

Radovi Zavoda za informacijske studije

Knjiga 22.



Recenzenti:

prof. dr. sc. Miroslav Tuđman

prof. dr. sc. Nenad Prelog

Izdavač:

Zavod za informacijske studije,

Odsjeka za informacijske i komunikacijske znanosti,

Filozofskog fakulteta Sveučilišta u Zagrebu.

Za izdavača:

Sonja Špiranec

Tehnički urednik:

Hrvoje Stančić

Dizajn naslovnice:

Hrvoje Stančić

Lektura i korektura:

Brankica Bošnjak Terzić

Tisak:

Naklada:

300 primjeraka

CIP zapis dostupan u računalnom katalogu
Nacionalne i sveučilišne knjižnice u Zagrebu
pod brojem 816708.

ISBN 978-953-175-420-0

Computational Language Analysis

Computer-Assisted Translation (CAT)
and **e-Language Learning (eLL)**

edited by
Sanja Seljan

Zavod za informacijske studije
Zagreb, rujan 2012.

Contents

Introduction	1
1. COMPUTATIONAL LINGUISTIC MODELS AND LANGUAGE TECHNOLOGIES FOR CROATIAN.....	5
Information Technology in Machine Translation and in e-Language Learning of Croatian (<i>Sanja Seljan</i>)	7
Computational Linguistic Models and Language Technologies for Croatian (<i>Bojana Dalbelo Bašić, Zdravko Dovedan Han, Ida Raffaelli, Sanja Seljan, Marko Tadić</i>).....	17
Why Machine-Assisted Translation (MAT) Tools for Croatian? (<i>Sanja Seljan, Damir Pavuna</i>)	31
Digitisation and Language Technologies in the Learning Process of Information Sciences - Approaching the EU Standards (<i>Hrvoje Stančić, Sanja Seljan, Jadranka Lasić-Lazić</i>)	43
Integration of the project “Information Technology in Computer-Assisted Translation of Croatian and in e-Language Learning” into curriculum (<i>Sanja Seljan, Nataša Pavlović</i>).....	53
2. TRANSLATION MEMORIES	61
2.1. Sentence Alignment - Corpus, Tools, Methods.....	62
Sentence Alignment as the Basis for Translation Memory Database (<i>Sanja Seljan, Angelina Gašpar, Damir Pavuna</i>)	63
Corpus-Based Comparison of Contemporary Croatian, Serbian and Bosnian (<i>Božo Bekavac, Sanja Seljan, Ivana Simeon</i>).....	77
Towards Obtaining High Quality Sentence-Aligned English-Croatian Parallel Corpus (<i>Marija Brkić, Maja Matetić, Sanja Seljan</i>).....	87

Evaluating Sentence Alignment on Croatian-English Parallel Corpora (<i>Sanja Seljan, Željko Agić, Marko Tadić</i>).....	101
Corpus Aligner (CorAl) Evaluation on English-Croatian Parallel Corpora (<i>Sanja Seljan, Marko Tadić, Željko Agić, Jan Šnajder, Bojana Dalbelo Bašić, Vjekoslav Osmann</i>).....	113
2.2. Using Translation Memories.....	121
Translation Memory Database in the Translation Process (<i>Sanja Seljan, Damir Pavuna</i>)	123
Using Translation Memory to Speed up Translation Process (<i>Marija Brkić, Sanja Seljan, Božena Bašić Mikulić</i>)	135
3. TERMINOLOGY EXTRACTION	147
First Steps in Term and Collocation Extraction from English-Croatian Corpus (<i>Sanja Seljan, Angelina Gašpar</i>)	149
Comparative Analysis of Automatic Term and Collocation Extraction (<i>Sanja Seljan, Bojana Dalbelo Bašić, Jan Šnajder, Davor Delač, Matija Šamec-Gjurin, Dina Crnec</i>)	157
4. EVALUATION OF TRANSLATION TECHNOLOGY TOOLS & TRANSLATOR'S EDUCATION.....	167
Translation Technology as Challenge in Education and Business (<i>Sanja Seljan</i>)	169
Translator's Educational Perspective in Accession Country (<i>Vlasta Kučić, Sanja Seljan</i>).....	185
Evaluation of Electronic Translation Tools through Quality Parameters (<i>Vlasta Kučić, Sanja Seljan, Ksenija Klasnić</i>).....	199
5. MACHINE TRANSLATION EVALUATION.....	211
Evaluation of Free Online Machine Translations for Croatian-English and English-Croatian Language Pairs (<i>Sanja Seljan, Marija Brkić, Vlasta Kučić</i>).213	
Evaluation of the Statistical Machine Translation Service for Croatian-English (<i>Marija Brkić, Tomislav Vičić, Sanja Seljan</i>).....	229

Machine Translation Evaluation for Croatian-English and English-Croatian Language Pairs (<i>Marija Brkić, Sanja Seljan, Maja Matetić</i>)	243
BLEU Evaluation of Machine-Translated English-Croatian Legislation (<i>Sanja Seljan, Tomislav Vičić, Marija Brkić</i>)	255
6. COMPUTER-ASSISTED LANGUAGE LEARNING.....	267
6.1. e- Learning Platform.....	268
CALL (Computer-Assisted Language Learning) and Distance Learning (<i>Sanja Seljan, Mihaela Banek Zorica, Sonja Špiranec, Jadranka Lasić-Lazić</i>)	269
Quality Parameters for the e-Learning Omega System (<i>Ksenija Klasnić, Sanja Seljan, Hrvoje Stančić</i>)	281
Teaching English for Special Purposes Aided by e-Learning Platform (<i>Biserka Fučkan Držić, Sanja Seljan, Jelena Mihaljević Džigunović, Jadranka Lasić-Lazić, Hrvoje Stančić</i>)	295
6.2. Evaluation of e-Language Learning Activities.....	310
Evaluation of Classroom-based Online Multimedia Language Assessment (<i>Ana Cetinić, Sanja Seljan</i>)	311
Computer Learning of Small Math Using MATΣMATX in English Class (<i>Petra Mitrović, Sanja Seljan</i>)	319
Une expérience interculturelle en tandem par Internet (<i>Yvonne Vrhovac, Sanja Seljan, Martina Mencer Salluzzo, Bojan Prosenjak</i>)	329
L'enseignement hybride vers l'enseignement EAO en FLE dans le contexte croate (<i>Sanja Seljan, Yvonne Vrhovac, Martina Mencer Salluzzo</i>)	341
List of authors	353
List of tables, charts and figures	357

List of tables, charts and figures

Tables

Table 1: Number of translated pages in 1990s.....	33
Table 2: Size of parallel texts	66
Table 3: Automatic alignment	70
Table 4: Alignment: automatic vs. manual	71
Table 5: Alignment: automatic, manual.....	73
Table 6: Differences at phonological level.....	79
Table 7: Differences at morphological level.....	80
Table 8: Differences in name forms	80
Table 9: Differences at lexical level	81
Table 10: Differences in acronyms	82
Table 11: Differences at syntactic level	83
Table 12: Differences at various levels	84
Table 13: Corpus excerpts	88
Table 14: Corpus statistics	98
Table 15: Most frequent words	99
Table 16: Segment counts in sentence-aligned parallel corpus.....	99
Table 17: Legislative documents subcorpus statistics.....	103
Table 18: Cro-Eng subcorpus statistics.....	104
Table 19: Alignment and sentence track F1-measure on Croatian-English parallel corpora	109
Table 20: Document stats	116
Table 21: Alignment accuracy.....	118
Table 22: Time spent for each phase	144
Table 23: Corpus statistics	150
Table 24: Number of N-grams for two tools.....	151
Table 25: English POS-patterns of automatically extracted term candidates.....	153
Table 26: English POS-patterns for manually created reference list.....	154
Table 27: Number of n-grams in the reference list	162
Table 28: Local grammars.....	163
Table 29: Results of extraction evaluation.....	164
Table 30: Pages translated by DGT	193
Table 31: Total number of mistakes and paired samples statistics.....	204
Table 32: Paired samples t-test of statistically significant difference between average number of mistakes	205
Table 33: Number of lexical mistakes and paired samples statistics.....	205
Table 34: Paired samples t-test of statistically significant difference between average number of lexical mistakes	206
Table 35: Number of spelling and punctuation mistakes and paired samples statistics.....	206

Table 36: Paired samples t-test of statistically significant difference between average number of spelling and punctuation mistakes	206
Table 37: Number of syntactic and stylistic mistakes and paired samples statistics.....	207
Table 38: T-test of statistically significant differences.....	207
Table 39: Percentage of translation improvements	208
Table 40: Text statistics.....	216
Table 41: Average grades in Cro-En / En-Cro translations.....	219
Table 42: The level of agreement per domain and per system for Cro-En translation services	224
Table 43: The level of agreement per domain with regard to the criteria of fluency and accuracy for En-Cro translations by GT	225
Table 44: Fluency and adequacy scale.....	238
Table 45: Score frequencies according to fluency criteria	239
Table 46: Score frequencies according to adequacy criteria.....	239
Table 47: Association of two criteria with regard to different evaluators	240
Table 48: F-measure obtained for four different MT systems in four different domains.....	249
Table 49: Overall scores obtained for four different MT systems for the Croatian-English language pair	249
Table 50: BLEU scores obtained for four different MT systems in four different domains.....	249
Table 51: NIST scores obtained for four different MT systems in four different domains	250
Table 52: Correlation between automatic metrics and human assessments averaged over all systems for the Croatian-to-English translation task.....	251
Table 53: Correlation between automatic metrics and human assessments averaged over all systems for the English-to-Croatian translation task.....	252
Table 54: Test set statistics.....	258
Table 55: Average human fluency and adequacy criteria	260
Table 56: Interpretation of Fleiss' kappa values	260
Table 57: Fleiss kappa on human evaluation.....	261
Table 58: Error categories and error examples.....	261
Table 59: Error categories and number of errors per category.....	262
Table 60: BLEU scores with a single and multiple reference sets.....	262
Table 61: Variables for the research.....	302
Table 62: Number of students in both groups.....	303
Table 63: Average results	314
Table 64: Test scores overview.....	327
Tableau 65: Evaluation des assertions.....	349

Charts

Chart 1: The number of students owing computer	48
Chart 2: Students' Internet usage.....	48
Chart 3: The frequency of student's Internet usage.....	48
Chart 4: Access to Internet	48
Chart 5: The percentage of students having e-mail address.....	49
Chart 6: The percentage of students having personal web-page.....	49
Chart 7: Students' assessment of their knowledge of MS Windows OS	49
Chart 8: Students' assessment of their knowledge of MS Word	49
Chart 9: Effectiveness of TM.....	142
Chart 10: Time spent in translation process	144
Chart 11: Total time for traditional translation and TM	144
Chart 12: Eng-Cro N-gram term bases.....	152
Chart 13: Cro-Eng N-gram term bases.....	152
Chart 14: Previous experience in translation.....	175
Chart 15: Attending language technology courses.....	176
Chart 16: Types of text translated by free Internet translation tools.....	176
Chart 17: Use of free translation resources	177
Chart 18: Average grades for free online translation tools	177
Chart 19: Evaluation of Croatian language resources and tools	178
Chart 20: Preferred translation tool	178
Chart 21: Average grades for Google Translate.....	179
Chart 22: Average grades	179
Chart 23: Distribution of grades	180
Chart 24: Preferred way of packages of resource/tools	180
Chart 25: Further education for the translator's profession	181
Chart 26: Use of free translation resources	195
Chart 27: Average grades given to free online translation tools	196
Chart 28: Average grades for free language resources on the Internet	217
Chart 29: Average grades for online free Croatian language tools and resources.....	218
Chart 30: Average grades for four selected online translation services	218
Chart 31: Desirable resources of the appropriate quality.....	219
Chart 32: Average scores for 4 free online translation services for Cro-En	220
Chart 33: Average scores for GT in En-Cro and Cro-En directions	220
Chart 34: <i>Fluency</i> and <i>adequacy</i> average judgements.....	240
Chart 35: Correlation between fluency and adequacy criteria and error type.....	263
Chart 36: Human and BLEU scores on a 0-1 scale on short and long sentences separately, and with respect to the number of reference sets	263
Chart 37: The use of IT in educational purposes (average score on the scale from 1-never to 5-very often)	287
Chart 38: Frequency of the use of the computers.....	288
Chart 39: The degree of teachers' computer skills	289
Chart 40: Satisfaction with the quantity and the quality of the use of Omega.....	291
Chart 41: Achievements in single elements of oral evaluation.....	304

Chart 42: Achievement in oral evaluation referring to learning model.....	304
Chart 43: Different information sources (online vs. offline).....	305
Chart 44: Evaluation of communication quality between students and teacher (online vs. offline).....	306
Chart 45: Agreement with the statement “Moodle could replace all teaching models.”.....	307
Chart 46: Regions.....	314
Chart 47: Essay 1.....	315
Chart 48: Essay 2.....	315
Chart 49: Average success.....	316
Chart 50: Affective attitudes.....	317

Figures

Figure 1: CorAl tool.....	93
Figure 2: Bilingual Sentence Aligner tool.....	94
Figure 3: Model 1.....	95
Figure 4: Gargantuta tool.....	96
Figure 5: A plot of word frequency versus rank: (a) the English side of the parallel corpus (b) the Croatian side of the parallel corpus.....	97
Figure 6: Croatian-English parallel corpus sample.....	105
Figure 7: CORAL screenshot.....	106
Figure 8: Example reference and system alignment.....	107
Figure 9: Example alignments.....	116
Figure 10: Matches found by DVX when the second source text was inserted.....	141
Figure 11: Matches found by DVX after the third source text was inserted.....	141
Figure 12: Translator’s competences according to DGT.....	187
Figure 13: Ranking of four systems in the Croatian-to-English translation task (Google Translate, TranStar, Translation Guide and InterTran) according to three automatic metrics (BLEU, NIST and F-measure) and human assessments.....	251
Figure 14: System-level correlations between F-measure, BLEU, NIST and human assessments in the Croatian-to-English translation task. The size of a shape for a system depends on the majority-agreement-ranking of four systems, where the best system, i.e. TranStar, has the biggest shape.....	252
Figure 15: System-level correlation between automatic metrics and human accuracy and fluency assessments in the English-to-Croatian translation task.....	252
Figure 16: Google Translate correlation between two translation directions according to three automatic metrics and human assessments.....	253
Figure 17: Omega e-learning system.....	277
Figure 18: Examples of the recorded lecture.....	278
Figure 19: Example from the online glossary of geodetic terms.....	299
Figure 20: MATΣMATX interface – main screen.....	321
Figure 21: Help element – even numbers revision.....	323

Radovi Zavoda za informacijske studije

Knjige

1. Informacijske znanosti i znanje
2. M. Tuđman, Obavijest i znanje
3. A. Stipčević, Cenzura u knjižnicama
4. S. Jelušić, Struktura i organizacija knjižničnih sustava
5. Obrada jezika i prikaz znanja
6. I. Maroević, Uvod u muzeologiju
7. T. Aparac-Gazivoda, Teorijske osnove knjižnične znanosti
8. J. Lasić-Lazić, Znanje o znanju
9. B. Tepeš, Računarska lingvistika
10. Z. Dovedan, Formalni jezici: sintaksna analiza
11. Zbornik radova "Težakovi dani"
12. Modeli znanja i obrada prirodnog jezika
13. D. Kovačević, J. Lasić-Lazić, J. Lovrinčević, Školska knjižnica – korak dalje
14. Odabrana poglavlja iz organizacije znanja
15. Informacijske znanosti u procesu promjena
16. J. Lovrinčević, D. Kovačević, J. Lasić- Lazić, M. Banek Zorica, Znanjem do znanja
17. J. Lasić-Lazić, M. László, D. Boras, Informacijsko čitanje
18. S. Špiranec, M. Banek Zorica, Informacijska pismenost: Teorijski okvir i polazišta
19. H. Stančić, Digitalizacija
20. N. Mikelić Preradović, Učenjem do društva znanja
21. D. Kovačević, J. Lovrinčević, Školski knjižničar

Spomenice

1. Aleksandru Stipčeviću s poštovanjem
2. Ivi Maroeviću baštini u spomen



Sanja Seljan, Ph.D, Associate Professor is employed at the Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb. Her research interests are computer-assisted translation (CAT) and machine translation (MT), localization and translation memories (TM), natural language processing (NLP), Lexical-Functional Grammar (LFG) and Computer-Assisted Language Learning (CALL). In 1997 she did her master's thesis "Problems of syntactic analysis of Natural Language processing" and in 2003 her doctoral thesis "Lexical-Functional Grammar of the Croatian Language: theoretical and practical models". In 2004 she was elected as Assistant Professor and in 2008 as Associate Professor.

She has published about 50 research papers from the area of language and technology, participated in more than 60 conferences and workshops, gave 5 invited lectures and 6 workshops.

She has participated in three international and two domestic projects. Currently, she is the leader of the project "*Information Technology In Computer-Assisted Translation (CAT) of Croatian and in e-Language Learning (eLL)*" funded by the Croatian Ministry of Science, Education and Sport and she also participates in two more international projects. She is a member of different associations related to language technologies. She fluently speaks English and French and uses Italian.