	0.00%	0.00%	<i>polish</i>	<i>minister</i>
<i>polskie</i>	<i>pieniqdze</i>	2.78%	2.86%	<i>polish</i>	<i>pennies</i>

Stimuli PL		Intelligibility		Closest GER transfer base	
A	N	AN	NA	A	N
<i>prywatna</i>	<i>policja</i>	55.88%	36.36%	<i>private</i>	<i>police</i>
<i>prywatna</i>	<i>rzecz</i>	3.03%	0.00%	<i>private</i>	<i>sache</i>
<i>prywatny</i>	<i>numer</i>	81.25%	54.05%	<i>private</i>	<i>number</i>
<i>prywatny</i>	<i>szpital</i>	37.14%	68.75%	<i>private</i>	<i>spital</i>
<i>publiczna</i>	<i>informacja</i>	69.70%	58.33%	<i>public</i>	<i>information</i>
<i>publiczny</i>	<i>program</i>	51.52%	63.89%	<i>public</i>	<i>program</i>
<i>rosyjska</i>	<i>rada</i>	0.00%	6.06%	<i>russian</i>	<i>council</i>
<i>rosyjski</i>	<i>film</i>	54.05%	44.12%	<i>russian</i>	<i>film</i>
<i>rosyjski</i>	<i>prezydent</i>	52.78%	51.43%	<i>russian</i>	<i>president</i>
<i>specjalna</i>	<i>ustawa</i>	2.70%	0.00%	<i>special</i>	<i>statut</i>
<i>specjalna</i>	<i>woda</i>	27.03%	41.18%	<i>special</i>	<i>water</i>
Mean		29.84%	26.81%		
SD		30.43%	27.89%		

Table A 6: NP stimuli presented to German readers, their correct DE and closest GER translations.

4.4. Highly Predictable Target Words

The following high-constraint, high cloze probability sentences were originally published as a resource by Block & Baldwin (2010). For the present study, the originally EN sentences were translated into PL in such a manner that the highly predictable target words remain on the last position in the sentences, although a translation variant with another word order might have been more appropriate in some sentences. The 149 stimuli sentences together with their original EN versions are made available under https://www.coli.uni-saarland.de/%7Etania//ta-pub/CICLing2019_PL_sentences_resource.xlsx. The PL sentences were presented to Czech respondents as stimuli in cloze translation experiments. The respondents were asked to translate only the last word in the PL sentence – the target word is in brackets in each sentence. The colour code in Table A 7 follows the code introduced in Table 53 (Chapter VI).

Stimuli PL as used in experiment	Intelligibility in %		Type of error in %			
	In context	No context	Context	Similar	Association	Interference Ln
Babka zapisała wszystko swojemu synowi w swoim {testamencie}.	83.33	100.00	6.67	0.00	0.00	0.00
Większość kotów bardzo dobrze widzi {nocą}.	96.67	100.00	0.00	0.00	0.00	0.00
Aby promować swój album, zespół udał się w {trasę}.	13.33	100.00	6.67	40.00	0.00	30.00
Sarah widziała zwierzęta z całego świata w {ZOO}.	100.00	100.00	0.00	0.00	0.00	0.00
Idąc przez ciemny pokój, uderzyłem się w nogę, w {palec}.	80.00	100.00	0.00	0.00	0.00	0.00
Jessie zaliczyła wyścig w powolnym {tempie}.	90.00	100.00	3.33	3.33	0.00	0.00
Benny próbował realizować nowe postanowienie co {roku}.	100.00	100.00	0.00	0.00	0.00	0.00
Ukończyła naukę jako najlepsza w swojej {klasie}.	96.77	75.76	0.00	0.00	3.23	0.00
Wiosna była Jo ulubioną porą {roku}.	100.00	100.00	0.00	0.00	0.00	0.00
Przyniósł swoją przynętę nad jezioro, żeby złowić {rybę}.	77.42	100.00	0.00	9.67	0.00	0.00
Chociaż Keith dobrze grał w kręgle, nie miał najwyższej liczby {punktów}.	58.06	100.00	9.67	3.23	0.00	16.13
Służąca starła kurz z książek na {regale}.	83.33	90.91	0.00	0.00	3.33	0.00
Że był wściekły, rozpoznała po tonie jego {głosu}.	93.33	26.67	0.00	0.00	0.00	3.33
Poszła do piekarni po bochen {chleba}.	100.00	100.00	0.00	0.00	0.00	0.00
Bob oświadczył się i dał jej diamentowy {pierścionek}.	90.00	45.45	0.00	0.00	6.67	0.00
Dentysta zaleca myć zęby dwa razy na {dzień}.	66.67	80.00	0.00	26.67	16.67	0.00
Połuźnił krawat na swojej {szyji}.	50.00	42.42	20.00	6.67	9.67	0.00
Zapłacili za swoje dania, ale zapomnieli zostawić {napiwku}.	10.00	6.06	0.00	36.67	3.33	0.00
Aby opłacić czesne za studia, wzięła dwa studenckie {kredyty}.	96.67	96.97	0.00	0.00	0.00	0.00
Nie miała przy sobie zegarka, więc zapytała o {czas}.	100.00	96.67	0.00	0.00	0.00	0.00
Sherry musiała zyczać z ust, ponieważ była {głucha}.	100.00	87.88	0.00	0.00	0.00	0.00
Siedzieli razem, nie mówiąc ani jednego {słowa}.	100.00	100.00	0.00	0.00	0.00	0.00
Uderzywszy w górę lodową, statek zaczął {tonać}.	30.00	27.27	23.33	3.33	30.00	0.00
Bez okularów słońce raniło Eriki {oczy}.	86.67	78.79	0.00	13.33	0.00	0.00
Kulejący koń odczuwał najwyraźniej duży {ból}.	16.67	27.27	0.00	0.00	0.00	60.00
Aby zapobiec kontuzjom w futbolu amerykańskim, wszyscy zawodnicy muszą nosić naramienne {ochraniacze}.	86.67	33.33	0.00	13.33	0.00	0.00
Po kłótni Ann poszła do swojego pokoju i trzysnęła {drzwiami}.	66.67	6.67	0.00	0.00	16.67	0.00
Woda i światło słoneczne pomagają roślinom {rosnąć}.	90.32	12.12	0.00	3.23	6.45	0.00
Nosiła kolorowy szal na swojej {szyji}.	74.19	42.42	6.45	6.45	3.23	0.00
Sportowiec lubi chodzić na podnoszenie ciężarów na {siłownię}.	58.06	30.30	0.00	22.58	12.90	0.00
Spodziewając się telefonu od Jeffa, czekała, że telefon {zadzwoni}.	96.77	93.94	0.00	0.00	0.00	0.00
Po wyjściu z samochodu, zamknęła {drzwi}.	93.55	6.67	0.00	3.23	0.00	3.23
Pod prysznicem umył się {mydłem}.	100.00	96.97	0.00	0.00	0.00	0.00
Potrzebowałbyś płaszcz przeciwdeszczowego, żebyś nie był {mokry}.	100.00	93.94	0.00	0.00	0.00	0.00
Na spotkanie Tom przyniósł na długiej łódzce {różę}.	80.65	93.94	0.00	12.9	0.00	0.00

Stimuli PL as used in experiment	Intelligibility in %		Type of error in %			
	In context	No context	Context	Similar	Association	Interference Ln
Rok po śmierci matki Bill odwiedził jej {grób}.	96.77	84.85	0.00	0.00	0.00	0.00
Dzieci wyszły na dwór, żeby się {bawić}.	96.77	51.52	0.00	0.00	0.00	0.00
Nauczyciel zapisał problem na {tablicy}.	87.01	72.73	3.23	9.68	0.00	0.00
W nocy stara kobieta zamknęła {drzwi}.	38.71	6.67	0.00	9.68	0.00	0.00
Jej praca była łatwa większą część {czasu}.	96.77	96.67	0.00	0.00	0.00	0.00
Gdy idziesz do łóżka, wyłącz {światło}.	96.77	93.94	0.00	3.23	0.00	0.00
Po jedzeniu umył ręce {mydłem}.	100.00	96.97	0.00	0.00	0.00	0.00
W cichym kinie telefon Kima {zadzwoił}.	100.00	93.94	0.00	0.00	0.00	0.00
W pomieszczeniu było ciemno, więc zapaliła {światło}.	93.55	93.94	6.45	0.00	0.00	0.00
Farmer spędził rano dojąc swoje {krowy}.	96.77	66.67	3.23	3.23	3.23	0.00
Po każdym jedzeniu dobrze jest umyć {zęby}.	74.19	42.42	22.58	0.00	0.00	0.00
W pokoju było głośno, tak że musiałam krzyczeć, żeby być {usłyszana}.	64.52	90.91	0.00	3.23	9.68	0.00
Bill poszedł do dentysty, żeby skontrolować swoje {zęby}.	96.77	42.42	0.00	0.00	3.23	0.00
Jenny zapaliła świeczki na urodzinowym {torcie}.	41.94	96.97	54.84	0.00	16.13	16.13
Gdy rozległ się alarm, strażak spuścił się w dół po {rurze}.	12.90	75.76	41.94	77.42	3.23	0.00
Na nocnym niebie lepiej widać wszystkie te {gwiazdy}.	100.00	87.88	0.00	0.00	0.00	0.00
Nie mogła kupić koszulki, bo nie {pasowała}.	83.87	60.61	9.68	0.00	6.45	0.00
Wpłacił swój nowy czek z wynagrodzeniem do {banku}.	100.00	86.67	0.00	0.00	0.00	0.00
Młode ptaki były gotowe, aby opuścić {gniazdo}.	96.67	100.00	3.33	3.33	0.00	0.00
Poszła do fryzjera, żeby ufarbować {włosy}.	93.33	96.97	0.00	3.33	0.00	0.00
Sukces jest często po prostu kwestią ciężkiej {pracy}.	50.00	93.33	0.00	43.33	0.00	0.00
W pierwsze pole proszę wpisać swoje {imię}.	70.00	33.33	16.67	3.33	0.00	0.00
Ponieważ jechała nocą zmniejsza widoczność, lepiej włączyć {światła}.	83.33	93.94	3.33	10.00	3.33	0.00
Miał na sobie grubą kurtkę, bo było {zimno}.	93.33	96.97	0.00	0.00	0.00	0.00
Dzieci wybiegły się {bawić}.	86.67	51.52	0.00	13.33	0.00	0.00
W Walentynki kobieta dostała jedną czerwoną {różę}.	100.00	93.94	0.00	0.00	0.00	0.00
Dziecko Parkerów mogło już powiedzieć trzy {słowa}.	93.33	100.00	0.00	0.00	0.00	0.00
Kartka urodzinowa była wesoła i doprowadziła mnie do {śmiechu}.	96.67	96.97	3.33	0.00	0.00	0.00
Nie mogłam przypomnieć sobie jego {imienia}.	80.00	33.33	20.00	20.00	0.00	0.00
Ray upadł i obdarł sobie {kolana}.	100.00	100.00	0.00	0.00	0.00	0.00
Po tym, jak wdychała dym z ognia, potrzebowała świeżego {powietrza}.	76.67	70.00	0.00	13.33	0.00	0.00
Student poszedł do biblioteki przeczytać {książkę}.	93.33	23.08	0.00	0.00	0.00	0.00
Aby założyć ogród, musisz najpierw wysiać {nasiona}.	33.33	0.00	30.00	16.67	13.33	3.33
Po tym, jak wdychała dym z ognia, potrzebowała świeżego {powietrza}.	93.33	70.00	0.00	0.00	0.00	0.00
Gdy niemowlęta są głodne, mogą często {płakać}.	86.67	73.33	0.00	6.67	0.00	0.00
Dan nazbierał więcej drzewa na {ognisku}.	96.67	100.00	0.00	0.00	0.00	0.00
Je w restauracjach, bo jest mizernym {kucharzem}.	100.00	100.00	0.00	0.00	0.00	0.00
Johnowi było przykro, ale to nie była jego {wina}.	100.00	76.67	0.00	0.00	0.00	0.00

Stimuli PL as used in experiment	Intelligibility in %		Type of error in %			
	In context	No context	Context	Similar	Association	Interference Ln
Wykład trwa około {godziny}.	100.00	46.67	0.00	0.00	0.00	0.00
Hodowali świnie w swoim {gospodarstwie}.	93.33	100.00	0.00	0.00	0.00	0.00
Dziecko z dobrze usytuowanej rodziny uczęszczało do prywatnej {szkoły}.	96.67	92.00	0.00	0.00	0.00	0.00
Jej nowe buty miały zły {rozmiar}.	96.67	53.33	0.00	0.00	3.33	0.00
Podczas świąt lepiej niż dostawać jest {dawać}.	80.00	80.00	3.33	3.33	0.00	0.00
Bradley woli koty od {psów}.	96.67	90.00	0.00	0.00	0.00	0.00
Na śniadanie Jim chciał boczek i {jajka}.	50.00	20.00	30.00	20.00	6.67	0.00
Młody ptak był gotów, żeby uczyć się {latać}.	93.33	20.83	0.00	0.00	6.67	0.00
Nie mogła pić kawy, bo była ona za {gorąca}.	30.00	16.67	46.67	13.33	13.33	0.00
Gdy Colin zobaczył dym, zadzwonił do straży pożarnej i zgłosił {pożar}.	96.67	96.67	0.00	0.00	0.00	0.00
Zauważyłam, że nie mam parasola, gdy zaczęło {padać}.	96.67	63.33	0.00	43.33	0.00	0.00
Wyjście było oznakowane dużym {napisem}.	100.00	93.33	0.00	0.00	0.00	0.00
Zrobiła sobie kanapkę i frytki na {obiad}.	100.00	93.33	0.00	0.00	0.00	0.00
Film był tak smutny, że publiczność {plakała}.	100.00	73.33	0.00	0.00	0.00	0.00
John wziął swojego psa na {spacer}.	81.25	16.67	3.13	0.00	3.13	3.13
Przeczcił stronę swojej ulubionej {książki}.	56.25	23.08	12.50	12.50	9.38	3.13
Pojechałbym, ale w moim aucie nie ma {paliwa}.	90.63	100.00	6.25	3.13	0.00	0.00
Księgowy uprasował koszulę, zanim poszedł do {pracy}.	100.00	93.33	0.00	0.00	0.00	0.00
Śmierci z ubiegłego tygodnia miały nieprzyjemny {zapach}.	100.00	100.00	0.00	0.00	0.00	0.00
Książniczka mogła poślubić tylko {księcia}.	6.25	10.00	12.50	18.75	18.75	6.25
Kłamała, że zgubiła świadectwo, żeby ukryć złe {oceny}.	0.00	9.09	3.23	74.19	6.45	0.00
Umyła brudne naczynia w {zlewozmywaku}.	60.00	6.06	36.67	33.33	0.00	0.00
Brudne talerze piętrzyły się w {zlewozmywaku}.	12.90	6.06	45.16	25.81	3.23	0.00
Chcąc mieć kolorowy pokój, kupił wiadro {farby}.	100.00	96.67	0.00	0.00	0.00	0.00
Wspaniała kelnerka otrzymała znakomity {napiwek}.	16.67	6.06	50.00	33.33	30.00	0.00
Ojciec rozebrał indyka {nożem}.	100.00	100.00	0.00	0.00	0.00	0.00
Przy kolacji pokroił swoje jedzenie {nożem}.	100.00	100.00	0.00	0.00	0.00	0.00
Lubiła grać na gitarze, więc dołączyła do {zespołu}.	20.00	3.03	13.33	63.33	3.33	0.00
John zamiótł podłogę {szczotką}.	3.13	0.00	9.38	28.13	28.13	0.00
Dana poproszono, aby został nowym coachem {zespołu}.	26.67	9.09	0.00	63.33	10.00	0.00
George musi swojego psa trzymać na {smyczy}.	75.00	3.33	12.50	9.38	9.38	0.00
Amber poszedł do salonu, żeby kupić nowy {samochód}.	65.63	34.78	28.13	3.13	0.00	0.00
Miała grypę i potrzebowała się napić gorącej {herbaty}.	53.13	16.67	28.13	9.38	0.00	12.50
Aby się uczyć, Karen usiadła przy swoim {biurku}.	37.50	0.00	25.00	25.00	3.13	3.13
Bill wskoczył do jeziora i zrobił wielki {plusk}.	30.00	0.00	26.67	3.33	6.67	0.00
Było wystarczająco wietrznie, żeby puścić {latawiec}.	32.26	3.03	25.81	22.58	25.81	0.00
Ellen lubi poezję, malarstwo i inne formy {sztuki}.	54.84	3.03	25.81	0.00	6.45	0.00
W tanim piórze szybko zabrakło {atramentu}.	9.68	9.09	0.00	22.58	25.81	0.00
Duch z butelki obiecał mężczyźnie spełnić jedno {życzenie}.	38.71	9.09	0.00	35.48	22.58	0.00

Stimuli PL as used in experiment	Intelligibility in %		Type of error in %			
	In context	No context	Context	Similar	Association	Interference Ln
Surfujący boją się, że zostaną ugryzieni przez {rekiną}.	10.00	3.33	60.00	66.67	0.00	0.00
Dobrym sposobem zachowania dobrej kondycji jest jazda na {rowerze}.	78.13	10.00	21.88	12.50	0.00	3.13
Nagrzała piekarnik i natłuściła {patelnię}.	18.75	0.00	65.63	12.50	3.13	0.00
Wysłał list bez {znaczką}.	70.97	3.03	0.00	16.13	9.68	0.00
Ponieważ błyskało, nie mogła iść na basen {pływać}.	83.33	18.18	6.67	0.00	3.33	0.00
Księżniczka pewnego dnia zostanie {królową}.	26.67	48.48	6.67	56.67	3.33	0.00
Aby zapłacić za samochód, Al po prostu wypisał {czek}.	100.00	33.33	0.00	0.00	0.00	0.00
Katie wsadziła kwiaty do drogiego {wazonu}.	50.00	15.15	43.33	40.00	10.00	0.00
Księżyc w pełni rozświetlił nocne {niebo}.	93.33	23.33	6.67	0.00	0.00	0.00
Kotek bawił się kłębkami {nici}.	51.61	3.03	9.68	32.26	6.45	0.00
Paczka została wysłana {pocztą}.	51.61	24.24	0.00	41.94	3.23	0.00
Gdy tych dwoje się spotkało, jedno z nich wyciągało {rękę}.	12.90	3.03	0.00	74.19	3.23	0.00
Mój ulubiony czas wiosną jest, gdy kwiaty {kwitną}.	66.67	16.67	0.00	26.67	6.67	0.00
Po zagrabieniu ogrodu Pat wskoczyła w górę {łłści}.	26.67	26.67	43.33	16.67	6.67	0.00
Rycerz szykował się do walki i wyciągnął swój {miecz}.	83.33	33.33	13.33	0.00	3.33	13.33
Jane powiesiła kolorowy obraz na {ścianie}.	96.67	16.67	0.00	0.00	0.00	0.00
Pierzaste białe obłoki są wysoko na {niebie}.	93.33	23.33	0.00	0.00	0.00	0.00
Na swoje urodziny Jan upiekł {ciasto}.	43.33	0.00	0.00	6.67	0.00	0.00
Najpierw kobieta odmówiła, ale zmieniała {zdanie}.	43.33	3.33	30.00	30.00	3.33	0.00
Uszkodzenie opony zmusiło Katie do zatrzymania się na skraju {drogi}.	3.33	3.03	16.67	63.33	3.33	3.33
Zimna woda na dworze oznaczała, że nadszedł czas, aby włączyć {ogrzewanie}.	26.67	9.09	10.00	46.67	10.00	0.00
Podczas kierowania samochodem powinnaś trzymać wzrok na {drodze}.	20.00	0.00	20.00	36.67	13.33	0.00
Po rozbiciu się okrętu marynarzowi urosła długa {broda}.	6.25	0.00	18.75	68.75	6.25	0.00
Joe nie był zadowolony z ubrania i postanowił się {przebrać}.	6.67	0.00	30.00	56.67	20.00	0.00
W pokoju było zimno, więc włączyli {ogrzewanie}.	45.16	9.09	3.23	48.39	3.23	0.00
Obawiał się pracy na nocnej {zmianie}.	93.33	0.00	3.33	0.00	3.33	0.00
Aby dowiedzieć się czegoś o swoich przodkach, narysowali genealogiczne {drzewo}.	36.67	0.00	3.33	56.67	0.00	0.00
Cid potrzebował paska, żeby przytrzymać swoje {spodnie}.	60.00	0.00	0.00	30.00	0.00	0.00
Wyszła za mąż tylko dla pieniędzy a nie z {miłości}.	16.67	6.67	66.67	66.67	10.00	0.00
Moderator w radiu poinformował o pilnym {doniesieniu}.	20.00	6.67	30.00	16.67	6.67	0.00
Każdej jesieni liście opadają z {drzew}.	73.33	0.00	6.67	20.00	0.00	0.00
Włączył kanał 13, żeby obejrzeć codzienne {wiadomości}.	3.23	6.06	0.00	93.55	3.23	0.00
Miał długi dzień i był w złym {nastroju}.	45.16	3.03	0.00	41.94	12.90	0.00
Dostarczyli projekt w terminie {wyznaczonym}.	29.03	0.00	9.68	61.29	0.00	0.00
Pianino brzmiało okropnie i {nie stroiło}.	10.00	23.33	63.33	13.33	3.33	0.00
Był tak pewny, że ten koń wyścigowy wygra, że zrobił {zakład}.	13.33	0.00	6.67	76.67	0.00	0.00
Aby zawiesić obraz Ted potrzebował młotka i {gwoździa}.	53.33	3.03	20.00	13.33	3.33	0.00
Mean	67.99	49.73	9.09	15.13	3.85	1.21

Table A 7: Sentences with highly predictable target words in cloze translation experiments.

5. Target Words in Highly Predictive Context Categorised as FFs

5.1. False Friends that are also Cognates—FF-C

PL target word	Inflected form (if different)	Frequently mistaken for	Mistaken		No answer	
			No context	In context	No context	In context
<i>czek</i> 'cheque'	identical	<i>čech</i> 'Czech person'	43.33%	0.00%	10.00%	0.00%
<i>królowa</i> 'queen'	<i>królową</i>	<i>králova</i> 'the king's [A]'	24.24%	40.00%	3.03%	10.00%
<i>kwitnąć</i> 'to bloom'	<i>kwitną</i>	<i>květináč</i> 'flowerpot'	73.33%	0.00%	0.00%	0.00%
<i>liście</i> 'leaves'	<i>liści</i>	<i>liška</i> 'fox'	36.67%	43.00%	16.67%	13.33%
<i>miecz</i> 'sword'	identical	<i>míč</i> 'ball'	40.00%	13.33%	3.03%	0.00%
<i>nić</i> 'thread'	<i>nici</i> [gen]	<i>nic</i> 'nothing'	90.91%	6.45%	0.00%	6.45%
<i>niebo</i> 'sky'	identical	<i>nebo</i> 'or'	73.33%	0.00%	3.03%	0.00%
	<i>niebie</i>			0.00%		3.33%
<i>plywać</i> 'to swim'	identical	<i>plivat</i> 'to spit'	51.52%	0.00%	15.15%	6.67%
<i>poczta</i> 'post (office)'	<i>poczty</i> [instr]	<i>pocta</i> 'honour'	63.64%	38.71%	3.03%	3.23%
<i>ręka</i> 'hand'	<i>rękę</i> [accu]	<i>řeka</i> 'river'	90.91%	64.52%	0.00%	9.68%
<i>ściana</i> 'wall'	<i>ścianie</i>	<i>scěna</i> 'scene'	50.00%	0.00%	13.33%	3.33%
<i>wazon</i> 'vase'	<i>wazonu</i>	<i>vagon</i> 'wagon'	48.48%	36.67%	18.18%	0.00%
<i>znaczek</i> 'stamp'	<i>znaczką</i>	<i>značka</i> 'sign'	69.70%	12.90%	3.03%	0.00%

PL target word	Variance of answers		Correct CS			Correct minus false	No answer + variance	(Answer + variance) - correct
	No context	In context		No context	In context			
<i>czek</i> 'cheque'	20.00%	0.00%	<i>šek</i>	33.33%	100.00%	-10.00%	30.00%	-13.33%
<i>królowa</i> 'queen'	21.21%	20.00%	<i>královna</i>	48.48%	26.67%	24.24%	24.24%	0.00%
<i>kwitnąć</i> 'to bloom'	13.33%	23.33%	<i>kvěst</i>	16.67%	66.67%	-56.67%	13.33%	-60.00%
<i>liście</i> 'leaves'	13.33%	23.00%	<i>listí</i>	26.67%	26.67%	-10.00%	30.00%	-6.67%
<i>miecz</i> 'sword'	23.33%	13.33%	<i>meč</i>	33.33%	83.33%	-6.67%	26.36%	-13.64%
<i>nić</i> 'thread'	12.12%	29.03%	<i>nit</i>	3.03%	51.61%	-87.88%	12.12%	-78.79%
<i>niebo</i> 'sky'	6.67%	10.00%	<i>nebe</i>	23.33%	93.33%	-50.00%	9.70%	-63.64%
		10.00%			93.33%		9.70%	
<i>plywać</i> 'to swim'	21.21%	13.33%	<i>plavat</i>	18.18%	83.33%	-33.33%	36.36%	-15.15%
<i>poczta</i> 'post (office)'	15.15%	16.13%	<i>pošta</i>	24.24%	51.61%	-39.39%	18.18%	-45.45%
<i>ręka</i> 'hand'	9.09%	19.35%	<i>ruka</i>	3.03%	12.90%	-87.88%	9.09%	-81.82%
<i>ściana</i> 'wall'	23.33%	6.67%	<i>stěna</i>	16.67%	96.67%	-33.33%	36.67%	-13.33%
<i>wazon</i> 'vase'	24.24%	23.33%	<i>váza</i>	15.15%	50.00%	-33.33%	42.42%	-6.06%
<i>znaczek</i> 'stamp'	21.21%	12.90%	<i>známka</i>	3.03%	70.97%	-66.67%	24.24%	-45.45%

Table A 8: Target words classified as FF-C.

5.2. False Friends that are Cognates in Another Context - FF-OC

PL target word	Inflected form (if different)	Frequently mistaken for	Mistaken		No answer		Variance of answers	
			No context	In context	No context	In context	No context	In context
<i>broda</i> 'beard'	identical	<i>brod</i> 'ford', <i>brada</i> 'chin'	70.00%	56.67%	20.00%	0.00%	16.67%	36.67%
<i>ciasto</i> 'cake'	identical	<i>často</i> 'often'	73.33%	6.67%	3.03%	0.00%	13.33%	13.33%
<i>droga</i> 'road'	<i>drogi</i>	<i>droga</i> 'drug', <i>lék</i> 'medicine'	78.79%	53.33%	3.03%	13.33%	15.15%	33.33%
	<i>drodze</i>			3.33%		13.33%		56.67%
<i>ogrzewanie</i> 'heating'	identical	<i>ohřívání</i> 'heating up [N]'	63.63%	29.03%	9.09%	3.23%	24.24%	35.48%
				33.33%		6.67%		40.00%
<i>przebrać</i> 'to change clothes'	identical	<i>přebat</i> 'to pick over'	60.61%	16.67%	6.06%	13.33%	30.30%	63.33%
<i>zdanie</i> 'opinion'	identical	<i>zdání</i> 'appearance'	56.67%	10.00%	3.33%	0.00%	23.33%	26.67%
<i>zmiana</i> 'shift'	<i>zmianie</i>	<i>změna</i> 'change'	83.33%	93.33%	0.00%	0.00%	16.67%	10.00%

PL target word	Inflected form (if different)	Correct CS				Correct - false	No answer + variance	(Answer + variance) - correct
		Present context	Cognate in other context	No context	In context			
<i>broda</i> 'beard'	ident cal	<i>vousy</i> 'beard'	<i>brada</i> 'chin'	0.00%	6.25%	-70.00%	36.67%	-33.33%
<i>ciasto</i> 'cake'	ident cal	<i>koláč</i> 'cake'	<i>těsto</i> 'dough'	6.67%	96.67%	-66.67%	15.36%	-56.97%
<i>droga</i> 'road'	<i>drogi</i>	<i>silnice</i> 'road'	<i>dráha</i> 'lane, track'	3.03%	3.33%	-75.76%	18.18%	-60.61%
	<i>drodze</i>				20.00%			
<i>ogrzewanie</i> 'heating'	ident cal	<i>topení</i> 'heating'	<i>ohřívání</i> 'heating up [N]'	9.09%	45.16%	-51.52%	33.33%	-30.30%
					26.67%			
<i>przebrać</i> 'to change clothes'	ident cal	<i>převléct se</i> 'to change'	<i>přebat</i> 'to pick over'	0.00%	6.67%	-60.61%	35.36%	-24.24%
<i>zdanie</i> 'opinion'	identical	<i>názor</i> 'opinion'	<i>zdání</i> 'impression'	3.33%	43.33%	-53.33%	26.67%	-30.00%
<i>zmiana</i> 'shift'	<i>zmianie</i>	<i>směna</i> 'shift'	<i>změna</i> 'change'	0.00%	93.33%	-83.33%	16.67%	-66.67%

Table A 9: Target words classified as FF-OC.

5.3. False Friends that Allow for Correct Associations - FF-A

PL target word	Inflected form if different from base form	Frequently mistaken for	Mistaken		No answer	
			No context	In context	No context	In context
<i>doniesienie</i> 'message'	<i>doniesieniu</i>	<i>donešení</i> 'delivery'	43.33%	6.67%	13.33%	23.33%
<i>drzewo</i> 'tree'	identical <i>drzew</i>	<i>dřevo</i> 'wood'	96.97%	56.67%	0.00%	3.33%
				15.00%		3.33%
<i>miłość</i> 'love'	<i>miłości</i>	<i>milost</i> 'mercy'	83.33%	60.00%	3.33%	6.67%
<i>spodnie</i> 'pants'	identical	<i>spodky</i> 'underpants'	80.00%	6.67%	3.33%	3.33%
<i>wiadomości</i> 'news'	identical	<i>vědomost(i)</i> 'knowledge'	75.76%	87.10%	3.03%	0.00%

PL target word	Variance of answers		Correct CS			Correct minus false	No answer + variance	(Answer + variance) - correct
	No context	In context	CS	No context	In context			
<i>doniesienie</i> 'message'	20.00%	40.00%	<i>zpráva</i>	6.67%	20.00%	-36.67%	33.33%	-10.00%
<i>drzewo</i> 'tree'	6.06%	13.33%	<i>strom</i>	0.00%	36.67%	-96.97%	6.06%	-90.91%
		20.00%			73.33%	-96.97%	6.06%	6.06%
<i>miłość</i> 'love'	16.67%	20.00%	<i>láska</i>	6.67%	16.67%	-76.67%	20.00%	-63.33%
<i>spodnie</i> 'pants'	20.00%	20.00%	<i>kalhoty</i>	0.00%	60.00%	-80.00%	23.33%	-56.67%
<i>wiadomości</i> 'news'	21.21%	16.13%	<i>zprávy</i> 'news'	6.06%	3.23%	-69.70%	24.24%	-51.52%

Table A 10: Target words classified as FF-A.

5.4. False Friends—FF

PL target word	Inflected form	Frequently mistaken for	Mistaken		No answer		Variance of answers	
			No context	In context	No context	In context	No context	In context
<i>nastrój</i> 'mood'	<i>nastroju</i>	<i>nástroj</i> 'instrument'	96.97%	38.71%	0.00%	0.00%	6.06%	32.26%
<i>wyznaczony</i> 'appointed'	<i>wyznaczonym</i>	<i>wyznačený</i> 'characterized'	54.55%	48.39%	15.15%	0.00%	21.21%	38.71%
<i>zakład</i> 'bet'	identical	<i>základ</i> 'base'	100.00%	70.00%	0.00%	6.67%	0.00%	12.90%
<i>stroić</i> 'to tune'	<i>nie stroiło</i>	<i>stroj</i> 'machine'	53.33%	3.33%	6.67%	16.67%	23.33%	36.67%
<i>gwóźdź</i> 'nail'	<i>gwóździa</i>	<i>hvozd</i> 'forest'	45.45%	0.00%	18.18%	10.00%	24.24%	20.00%

PL target word	Correct CS			Correct - false	No answer + variance	(Answer + variance) - correct
	CS	No context	In context			
<i>nastrój</i> 'mood'	<i>nálada</i>	3.03%	45.16%	-93.94%	6.06%	-90.91%
<i>wyznaczony</i> 'appointed'	<i>určený</i>	40.91%	29.03%	-13.64%	36.36%	-18.18%
<i>zakład</i> 'bet'	<i>sázka</i>	0.00%	12.90%	-100.00%	0.00%	-100.00%
<i>stroić</i> 'to tune'	<i>ladit</i>	23.33%	10.00%	-30.00%	30.00%	-23.33%
<i>gwóźdź</i> 'nail'	<i>hřebík</i>	3.03%	53.33%	-42.42%	42.42%	-3.03%

Table A 11: Target words categorised as FF.

6. Monolingual Cloze Tests

6.1. Task in Monolingual Cloze Tests

You will be presented with about 30 sentences containing gaps. Your task will be to fill these gaps with whatever you spontaneously consider best.

There is no wrong or right. Even if some of the gaps could be filled with anything, please write down what seems most appropriate to you or what you think of first.

Completing the questionnaire will take approx. 10 minutes.

Thank you in advance for your participation!

This survey is part of a dissertation within the framework of the linguistic research project INCOMSLAV – Mutual Intelligibility and Surprisal in Slavic Intercomprehension – at Saarland University. More information and links can be found in the imprint.

Klára Jágrová
Doctoral Researcher

Figure A 3: Instruction as presented to respondents in the monolingual cloze test (EN version).

6.2. Stimuli

The PL stimuli in Table A 12 and their CS translations (Table A 13) were presented in regular monolingual online cloze tests to Polish and Czech respondents, respectively. The respondents were asked to fill the gaps with whatever they find most suitable. No time limit was given. The sentences appeared in the same order as they are listed here. The cloze probability is the percentage of the most frequent response.

Condition	Stimuli	Cloze probability	Most frequent response	Cloze probability of word in original PL stimulus
1	Gdyby nie było ... , czytałbym Ci z oczu.	21.88%	<i>ciemno</i>	18.75%
2	Gdyby nie było książek,	21.21%	<i>nie byłoby wiedzy (o wydarzeniach z minionych lat)</i>	0.00%
3	Gdyby nie było książek, czytałbym Ci z	15.63%	<i>oczcu</i>	15.63%
4	Gdyby nie było książek, czytałbym ... z oczu.	27.59%	<i>ludziom</i>	6.90%
1	W 2000 roku wzrósł do ponad 900 mln. marek obrót ... , w procesie produkcji których nie używano substancji zagrażających środowisku naturalnemu wilka.	25.00%	<i>drewna / drewnem</i>	6.25%
2	W 2000 roku wzrósł do ponad 900 mln. marek obrót towarów , w procesie produkcji których nie używano substancji ... środowisku naturalnemu wilka.	78.79%	<i>szkodzących, szkodliwych, zagrażających</i>	78.79%
3	W 2000 roku wzrósł do ponad 900 mln. marek obrót towarów, w procesie produkcji których nie używano substancji zagrażających środowisku naturalnemu	15.63%	<i>Niemiec</i>	0.00%
1	Kolegium dało mi pozwolenie, aby zrealizować ten projekt nad	21.88%	<i>rzeką, Wisłą, Odrą</i>	6.25%
2	Kolegium dało mi pozwolenie, aby zrealizować ten projekt	15.63%	<i>natychmiast(owo)</i>	0.00%
1	Praga to ważny ... komunikacyjny.	28.13%	<i>węzeł</i>	28.13%
2	Praga to ... węzeł komunikacyjny.	21.88%	<i>główny</i>	18.18%
1	Czy pani będzie ... ? Czy chcielibyście, aby stały się one gwiazdami?	15.63%	<i>ślawna</i>	0.00%
2	Czy pani będzie głosowała? Czy chcielibyście, aby stały się one ... ?	28.13%	<i>ważne</i>	0.00%
3	Czy pani ... głosowała? Czy chcielibyście, aby stały się one gwiazdami?	28.13%	<i>już</i>	3.13%
1	Kupiliśmy nie tylko czerstwy chleb, ale jeszcze gorzej - też stary żółty	96.88%	<i>ser</i>	0.00%
2	Kupiliśmy nie tylko ... chleb, ale jeszcze gorzej - też stary żółty samochód.	15.63%	<i>stary</i>	24.24%
3	Kupiliśmy nie tylko czerstwy chleb, ale jeszcze ... - też stary żółty samochód.	9.38%	<i>dodatkowo</i>	0.00%
4	Kupiliśmy nie tylko czerstwy chleb, ale jeszcze gorzej - też stary ... samochód.	31.03%	<i>zardzewiały</i>	0.00%
1	Teraz rosną również ... odbycia interesujących praktyk w kraju.	53.13%	<i>możliwości / szanse</i>	53.13%
2	Teraz ... również możliwości odbycia interesujących praktyk w kraju.	96.88%	<i>istnieją / (wiele) istnieje</i>	0.00%
1	Nie widziałam, że jego żona pokazuje ręką, żebyśmy poszli do	25.00%	<i>sklepu</i>	0.00%
2	Nie widziałam, że jego żona pokazuje ... , żebyśmy poszli do rektora.	53.13%	<i>nam</i>	6.06%
3	Nie ... , że jego żona pokazuje ręką, żebyśmy poszli do rektora.	18.75%	<i>widział(em)</i>	18.75%
1	Skąd jesteś ... , że za pięćdziesiąt lat ludzie nie będą już latali samolotem?	68.75%	<i>pewny/-a / upewniony</i>	9.38%

Condition	Stimuli	Cloze probability	Most frequent response	Cloze probability of word in original PL stimulus
2	Skąd jesteś przekonana, że za pięćdziesiąt lat ludzie nie będą już latali ... ?	25.00%	<i>samolotem, samolotami</i>	25.00%
3	... jesteś przekonana, że za pięćdziesiąt lat ludzie nie będą już latali samolotem?	59.38%	<i>czy</i>	9.38%
1	OCZEKIWANIA: ... w pracy przy produkcji mięsa; pełna dyspozycyjność od poniedziałku do piątku.	37.50%	<i>doświadczenie</i>	37.50%
2	OCZEKIWANIA: doświadczenia w pracy przy produkcji ... ; pełna dyspozycyjność od poniedziałku do piątku.	68.75%	<i>maszyn</i>	3.03%
3	OCZEKIWANIA: doświadczenia w pracy przy produkcji mięsa; ... dyspozycyjność od poniedziałku do piątku.	28.13%	<i>pełna</i>	28.13%
1	OBŚŁUGA ... - ZAKRES OBOWIĄZKÓW: znajomość języka polskiego; ekspozycja towarów; gotowość do pracy zmianowej; czynności porządkowe	40.63%	<i>klienta / klientów</i>	28.13%
2	OBŚŁUGA SKLEPU - ZAKRES OBOWIĄZKÓW: znajomość języka polskiego; ekspozycja towarów; gotowość do pracy ... ; czynności porządkowe	18.18%	<i>zmianowej, na zmiany</i>	18.18%
3	OBŚŁUGA SKLEPU - ZAKRES OBOWIĄZKÓW: znajomość języka polskiego; ... towarów; gotowość do pracy zmianowej; czynności porządkowe	25.00%	<i>wykładanie, wystawianie, rozkładanie</i>	25.00%
1	NAPÓJ Z MIĘTY I MIODU: mięta zielona suszona: 25g; miód kwiatowy: 50g; cytryna: 1 szt.; ... konsumpcyjny: 5 kostek; sok z brzozy: 100ml; jarzębiny: 50g.	40.63%	<i>cukier</i>	21.88%
2	NAPÓJ Z MIĘTY I MIODU: mięta zielona suszona: 25g; miód kwiatowy: 50g; cytryna: 1 szt.; lód konsumpcyjny: 5 kostek; ... z brzozy: 100ml; jarzębiny: 50g.	40.63%	<i>sok, wyciąg, ekstrakt</i>	40.63%
3	NAPÓJ Z MIĘTY I MIODU: mięta zielona suszona: 25g; miód kwiatowy: 50g; cytryna: 1 szt.; lód konsumpcyjny: 5 kostek; sok z ... : 100ml; jarzębiny: 50g.	34.38%	<i>cytryny</i>	0.00%
4	NAPÓJ Z MIĘTY I MIODU: mięta zielona suszona: 25g; miód kwiatowy: 50g; ... : 1 szt.; lód konsumpcyjny: 5 kostek; sok z brzozy: 100ml; jarzębiny: 50g.	34.48%	<i>cytryna</i>	34.48%
1	OFERTA: realne ... awansu w firmie; 12,00 brutto/godzinę + premie miesięczne.	46.15%	<i>szanse</i>	46.15%
2	OFERTA: realne możliwości awansu w firmie; 12,00 brutto za ... + premie miesięczne.	90.91%	<i>godzinę</i>	90.91%
1	Zakaz palenia wyrobów ... w pojeździe.	96.88%	<i>tytoniowych</i>	96.88%
2	Zakaz palenia ...	36.36%	<i>papierosów</i>	24.24%
1	Poszła do ... i kupiła znaczek.	46.88%	<i>sklepu</i>	46.88%
2	Poszła do sklepu i kupiła ...	21.21%	<i>chleb</i>	0.00%
3	Poszła ... i kupiła znaczek.	81.25%	<i>na pocztę</i>	0.00%

Condition	Stimuli	Cloze probability	Most frequent response	Cloze probability of word in original PL stimulus
1	... zabiera max. 65 osob i dysponuje dwoma pokładami.	34.38%	<i>autokar / autobus</i>	28.13%
2	Statek zabiera max. 65 osob i dysponuje dwoma ...	46.88%	<i>pokładami, piętrami</i>	46.88%
3	Statek zabiera max. 65 osob i dysponuje ...	6.90%	<i>szalupami</i>	0.00%
1	W czasie pracy klimatyzacji okna są ...	90.63%	<i>zamknięte / pozamykane</i>	90.63%
1	Biuro Trojmiasto.pl - Wynajem ... - Nowoczesne przestrzenie - Wyjątkowe lokalizacje	34.38%	<i>mieszkań</i>	18.75%
2	Biuro Trojmiasto.pl - Wynajem biur - Nowoczesne przestrzenie - ... lokalizacje	34.38%	<i>dogodne</i>	3.03%
1	... , szalunki, zsypy do gruzu, sprzęt budowlany, sprzedaż	15.63%	<i>rusztowania</i>	15.63%
2	rusztowania, szalunki, zsypy do gruzu, sprzęt budowlany, ...	15.63%	<i>betoniarki</i>	0.00%
1	PROSZĘ NIE ... LIZAKÓW!	25.00%	<i>lizać</i>	0.00%
2	PROSZĘ NIE WYCIĄGAĆ ... I CUKIERKÓW!	90.63%	<i>batoników, batonów</i>	6.06%
3	PROSZĘ NIE WYCIĄGAĆ ... !	18.75%	<i>telefonów / telefonu / komórek</i>	0.00%
1	NOWOŚĆ DO PRANIA – Czystość, która ... pachnie - NOWY ZAPACH	34.38%	<i>pięknie</i>	34.38%
2	NOWOŚĆ DO PRANIA – Czystość, która pięknie ... - NOWY ZAPACH	34.38%	<i>pachnie</i>	34.38%
3	NOWOŚĆ DO PRANIA – Czystość, która pięknie pachnie - NOWY ...	25.00%	<i>zapach</i>	25.00%
4	NOWOŚĆ DO PRANIA – Czystość, która pięknie pachnie - NOWA ...	31.03%	<i>formuła</i>	0.00%
1	... MISIE – SMAK RADOŚCI – BEZ SZTUCZNYCH BARWNIKÓW	71.88%	<i>ZŁOTE / żelkowe / gumowe</i>	71.88%
2	ZŁOTE MISIE – ... RADOŚCI – BEZ SZTUCZNYCH BARWNIKÓW	15.63%	<i>dużo, wiele</i>	12.12%
3	ZŁOTE MISIE – SMAK RADOŚCI – BEZ ... BARWNIKÓW	71.88%	<i>sztucznych</i>	71.88%
1	POWIERZCHNIA ... DO WYNAJĘCIA	43.75%	<i>powierzchnia</i>	43.75%
2	POWIERZCHNIA REKLAMOWA DO ...	25.00%	<i>wynajęcia</i>	25.00%
1	Poszła do sklepu, żeby kupić jabłka i ...	25.00%	<i>banany</i>	3.13%
2	Poszła list bez ...	60.00%	<i>znaczka</i>	60.00%
1	Zakaz palenia e- ...	100.00%	<i>papierosów</i>	100.00%
1	Daleko jest mój ...	78.13%	<i>dom</i>	78.13%
2	Daleko jest moja ...	71.88%	<i>ojczyzna</i>	6.06%
1	Czy ktoś z was ma doświadczenie, co jest najlepsze jako podkładka pod ... przed przyczepą kempingową na kempingu?	18.75%	<i>namiot</i>	3.13%
2	Czy ktoś z was ma doświadczenie, co jest najlepsze jako podkładka pod stół przed ... na kempingu?	43.75%	<i>namiot / namiotem</i>	9.09%
1	22 grudnia 1882 – pierwsza elektrycznie oświetlona ... świata.	36.92%	<i>ulica</i>	15.38%

Condition	Stimuli	Cloze probability	Most frequent response	Cloze probability of word in original PL stimulus
1	A tutaj mamy następne wydanie naszego ... Technet.	65.63%	<i>czasopisma / magazynu</i>	65.63%
2	A tutaj mamy następne ... naszego magazynu Technet.	100.00%	<i>wydanie</i>	100.00%
1	Według tego przepisu możecie państwo przygotować dowolne ...	62.50%	<i>danie / dania</i>	0.00%
2	Według tego przepisu możecie państwo przygotować ...	13.79%	<i>ciasto</i>	0.00%
1	W wilgotnej ... żywność szybko się psuje.	31.25%	<i>lodowce</i>	12.50%

Table A 12: Results from the monolingual cloze test (PL).

Condition	Stimuli	Cloze probability	Most frequent response	Cloze probability of stimulus in original PL
1	Kdyby nebylo ... , četl bych Ti z očí.	14.71%	<i>knížek / knížky / knih</i>	14.71%
2	Kdyby nebylo knížek, ...	8.57%	<i>nebylo by rozumu.</i>	0.00%
3	Kdyby nebylo knížek, četl bych Ti z ...	18.18%	<i>ruky, novin</i>	6.06%
4	Kdyby nebylo knížek, četl bych ... z očí.	31.25%	<i>lidem</i>	9.38%
1	V roce 2000 narostl obrat ... , u kterého při procesu výroby není užíváno látek ohrožujících životní prostředí vlka, na více než 900 mil. marek.	20.59%	<i>produktů / zboží / výrobků</i>	20.59%
2	V roce 2000 narostl obrat zboží, u kterého při procesu výroby není užíváno látek ... životní prostředí vlka, na více než 900 mil. marek.	74.29%	<i>škodlivých, ohrožujících, poškozujících, ničících, narušujících</i>	74.29%
3	V roce 2000 narostl obrat zboží, u kterého při procesu výroby není užíváno látek ohrožujících životní prostředí ... , na více než 900 mil. marek.	12.12%	<i>planety</i>	0.00%
1	Kolegium mi dalo povolení, abych zrealizoval ten projekt u ...	8.82%	<i>vás</i>	20.59%
3	Kolegium mi dalo povolení, abych zrealizoval ten projekt ...	18.18%	<i>okamžitě</i>	0.00%
1	Praha je významný komunikační ...	35.29%	<i>uzel</i>	35.29%
2	Praha je ... komunikační uzel.	37.14%	<i>hlavní</i>	17.14%
1	Paní, budete ... ? Chtěli byste, aby ony se staly hvězdami?	17.65%	<i>hlasovat</i>	17.65%
2	Paní, budete hlasovat? Chtěli byste, aby ony se staly ... ?	22.86%	<i>vítězkami, výherci, výherkyně, výtězy</i>	2.86%
3	Paní, ... hlasovat? Chtěli byste, aby ony staly hvězdami?	39.39%	<i>budete</i>	39.39%

Condition	Stimuli	Cloze probabilitly	Most frequent response	Cloze probability of stimulus in original PL
1	Koupili jsme nejen tvrdý chléb, ale ještě hůř – také staré žluté ...	23.53%	<i>máslo</i>	0.00%
2	Koupili jsme nejen ... chléb, ale ještě hůře - též staré žluté auto.	60.00%	<i>starý, tvrdý, okoralý, tuhý</i>	60.00%
3	Koupili jsme nejen tvrdý chléb, ale ještě ... - též staré žluté auto.	6.06%	<i>chléb, chleba</i>	6.06%
4	Koupili jsme nejen tvrdý chléb, ale ještě hůře - též staré ... auto.	34.38%	<i>rezavé</i>	3.13%
1	Nyní rovněž rostou ... zajímavých praxí v zemi.	38.24%	<i>možnosti, příležitosti</i>	26.47%
2	Nyní rovněž ... možnosti zajímavých praxí v zemi.	31.43%	<i>nabízí(me)</i>	2.86%
1	Neviděla jsem, že jeho žena ukazuje rukou, abychom šli za ...	44.12%	<i>ní</i>	0.00%
2	Neviděla jsem, že jeho žena ukazuje ... , abychom šli k rektorovi.	14.29%	<i>na nás</i>	5.71%
3	Ne-... , že jeho žena ukazuje rukou, abychom šli k rektorovi.	21.21%	<i>forms of vidět</i>	21.21%
1	Proč jsi ... , že za padesát let lidé již nebudou létat letadlem?	26.47%	<i>řekl/řikal, myslíš/myslíte/(si) myslel(a),</i>	5.88%
2	Proč jsi přesvědčená, že za padesát let lidé již nebudou létat ... ?	48.57%	<i>letadlem, letadly, letadla, v letadlech</i>	48.57%
3	... jsi přesvědčená, že za padesát let lidé již nebudou létat letadlem?	48.48%	<i>opravdu</i>	12.12%
1	POŽADAVKY: pracovní ... při zpracování masa, ochota pracovat od pondělka do pátku.	20.59%	<i>nasazení / zkušenost(i)</i>	20.59%
2	POŽADAVKY: pracovní zkušenosti při zpracování ... , ochota pracovat od pondělka do pátku.	34.29%	<i>dat</i>	14.29%
3	POŽADAVKY: pracovní zkušenosti při zpracování masa, být k ... dispozici od pondělka do pátku.	21.21%	<i>okamžitě, neustále</i>	3.03%
1	OBSLUHA ... – ROZSAH POVINNOSTÍ: znalost polského jazyka, vystavování zboží, ochota práce na směny, udržování pořádku.	35.29%	<i>obchodu</i>	35.29%
2	OBSLUHA OBCHODU - ROZSAH POVINNOSTÍ: znalost polského jazyka, vystavování zboží, ochota práce ... , udržování pořádku.	22.86%	<i>přesčas</i>	17.14%
3	OBSLUHA OBCHODU - ROZSAH POVINNOSTÍ: znalost polského jazyka, ... zboží, ochota práce na směny, udržování pořádku.	54.55%	<i>znalost</i>	3.03%
1	NÁPOJ Z MÁTY A MEDU: sušená zelená máta: 25g; květový med: 50g; citron: 1 kus; konzumní ... 5 kostek; šťáva z břízy: 100ml; jeřabiny: 50g.	50.00%	<i>cukr</i>	11.76%
2	NÁPOJ Z MÁTY A MEDU: sušená zelená máta: 25g; květový med: 50g; citron: 1 kus; konzumní led: 5 kostek; ... z břízy: 100ml; jeřabiny: 50g.	42.86%	<i>extrakt, výtažek, šťáva</i>	42.86%
3	NÁPOJ Z MÁTY A MEDU: sušená zelená máta: 25g; květový med: 50g; citron: 1 kus; konzumní led: 5 kostek; šťáva z ... : 100ml; jeřabiny: 50g.	27.27%	<i>citrónu</i>	0.00%

Condition	Stimuli	Cloze probability	Most frequent response	Cloze probability of stimulus in original PL
4	NÁPOJ Z MÁTY A MEDU: sušená zelená máta: 25g; květový med: 50g; ... : 1 kus; konzumní led: 5 kostek; stáva z břízy: 100ml; jeřabiny: 50g.	28.13%	<i>citrón</i>	28.13%
1	NABÍZÍME: reálná ... postupu ve firmě, 12,00 hrubého/hodinu + měsíční prémie	64.71%	<i>možnost, příležitost</i>	64.71%
2	NABÍZÍME: reálná možnost postupu ve firmě, 12,00 hrubého za ... + měsíční prémie	57.14%	<i>hodinu</i>	57.14%
1	Zákaz kouření ... výrobků ve vozidle.	64.71%	<i>tabákových</i>	64.71%
2	Zákaz kouření ...	17.14%	<i>v restauraci</i>	11.43%
1	Zašla do ... a koupila známku.	47.06%	<i>trafiky</i>	38.24%
2	Zašla do obchodu a koupila ...	14.29%	<i>chléb</i>	0.00%
3	Zašla ... a koupila známku.	63.64%	<i>na poštu</i>	3.03%
1	... pobere max. 65 osob a disponuje dvěma palubami.	61.76%	<i>lod'</i>	61.76%
2	Loď pobere max. 65 osob a disponuje dvěma ...	25.71%	<i>motory</i>	8.57%
3	Loď pobere max. 65 osob a disponuje dvěma ...	27.27%	<i>palubami</i>	27.27%
4	Loď pobere max. 65 osob a disponuje ...	9.38%	<i>záchrannými čluny</i>	9.38%
1	Při spuštěné klimatizaci jsou okna ...	97.06%	<i>zavřené</i>	97.06%
1	Kancelář Trojměstí – Pronájem ... – Moderní prostory – Výjimečné lokality.	20.59%	<i>bytů</i>	14.71%
2	Kancelář Trojměstí – Pronájem kanceláří – Moderní prostory – ... lokality.	11.43%	<i>dobré</i>	8.57%
1	..., bednění, shozy na suť, stavební úklid, prodej.	17.65%	<i>lešení</i>	17.65%
2	lešení, bednění, shozy na suť, stavební úklid, ...	12.12%	<i>míchačka</i>	0.00%
1	Prosím ne- ... lízátka.	17.65%	<i>lízat / olizovat / lízej / ližte</i>	0.00%
2	Prosím nevytahovat ... a bonbóny!	25.71%	<i>lízátka</i>	25.71%
3	Prosím nevytahovat ... !	15.15%	<i>zbraně</i>	0.00%
1	NOVINKA NA PRANÍ - Čistota, která ... voní. NOVÁ VŮNĚ	17.65%	<i>krásně</i>	5.88%
2	NOVINKA NA PRANÍ - Čistota, která pěkně ... NOVÁ VŮNĚ	62.86%	<i>(za-/pro-)voní</i>	62.86%
3	NOVINKA NA PRANÍ - Čistota, která pěkně voní. NOVÁ ...	33.33%	<i>aviváž</i>	12.12%
4	NOVINKA NA PRANÍ - Čistota, která pěkně voní. NOVÝ ...	28.13%	<i>prášek</i>	0.00%
1	... MEDVÍDČI – PŘÍCHUŤ RADOSTI – BEZ UMĚLÝCH BARVIV	91.18%	<i>gumoví / želatinoví / Haribo</i>	91.18%
2	ZLATÍ MEDVÍDČI – ... RADOSTI – BEZ UMĚLÝCH BARVIV	42.86%	<i>plno, plní, spousta</i>	0.00%
3	ZLATÍ MEDVÍDČI – PŘÍCHUŤ RADOSTI – BEZ ... BARVIV	54.55%	<i>umělých</i>	54.55%
1	REKLAMOVÁ ... K PRONÁJMU	82.35%	<i>plocha</i>	82.35%
2	REKLAMOVÁ PLOCHA K ...	94.29%	<i>pronájmu</i>	94.29%
1	Zašla do obchodu koupit jablka a ...	61.76%	<i>hrušky</i>	0.00%
2	Poslala dopis bez ...	71.43%	<i>známky</i>	71.43%

Condition	Stimuli	Cloze probability	Most frequent response	Cloze probability of stimulus in original PL
1	Zákaz kouření e- ...	100.00%	<i>cigaret</i>	100.00%
1	Daleko je můj ...	73.53%	<i>domov / dům / byt</i>	73.53%
2	Daleko je moje ...	17.14%	<i>láska vs. domovina</i>	0.00%
1	Máte někdo zkušenosti, co je nejvhodnější pod ... na sezení venku v kempu u karavanu?	32.35%	<i>zadek</i>	2.94%
2	Máte někdo zkušenosti, co je nejvhodnější pod stůl na sezení venku v kempu u ... ?	60.00%	<i>ohně, ohniště</i>	2.86%
1	22. prosince 1882 – první elektricky osvětlený vánoční ...	94.12%	<i>strom</i>	94.12%
2	22. prosince 1882 – první elektricky osvětlený ...	22.86%	<i>dům</i>	14.29%
1	A máme tu další číslo Technet ...	29.41%	<i>magazínu</i>	29.41%
2	A máme tu další ... Technet magazínu.	74.29%	<i>vydání, číslo</i>	74.29%
1	Podle tohoto receptu můžete připravit libovolnou ...	11.76%	<i>buchtu</i>	8.82%
4	Podle tohoto receptu můžete připravit libovolné ...	25.00%	<i>množství</i>	0.00%
1	Ve vlhkém ... se potraviny rychle kazí.	91.18%	<i>prostředí</i>	0.00%

Table A 13: Results from the monolingual cloze test (CS).

7. Correlations and Statistical Models

7.1. Intelligibility of the 100 Most Frequent PL Ns

The selected model is marked grey.

Model		Coefficient	SE	t	p	Adjusted R ²	F crit	F
1	Pron LD	-0.996	0.114	-8.682	< 0.0001	0.622	< 0.0001	46.594
	FF	-0.475	0.078	-6.09	< 0.0001			
	NC	0.338	0.109	3.098	< 0.01			
2	Pron LD	-0.787	-0.098	-8.064	< 0.0001	0.582	< 0.0001	58.850
	FF	-0.424	-0.080	-5.288	< 0.0001			
3	Pron LD	-0.988	0.138	-7.157	< 0.0001	0.598	< 0.0001	42.081
	FF	-0.444	0.079	-5.595	< 0.0001			
	Norm WAS	0.095	0.047	2.019	< 0.05			
4	Pron LD	-1.065	0.12	-8.911	< 0.0001	0.633	< 0.0001	36.730
	FF	-0.473	0.077	-6.148	< 0.0001			
	NC	0.335	0.108	3.119	< 0.01			
	Gender	0.163	0.091	1.797	< 0.1			

Table A 14: Regression models: intelligibility of the 100 most frequent PL Ns.

7.2. Intelligibility of NPs for Czech Readers—AN Condition

The selected model is marked grey.

AN	Coefficient	SE	t	p	Adjusted R ²	F crit	F
Total dist	-1.294	0.252	-5.135	0.000	0.520	0.000	16.685
SURP A+N	0.018	0.029	0.625	0.537			
NC total	0.082	0.258	0.319	0.752	0.491	0.001	6.586
FF	-0.245	0.225	-1.092	0.286			
Gender	0.106	0.252	0.418	0.680			
Total dist	-1.161	0.451	-2.572	0.017			
Surp A+N	0.032	0.033	0.968	0.343	0.507	0.000	8.469
NC total	0.065	0.250	0.258	0.798			
FF	-0.206	0.200	-1.027	0.314			
Total dist	-1.143	0.442	-2.587	0.016			
Surp A+N	0.027	0.030	0.896	0.379	0.511	0.000	11.108
NC total	0.042	0.248	0.169	0.867			
FF	-0.160	0.193	-0.831	0.414			
Total dist	-1.239	0.427	-2.905	0.007			
NC total	-0.480	0.193	-2.485	0.019	0.377	0.001	9.756
FF	-0.075	0.215	-0.346	0.732			
NC total	-0.537	0.112	-4.771	0.000	0.461	0.000	13.392
Trad LD	-0.767	0.367	-2.088	0.046			
NC total	-0.090	0.189	-0.478	0.637	0.517	0.000	16.506
Total dist	-1.185	0.419	-2.827	0.009			
NC total	-0.101	0.192	-0.526	0.603	0.506	0.000	10.917
Total dist	-1.106	0.441	-2.508	0.019			
Surp A+N	0.019	0.030	0.658	0.516			
FF	-0.172	0.150	-1.145	0.263			
Total dist	-1.066	0.320	-3.326	0.003	0.525	0.000	11.689
Surp A+N	0.027	0.030	0.891	0.381			
NC total	-0.373	0.185	-2.021	0.054	0.465	0.000	9.418
FF	-0.234	0.211	-1.112	0.276			
Pron LD	-0.906	0.386	-2.344	0.027			
NC total	0.042	0.248	0.169	0.867	0.511	0.000	11.108
FF	-0.160	0.193	-0.831	0.414			
Total dist	-1.239	0.427	-2.905	0.007			

Table A 15: Regression models: NP translation experiments—AN condition.

7.3. Intelligibility of NPs for Czech Readers—NA Condition

The selected model is marked grey.

NA	Coefficient	SE	t	p	Adjusted R ²	F crit	F
Total dist	-1.044	0.196	-5.320	0.000	0.583	0.000	21.307
Surp	0.051	0.021	2.391	0.024			
NC total	-0.155	0.198	-0.782	0.442	0.609	0.000	10.050
FF	-0.177	0.161	-1.100	0.282			
Gender	0.229	0.191	1.199	0.242			
Total dist	-0.613	0.355	-1.725	0.097			
Surp	0.068	0.022	3.045	0.006	0.603	0.000	11.993
NC total	-0.170	0.200	-0.849	0.404			
FF	-0.108	0.151	-0.712	0.483			
Total dist	-0.586	0.358	-1.638	0.114			
Surp	0.060	0.022	2.794	0.010	0.499	0.000	10.612
NC total	-0.045	0.219	-0.204	0.840			
FF	-0.117	0.170	-0.687	0.498			
Total dist	-0.938	0.376	-2.497	0.019			
NC total	-0.440	0.165	-2.671	0.013	0.401	0.000	10.723
FF	-0.052	0.184	-0.282	0.780			
NC total	-0.479	0.098	-4.888	0.000	0.458	0.000	13.267
Trad LD	-0.547	0.320	-1.710	0.099			
NC total	-0.141	0.166	-0.850	0.403	0.508	0.000	15.995
Total dist	-0.899	0.368	-2.444	0.021			
NC total	-0.259	0.154	-1.688	0.103	0.610	0.000	16.128
Total dist	-0.548	0.350	-1.564	0.130			
Surp	0.061	0.021	2.837	0.009			
FF	-0.189	0.117	-1.616	0.118	0.607	0.000	15.922
Total dist	-0.810	0.239	-3.385	0.002			
Surp	0.056	0.021	2.686	0.012			
FF	-0.139	0.128	-1.086	0.287	0.516	0.000	16.482
Total dist	-0.994	0.255	-3.905	0.001			

Table A 16: Regression models: NP translation experiments—NA condition.

7.4. Intelligibility of Target Words in Highly Predictive Context

	Intell. in context	Intell. without context	Cloze prob	Surp target	2gram surp	2gram drop	3gram surp	3gram drop	Sum surp sent	Mean surp sent
	PL									
Correct in context										
Correct without context	0.663									
Pron LD target	0.641	0.767								
Total dist target	0.680	0.772								
Cloze probability	0.051									
Surp PL target	0.186		0.006							
2gram surp PL	0.140		0.084							
2gram drop PL	0.121		0.098							
3gram surp PL	0.093		0.054							
3gram drop PL	0.125		0.064							
Sum surp sentence PL	0.215		0.068							
Mean surp sentence PL	0.211		0.069							
Sum pron LD sentence	0.372									
Mean pron LD sentence	0.400									
Surp CS target	0.191		0.014	0.554						
2gram surp CS	0.137		0.077	0.527						
2gram drop CS	0.083		0.114	0.658						
3gram surp CS	0.086		0.061	0.529						
3gram drop CS	0.113		0.096	0.602						
Surp sentence CS	0.150		0.088	0.732						
Mean surp sentence CS	0.156		0.018	0.535						
Sum total dist sentence	0.462									
Mean total dist sentence	0.476									
Mean pron LD 2gram	0.588									
Mean pron LD 3gram	0.534									
Mean total dist 2gram	0.621									
Mean total dist 3gram	0.582									
n words	0.064		0.071							
FF		0.522								
Lex dist		0.553								
Association		0.300								
FF / sentence	0.353									
NC / sentence	0.508									
FF / words	0.287									
NC / words	0.496									
Gender	0.272	0.281								

Table A 17: Correlation matrix (Pearson's r): intelligibility of target words with and without context and the different predictors.

Legend for Table A 17:

Abbreviation in table	Explanation
Pron LD target	Pron LD of the target word
Total dist target	If target word is non-cognate, total distance is 1; otherwise pron LD.
Surp PL target	Target word's surprisal as scored by the PL LM
2gram surp PL	Target word's + its preceding word's surprisal as scored by the PL LM
2gram drop PL	Preceding word's surprisal - target word's surprisal as scored by the PL LM
3gram surp PL	Target word's + its two preceding words' surprisal as scored by the PL LM
3gram drop PL	Two preceding words' surprisal - target word's surprisal as scored by the PL LM
Sum surp sentence PL	Sum of all surprisal values in a sentence as scored by the PL LM
Mean surp sentence PL	Mean surprisal value in a sentence as scored by the PL LM
Sum pron LD sentence	Sum of all pron LD values in a sentence
Mean pron LD sentence	Mean pron LD of a sentence
Surp CS target	Target word's surprisal as scored by the CS LM
2gram surp CS	Target word's + its preceding word's surprisal as scored by the CS LM
2gram drop CS	Preceding word's surprisal - target word's surprisal as scored by the CS LM
3gram surp CS	Target word's + its two preceding words' surprisal as scored by the CS LM
3gram drop CS	Two preceding words' surprisal - target word's surprisal as scored by the CS LM
Surp sentence CS	Sum of all surprisal values in a sentence
Mean surp sentence CS	Mean surprisal value in a sentence
Sum total dist sentence	Sum of all total distance values in a sentence
Mean total dist sentence	Mean total distance of a sentence
Mean pron LD 2gram	Mean pron LD of the target word and its preceding word
Mean pron LD 3gram	Mean pron LD of the target word and its two preceding words
Mean total dist 2gram	Mean total dist of the target word and its preceding word
Mean total dist 3gram	Mean total dist of the target word and its two preceding words
n words	Number of words per sentence
FF	False friend; Binary category: 0 – target word is no FF; 1 – target word is FF
Lex dist	0 – target word is non-cognate; 1 – target word is cognate in another context; 2 – target word is cognate
Association	0 – target word does not lead to correct association; 1 – target word allows correct association
FF / sentence	Number of false friends in sentence
NC / sentence	Number of non-cognates in sentence
FF / words	Number of false friends in sentence divided by n words
NC / words	Number of non-cognates in sentence divided by n words
Gender	1 if gender of target is different in PL and CS

Table A 18: Legend containing the abbreviations used in Table A 16.

7.5. Model for Intelligibility of Target Words in Highly Predictive Context

The selected model is marked grey.

Model		Coefficient	SE	<i>t</i>	<i>p</i>	Adjusted R^2	<i>F crit</i>	<i>F</i>
1	Total dist target	-0.492	0.073	-6.730	< 0.0001	0.500	< 0.0001	30.623
	Surp sentence PL	-0.008	0.003	-2.269	< 0.05			
	NC	-0.044	0.022	-2.026	< 0.05			
	Gender	-0.085	0.078	-1.092	0.277			
	FF / sentence	-0.058	0.038	-1.497	0.137			
2	Total dist target	-0.549	0.066	-8.363	< 0.0001	0.496	< 0.0001	49.522
	Surp sentence PL	-0.008	0.003	-2.437	< 0.05			
	NC	-0.043	0.022	-1.982	< 0.05			

Table A 19: Regression models: Intelligibility of highly predictable target words.

7.6. Model for Intelligibility of the Target Words Without Context

The selected model is marked grey.

Model		Coefficient	SE	t	p	Adjusted R ²	F crit	F
1	Total dist target	-0.796	0.086	-9.260	< 0.0001	0.641	< 0.0001	53.941
	Association	0.042	0.103	0.404	0.687			
	Cognate	-0.030	0.047	-0.638	0.525			
	Gender	-0.083	0.086	-0.966	0.336			
	FF	-0.241	0.051	-4.708	< 0.0001			
2	Association	0.008	0.103	0.082	0.935	0.642	< 0.0001	53.978
	Cognate	0.048	0.042	1.138	0.257			
	FF	-0.240	0.051	-4.684	< 0.0001			
	Gender	-0.024	0.087	-0.277	0.782			
	Dist target	-0.849	0.093	-9.179	< 0.0001			
3	Cognate	0.046	0.035	1.330	0.186	0.644	< 0.0001	67.940
	FF	-0.241	0.051	-4.737	< 0.0001			
	Gender	-0.021	0.081	-0.267	0.790			
	Dist target	-0.850	0.092	-9.261	< 0.0001			

Table A 20: Regression models: Intelligibility of the target words without context.

REFERENCES

- Bernardy, P., Lappin, S., & Lau, J. H. (2018). The influence of context on sentence acceptability judgements. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers)* (pp. 456-461). Melbourne, Australia: Association for Computational Linguistics. Retrieved from <https://pdfs.semanticscholar.org/d183/f5fba3bb1bfd2b5564cad5f7aebd3b1a1f3f.pdf>
- Berthele, R. (2011). On abduction in receptive multilingualism. Evidence from cognate guessing tasks [Print + Online]. In L. Wei (Ed.), *Applied linguistics review* (2nd ed., pp. 191-220). Berlin, New York: de Gruyter. doi:10.1515/9783110239331.191.
- Bidwell, C. E. (1963). *Slavic historical phonology in tabular form*. The Hague: Mouton & Co.
- Block, C. K., & Baldwin, C. L. (2010). Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. *Behavior Research Methods*, 42(3), 665-670. doi:10.3758/BRM.42.3.665
- Bloom, P. A., & Fischler, I. (1980). Completion norms for 329 sentence contexts. *Memory & Cognition*, 8(6), 631-642.
- Broda, B., & Piasecki, M. (2010). Parallel, massive processing in supermatrix – a general tool for distributional semantic analysis of corpora. In M. Ganzha & M. Paprzycki (Eds.), *Proceedings of the International Multiconference on Computer Science and Information Technology* (pp. 373-379). Wisła: Polskie Towarzystwo Informatyczne. doi:10.1504/ijdm.2013.051924
- Carlton, T. R. (1991). *Introduction to the phonological history of the Slavic languages*. Columbus, Ohio: Slavica Publishers, INC.
- Čermák, F., & Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 13(3), 411-427.
- Cetnarowska, B., Pysz, A., & Trugman, H. (2011). Accounting for some flexibility in a rigid construction. In P. Bański, B. Łukaszewicz, M. Opalińska & J. Zaleska (Eds.), *Generative investigations: Syntax, morphology and phonology* (pp. 24-57). Newcastle upon Tyne: Cambridge Scholars Publishing
- Cetnarowska, B. (2013). The representational approach to adjective placement in Polish. *Linguistica Silesiana*, 34, 7-22. ISSN 0208-4228

- Crocker, M., Demberg V., & Teich, E. (2015). Information density and linguistic encoding (IDEAL). *Künstliche Intell*, 30, 77-81. doi:10.1007/s13218-015-0391-y
- Cvrček, V., & Vondříčka, P. (2011a). SyD – Korpusový průzkum variant [Corpus analysis of variants]. Prague: FF UK. Retrieved from <http://syd.korpus.cz>
- Cvrček, V., & Vondříčka, P. (2011b). Výzkum variability v korpusech češtiny [Analysis of variability in corpora of Czech]. In F. Čermák (Ed.), *Korpusová lingvistika. 2. Výzkum a výstavba korpusů* (pp. 184-195). Prague: NLN.
- Czech National Corpus*: InterCorp (version 9). Retrieved February 03, 2017 from <http://treq.korpus.cz>
- Czech National Corpus*: Srovnávací frekvenční seznamy [Comparative frequency lists]. Ústav Českého národního korpusu FF UK, Praha, 2010. Retrieved January 01, 2016. <http://ucnk.ff.cuni.cz/srovnani10.php>
- Dankovičová, J. (1999). Czech. In International Phonetic Association (Eds.), *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet* (pp. 70-74). Cambridge, U.K.: Cambridge University Press.
- Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109, 193-210.
- Doyé, P. (2005). *Intercomprehension. Guide for the development of language education policies in Europe: from linguistic diversity to plurilingual education*. Reference study. Strasbourg: DG IV, Council of Europe
- Fischer, A., Jágrová, K., Stenger, I., Avgustinova, T., Klakow, D., & Marti, R. (2015). An orthography transformation experiment with Czech-Polish and Bulgarian-Russian. In B. Sharp, W. Lubaszewski & R. Delmonte (Eds.), *Natural Language Processing and Cognitive Science 2015 Proceedings* (pp. 115-126). Venezia: Libreria Editrice Cafoscarina.
- Fischer, A., Jágrová, K., Stenger, I., Avgustinova, T., Klakow, D., & Marti, R. (2016). Orthographic and morphological correspondences between related Slavic languages as a base for modeling of mutual intelligibility. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, ... S. Piperidis. (Eds.), *Language Resources and Evaluation Conference LREC 2016* (pp. 4202-4209), including linguistic resources. Paris: European Language Resources Association.
- Golubović, J. (2016). *Mutual intelligibility in the Slavic language area*. Groningen: University of Groningen

- Golubović, J., & Gooskens, C. S. (2015). Mutual intelligibility between West and South Slavic languages. *Russian Linguistics*, 39(3), 351-373. doi:10.1007/s11185-015-9150-9
- Gooskens, C. S. (2007). The contribution of linguistic factors to the intelligibility of closely related languages. *Journal of Multilingual and Multicultural Development*, 28(6), 445-467. doi:10.2167/jmmd511.0
- Gooskens, C. S. (2013). Methods for measuring intelligibility of closely related language varieties. In R. Bayley, R. Cameron & C. Lucas (Eds.), *The Oxford handbook of sociolinguistics* (pp. 195-213). Oxford: University Press. doi:10.1093/oxfordhb/9780199744084.013.0010
- Gooskens, C. S., & Swarte, F. (2017). Linguistic and extra-linguistic predictors of mutual intelligibility between Germanic languages. *Nordic Journal of Linguistics*, 40(2), 123-147. doi:10.1017/S0332586517000099
- Grupa Technologii Językowych G4.19 Politechniki Wrocławskiej. (2016). *Lista frekwencyjna* [Frequency list]. Retrieved September 08, 2016. <http://www.nlp.pwr.wroc.pl/narzedzia-i-zasoby/zasoby/lista-frekwencyjna>
- Gulan, T., & Valerjev, P. (2010). Semantic and related types of priming as a context in word recognition. *Review of Psychology*, 17(1), 53-58.
- Harley, T. (2007). *The psychology of language – from data to theory* (2nd ed.). New York/Hove: Psychology Press. <http://www.al-edu.com/wp-content/uploads/2014/05/Harley-Psychology-of-Language-From-Data-to-Theory.pdf>
- Havránek, B. (1964). *Slovník spisovného jazyka českého* [Dictionary of standard Czech]. Prague: Československá akademie věd, sekce jazyka a literatury.
- Heeringa, W., Golubovic, J., Gooskens, C. S., Schüppert, A, Swarte, F., & Voigt, S. (2013). Lexical and orthographic distances between Germanic, Romance and Slavic languages and their relationship to geographic distance. In C. S. Gooskens & R. van Bezoijen (Eds.), *Phonetics in Europe: Perception and production* (pp. 99-137). Frankfurt a. M.: Peter Lang.
- Heeringa, W., Swarte, F., Schüppert, A, & Gooskens, C. S. (2014). 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. doi:10.1017/S1470542714000166
- Heinz, C. (2009). Semantische Disambiguierung von false friends in slavischen L3: die Rolle des Kontexts [Semantic disambiguation of false friends in Slavic L3: the role of context]. *ZfSl* 54(2), 147-166.

- Heinz, C., & Kuße, H. (2015). *Slawischer Sprachvergleich für die Praxis* [Comparison of the Slavic languages in practice]. *Specimina philologiae Slavicae* 179. Leipzig: Biblion Media
- Hilton, N. H., Gooskens, C. S. & Schüppert, A. (2013). The influence of non-native morphosyntax on the intelligibility of a closely related language. *Lingua*, 137, 1-18. doi:10.1016/j.lingua.2013.07.007
- International Organization for Standardization. (1988). *Documentation – Transliteration of Slavic Cyrillic characters into Latin characters* (ISO Standard No. 9:1986). Retrieved from <https://www.iso.org/standard/3589.html>
- International Phonetic Association. (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge, U.K.: Cambridge University Press.
- Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cogn Psychol*, 61, 23-62. doi:10.1016/j.cogpsych.2010.02.002
- Jágrová, K. (2010). *Russisch-Deutsch-Tschechische Interferenzen* [Interferences between Russian, German, and Czech] (Unpublished state examination thesis). TU Dresden.
- Jágrová, K. (2016, December). The role of different factors for the intelligibility of written Polish for Czech readers. Paper presented at FDSL 12, Berlin
- Jágrová, K. (2018). Processing effort of Polish NPs for Czech readers – A+N vs. N+A. In W. Guz & B. Szymanek (Eds.), *Canonical and non-canonical structures in Polish. Studies in linguistics and methodology* (Vol. 12, pp. 123-143). Lublin: Wydawnictwo KUL.
- Jágrová, K., & Avgustinova, T. (2019). Intelligibility of highly predictable Polish target words in sentences presented to Czech readers. To appear in *Proceedings of CICLing: International Conference on Intelligent Text Processing and Computational Linguistics*. http://www.coli.uni-saarland.de/~tania/ta-pub/CICLing_preprint_Jagrova_Avgustinova_2019.pdf
- Jágrová, K., Avgustinova, T., Stenger, I., & Fischer, A. (2019). Language models, surprisal and fantasy in Slavic intercomprehension. *Computer Speech and Language* 53. 242-275. doi:10.1016/j.csl.2018.04.005
- Jágrová, K., & Stenger, I. (2019, September). *Čechische und russische Kognaten übersetzt von serbischen Studierenden* [Czech and Russian cognates translated by Serbian students]. Paper presented at 13. Deutscher Slavistentag, Trier

- Jágrová, K., Stenger, I., & Avgustinova, T. (2017). Polski nadal nieskomplicowany? Interkomprehensionsexperimente mit Nominalphrasen [Is Polish still uncomplicated? Intercomprehension experiments with noun phrases]. *Polnisch in Deutschland. Zeitschrift der Bundesvereinigung der Polnischlehrkräfte*, 5, 20-37.
- Jágrová, K., Stenger, I., & Avgustinova, T. (2019, September). *Slavische Interkomprehensionsmatrix* [Slavic intercomprehension matrix]. Poster presented at 13. Deutscher Slavistentag, Trier
- Jágrová, K., Stenger, I., Avgustinova, T., & Marti, R. (2016). Polski to język nieskomplicowany? Theoretische und praktische Interkomprehension der 100 häufigsten polnischen Substantive [Is Polish an uncomplicated language? Theoretical and practical intercomprehension of the 100 most frequent Polish nouns]. *Polnisch in Deutschland. Zeitschrift der Bundesvereinigung der Polnischlehrkräfte*, 4, 5-19.
- 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 J. Edmonds & M. Janebová (Eds.), *Proceedings of the Olomouc Linguistics Colloquium 2016: Olomouc Modern Language Series* (Vol. 5, pp. 401-416). Olomouc: Palacký University. <http://olinco.upol.cz/wp-content/uploads/2017/06/olinco-2016-proceedings.pdf>
- Karlík, P., Nekula, M., & Pleskalová, J. (2002). *Encyklopedický slovník češtiny* [Encyclopedic dictionary of Czech]. Prague: Nakladatelství lidové noviny
- Kazojć, J. (2010). *Otwarty słownik czesko-polski* [Open Czech-Polish dictionary]. V.03.2010 (c). Retrieved April 22, 2015 from <http://www.slowniki.org.pl/czesko-polski.pdf>
- Keller, F. (2010). Cognitively plausible models of human language processing. In *Proceedings of the ACL 2010 Conference Short Papers (ACLShort '10)* (pp. 60-67). Stroudsburg, PA: Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/P10-2012.pdf>
- Kneser, R., & Ney, H. (1995). Improved backing-off for M-gram language modeling. In *International Conference on Acoustics, Speech, and Signal Processing 1995* (Vol. 1, pp. 181-184). Detroit, MI: IEEE doi:10.1109/ICASSP.1995.479394
- Kosek, P. (2014). *Historická mluvnice češtiny – překlenovací seminář* [Historical grammar of Czech – bridging seminar]. Brno: Masarykova Univerzita. Retrieved from <https://digilib.phil.muni.cz/data/handle/11222.digilib/131101/monography.pdf>

- Křen, M. (2010). *Srovnávací frekvenční seznamy* [Comparative frequency lists]. Prague: Institute of the Czech National Corpus, Faculty of Arts, Charles University. Retrieved from <http://ucnk.ff.cuni.cz/index.php>.
- Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., ... Zásina, A. (2015). *SYN2015: reprezentativní korpus psané češtiny* [SYN2015: a representative corpus of written Czech]. Prague: Ústav Českého národního korpusu FF UK. <http://www.korpus.cz>
- Leiner, D. J. (2019). SoSci Survey (Version 3.1.06) [Computer software]. Available at <https://www.soscisurvey.de>
- Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory* 10, 707-710.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126-1177.
- Mayring, P. (2010). *Qualitative Inhaltsanalyse. Grundlagen und Technik* [Qualitative content analysis. Basics and techniques] (11th ed.). Weinheim: Beltz.
- Moberg, J., Gooskens, C. S., Nerbonne, J., & Vaillette, N. (2006). Conditional entropy measures intelligibility among related languages. In P. Dirix, I. Schuurman, V. Vandeghinste, & F. van Eynde (Eds.), *Computational linguistics in the Netherlands 2006: Selected papers from the 17th CLIN Meeting* (pp. 51-66). Utrecht: LOT. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.463.6211&rep=rep1&type=pdf>
- Möller, R., & Zeevaert, L. (2015). Investigating word recognition in inter-comprehension: Methods and findings. *Linguistics*, 53(2), 313-352. doi:10.1515/ling-2015-0006
- Muikku-Werner, P. (2014). Co-text and receptive multilingualism – Finnish students comprehending Estonian. *Nordic Journal of Linguistics*, 40(2), 99-113. doi:10.12697/jeful.2014.5.3.05
- Nábělková, M. (2007). Closely-related languages in contact: Czech, Slovak, “Czechoslovak”. *International Journal of the Sociology of Language*, 183, 53-73. doi:10.1515/IJSL.2007.004
- Obolonchikova, V. (2017). *Exploring syntactic distances for closely related Slavic languages* (Unpublished master’s thesis). Saarland University, Faculty of Mathematics and Computer Science.
- Rayner, K., & Well, A. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychon Bull Rev*, 3(4), 504-509. doi:10.3758/BF03214555

- Ringbom, H. (2007). *Cross-linguistic similarity in foreign language learning*. Clevedon: Multilingual Matters LTD.
- Reich, I., & Horch, E. (2017). *The Fragment Corpus (FraC)*. In *Proceedings of the 9th International Corpus Linguistics Conference*, Birmingham: University of Birmingham. Retrieved from <http://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper194.pdf>
- Schenker, A. M. (1993). Proto-Slavonic. In B. Comrie & G. G. Corbett (Eds.), *The Slavonic languages* (pp. 60-125). London/New York: Routledge.
- Sekerina, I. A., Laurinavichyute, A., Alexeeva, S., Bagdasaryan, K., & Kliegl, R. (2018). Russian Sentence Corpus: Benchmark measures of eye movements in reading in Russian. *Behavior Research Methods*, 51(3), 1161-1178. doi:10.3758/s13428-018-1051-6
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst Tech J* 27(379-423), 623-656.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Sikos, L., Greenberg, C., Drenhaus, H., & Crocker, M. (2017). Information density of encodings: The role of syntactic variation in comprehension. In *CogSci 2017: Annual Meeting of the Cognitive Science Society* (pp. 3168-3173). London. Retrieved from https://www.researchgate.net/profile/Les-Sikos/publication/346829439_Information_density_of_encodings_The_role_of_syntactic_variation_in_comprehension/links/5fd142a592851c00f8621752/Information-density-of-encodings-The-role-of-syntactic-variation-in-comprehension.pdf
- Šimandl, J. (2003). Od čtvrtku do pátku [From Thursday to Friday]. *Naše řeč*, 86(3), 161-164. Retrieved from <http://nase-rec.ujc.cas.cz/archiv.php?art=7737>
- Škrabal, M., & Vavřín, M. (2017a). The Translation Equivalents Database (Treq) as a lexicographer's aid. In I. Kosem et al. (Eds.), *Electronic Lexicography in the 21st Century: Proceedings of ELEX 2017 Conference* (pp. 124-137). Leiden: Lexical Computing CZ s. r. o.
- Škrabal, M., & Vavřín, M. (2017b). Databáze překladových ekvivalentů Treq [Translation equivalents database Treq]. *Časopis pro moderní filologii*, 99(2), 245-260.

- Šmerk, P., Pravdová, M., Beneš, M., Černá, A., Hlaváčková, A., Chromý, J., ... Uhlířová, L. (2009). *The internet language reference book*. Prague: LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. <http://hdl.handle.net/11858/00-097C-0000-0023-8BD2-2>
- Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition* 128(3), 302-319. doi:10.1016/j.cognition.2013.02.013
- Stenger, I., Avgustinova, T., & Marti, R. (2017). Levenshtein distance and word adaptation surprisal as methods of measuring mutual intelligibility in reading comprehension of Slavic languages. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue 2017* (pp. 304-317). Moscow: RSUH. Retrieved from <http://www.dialog-21.ru/media/3953/stengerietal.pdf>
- Stenger, I., Jágrová, K., Fischer, A., & Avgustinova, T. (2020). "Reading Polish with Czech eyes" or "How Russian can a Bulgarian text be?": Orthographic differences as an experimental variable in Slavic intercomprehension. In T. Radeva-Bork & Kosta, P. (Eds.), *Current developments in Slavic linguistics. Twenty years after (based on selected papers from FDSL II)* (pp. 483-500). Bern: Peter Lang. doi:10.3726/978-3-653-07147-4
- Stenger, I., Jágrová, K., Fischer, A., Avgustinova, T., Klakow, D., & Marti, R. (2017). Modelling the impact of orthographic coding on Czech-Polish and Bulgarian-Russian reading intercomprehension. *Nordic Journal of Linguistics*, 40(2), 175-199. doi:10.1017/S0332586517000130
- Szałek, M., & Nečas, J. (1993). *Czesko-Polska homonimia* [Czech-Polish homonymy]. Poznań: Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza.
- van Bezooijen, R., & Gooskens, C. S. (2005). Intertalig tekstbegrip. De begrijpelijkheid van Friese en Afrikaanse teksten voor Nederlandse lezers [Interlingual text comprehension. The intelligibility of Frisian and Afrikaans texts for Dutch readers]. *Nederlandse Taalkunde*, 10(2), 129-152.
- van Heuven, V. J., Gooskens, C. S. & van Bezooijen, R. (2015). Introducing Micrela: Predicting mutual intelligibility between closely related languages in Europe. In J. Navracics & S. Batyi (Eds.), *First and second language: Interdisciplinary approaches (Studies in psycholinguistics 6)* (pp. 127-45). Budapest: Tinta könyvkiado.

- Vanhove, J. (2014). *Receptive multilingualism across the lifespan. Cognitive and linguistic factors in cognate guessing* (Doctoral Dissertation, University of Fribourg, Switzerland). Retrieved from <https://core.ac.uk/download/pdf/20663762.pdf>
- Vanhove, J. (2015). The early learning of interlingual correspondence rules in receptive multilingualism. *International Journal of Bilingualism*, 20(5), 580-593. doi:10.1177/1367006915573338
- Vanhove, J., & Berthele, R. (2015). Item-related determinants of cognate guessing in multilinguals. In G. De Angelis, U. Jessner, & M. Kresić (Eds.), *Crosslinguistic influence and crosslinguistic interaction in multilingual language learning* (pp. 95-118). London: Bloomsbury. Retrieved from <https://core.ac.uk/download/pdf/43669306.pdf>
- Vasmer, M. (1964-1973). *Etimologičeskij slovar' ruskogo jazyka* [Etymological dictionary of Russian] (Vols. 1-4). Moscow: Progress.
- Vavřin, M., & Rosen, A. (2015). Treq. Prague: FF UK. Retrieved from <http://treq.korpus.cz>
- Žuravlev, A. F. (1974-2012). *Etimologičeskij slovar' slavjanskich jazykov. Praslavjanskij leksičeskij fond* [Etymological dictionary of Slavic languages. Proto-Slavic Lexical Stock] (Vols. 1-37). Moscow: Nauka.

Experiment software

The experiment website was developed in the scope of the INCOMSLAV project – Mutual Intelligibility and Surprisal in Slavic Intercomprehension – within the DFG-funded CRC 1102: Information Density and Linguistic Encoding at Saarland University.

Available from <http://intercomprehension.coli.uni-saarland.de>

Online documents

Swadesh list:

http://en.wiktionary.org/wiki/Appendix:Swadesh_lists_for_Slavic_languages
Retrieved April 22, 2015

Pan-Slavic list:

<http://www.eurocomslav.de/kurs/pwslav.htm>
Retrieved April 22, 2015

Internationalism list:

<http://www.eurocomslav.de/kurs/iwslav.htm>
Retrieved April 22, 2015

