University of Zadar

International conference

Corpora in Language Learning, Translation and Research

BOOK OF ABSTRACTS

University of Zadar

Zadar

August 28th - 29th, 2024

International conference

Corpora in Language learning, Translation and Research

Book of Abstracts

Publisher

University of Zadar

For the Publisher

Josip Faričić

Edited by

Larisa Grčić and Marija Brkić Bakarić

ISBN

978-953-331-509-6

PROGRAMME COMMITTEE

Maja Bratanić (Croatia), Elisa Corino (University of Turin, Italy), Paula de Santiago (Spain), Paula de Santiago (University of Valladolid, Spain), Cristina Dimulescu (Transilvania University of Brașov, Romania), Ivana Lalli Paćelat (University Jurja Dobrile of Pula), Lucia Načinović Prskalo (University of Rijeka), Andreea Nechifor (Transilvania University of Brașov, Romania), Špela Vintar (Slovenia)

ORGANIZING COMMITTEE

Larisa Grčić (University of Zadar, Croatia), Marija Brkić Bakarić (University of Rijeka, Croatia)

Program

August, 28th 2024

08:00-08:30	***Registration***
08:30-09:00	**Opening – welcome speech**
09:00-10:30	***Plenary lecture*** **Marek Łukasik** (Department of English Studies, Pomeranian Centre for Digital Humanities, Pomeranian University in Słupsk, Poland) *Corpus Linguistics in the Era of Artificial Intelligence*
10:30-12:00	**Session 1**
10:30-11:00	**Daniela Terenzi**, **Diego Brito** (Federal Institute of Education, Science and Technology of São Paulo - IFSP, Brazil), **Malila Carvalho de Almeida Prado** (BNU-HKBU United International College) *Study of clusters in aviation documents: corpus linguistics and international cooperation for ESP teaching* (ONLINE)

11:00-11:30	**Jasmina Jelčić Čolaković, Irena Bogunović** (Faculty of Maritime studies, University of Rijeka, Croatia), **Bojana Ćoso** (ISPP Cambodia) *A Corpus Study of Verb-Noun Dyads in Croatian: Towards a Psycholinguistic Database of Conceptual Metaphors in English and Croatian* (ONLINE)
11:30-12:00	**Sanja Marinov Vranješ** (Faculty of Economics, Business and Tourism, University of Split, Croatia) *The role of L2 proficiency in the outcomes and perceptions of data-driven learning of business and tourism students*
12:00-12:30	**Discussion**
12:30-14:30	**Lunch break**
	Session 2
14:30-15:00	**Artem Velychko** (University of Bialystok, Poland) *Building a Specialised Corpus on the Theme of Fishing* (ONLINE)
15:00-15:30	**Jana Kegalj** (Faculty of Maritime Studies, University of Rijeka, Croatia) *A corpus-based analysis of specialized spoken genre: a case study of maritime communications*
15:30-16:00	**Georgeta Bordea** (University La Rochelle, France), **Larisa Grčić,** (University of Zadar, Croatia) *Corpus-based expert profiling in the sustainability domain*
16:00-16:30	**Marija Brkić Bakarić** (Faculty of Informatics and Digital Technologies, University of Rijeka), **Ivana Lalli Paćelat** (Faculty of Humanities, Juraj Dobrila University of Pula) *Quick and Easy Term Extraction: Making Use of LLM-based Chatbots*
16:30-17:00	**Pedro Humánez-Berral** (University of Cantabria, Santander, Spain) *Potential vocabulary learning through children's movies for young EFL learners: A study using corpora and topic modeling*

17:00-17:30	**Discussion**

August 29th 2024

09:00-10:30	***Plenary lecture:*** **Elisa Corino** (Dipartimento di Lingue e Letterature straniere e Culture moderne, Università di Torino, Italy) *Data-driven learning: linguistic investigation from Sherlock Holmes to Chat GPT*
	Session 1
10:30-11:00	**Antonio Hermán-Carvajal** (University of Granada, Spain) *Questionnaires on Healthcare Workers' Mental Health: A Corpus-Based Study of Their Emotional Features for English-Spanish Translation* (ONLINE)
11:00-11:30	**Ivanka Rajh, Gorana Bikić-Carić** (Department for Romance Studies, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia) *Translators between the French passé simple and the Croatian aorist*
11:30-12:00	**Frane Malenica, Emilija Mustapić Malenica** (Department of English Studies)**, Jelena Gugić** (Faculty of Educational Sciences, University of Pula)**, Mojca Kompara Lukančić** (Faculty of Criminal Justice and Security, University of Maribor), **Jakov Proroković** (Department of Teachers and Preschool Teachers Education) *When Corpus-Driven Investigation Meets Experimental Design: The Role of Frequency Measures in Compound Processing*
12:00-12:30	**Discussion**
12:30-14:30	**Lunch break**

	Session 2
14:30-15:00	**Sandrine Peraldi** (University College Dublin, Ireland) *Examining online interactional commonalities and differences in the workplace: a multimodal corpus analysis of work meetings*
15:00-15:30	**Lucia Načinović Prskalo** (University of Rijeka, Faculty of Informatics and Digital Technologies) *Corpus Analysis in Addressing Gender Bias, Discrimination and Gender-Based Violence in Language: An Overview of the Possibilities*
15:30-16:00	**Discussion**
16:00-16:30	**Closing ceremony**

Marek Łukasik (Department of English Studies, Pomeranian Centre for Digital Humanities, Pomeranian University in Słupsk, Poland) <marek.lukasik@upsl.edu.pl>

Corpus Linguistics in the Era of Artificial Intelligence

Corpus linguistics (CL) has revolutionised numerous linguistic investigations, introducing a crucial element of objectivity to these studies and becoming the preferred method. Beyond the research realm, corpora have also been successfully applied in practical undertakings such as dictionary making, terminology extraction, language learning and teaching, and translation. The growing capacity of computers has made it possible to compile larger volumes of texts more quickly and efficiently, often in a fully-automated manner. Moreover, more developed CL tools offer more advanced features and user-friendly interfaces, becoming universal platforms with hundreds of preloaded corpora in dozens of languages.

With the introduction of widely available LLM tools, such as ChatGPT, linguists and other language professionals have started considering the possibility of employing these tools in their research and practical work. These tools have been tested, *inter alia*, in lexicography, terminology, language pedagogy, and translation, among others. Simultaneously, attempts have been made to combine the two approaches, with studies focussing on the application of GenAI tools in, for example, corpus approaches to discourse studies or the analysis of social media data. However, some studies yielded mixed results, highlighting some of the well-known limitations of GenAI technology.

From a research perspective, one of the major conundrums is repeatability and replicability of the studies, a feature that results from the non-deterministic nature of GenAI technology. Another issue is bias, which stems from the quality of the training. The lack of transparency of the processes (the black box phenomenon) and the inability to reference data sources are additional concerns.

In the talk, an attempt will be made to evaluate the application of CL and GenAI in specific linguistic studies and areas of practical application. This will be pursued through the literature review and the research results obtained by the author, mainly in the area of terminology classification, specialised translation and text excerption (for specific language data). Special emphasis will be put on low-resourced languages. The idea of integrating CL and AI will also be discussed and some examples of possible solutions will be provided. The benefits of the AI capability of analysing large datasets complementing the systematic nature of CL will also be highlighted.

References

Bonner, E. & Lege, R. & Frazier, E. (2023). Large Language Model-Based Artificial Intelligence in the Language Classroom: Practical Ideas For Teaching. *Teaching English with Technology*, 23(1), 23–41.

Curry, N. & Baker, P. & Brookes, G. (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. *Applied Corpus Linguistics*, 4, 1–9.

de Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, 36(4), 355–387.

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. Humanities and Social Sciences Communication (Nature), 10 (1).

Łukasik, M. (2023). Corpus linguistics and generative AI tools in term extraction: a case of Kashubian – a low-resource language. *Applied Linguistics Papers*, 27(4), 34–45.

Mátè, Á. et al. (2023). Machine Translation as an Underrated Ingredient? Solving Classification Tasks with Large Language Models for Comparative Research. *Computational Communication Research*, 5(2), 1–34.

Memon, A.B. et al. (2023). A corpus-based real-time text classification and tagging approach for social data. *Frontiers of Computer Science*, 6, 1–16.

Rees, G.P. & Lew, R. (2024). The Effectiveness of OpenAI GPT-Generated Definitions Versus Definitions from an English Learners' Dictionary in a Lexically Orientated Reading Task. *International Journal of Lexicography*, 37(1), 50–74.

Zappavigna, M. (2023). Hack your corpus analysis: How AI can assist corpus linguists deal with easy social media data. *Applied Corpus Linguistics*, 3, 1–5.

Elisa Corino (Dipartimento di Lingue e Letterature straniere e Culture moderne, Università di Torino, Italy) <elisa.corino@unito.it>

Data-driven learning: linguistic investigation from Sherlock Holmes to Chat GPT

More than 25 years ago, Tim Johns advocated the value of the data-driven learning-centered approach, calling "every student a Sherlock Holmes." Indeed, the best practices of data-driven learning (DDL) are perfectly aligned with current Second Language Acquisition (SLA) theories and practices, particularly constructivist and learner-centered approaches to language acquisition. It underpins the mandate of contemporary communicative language instruction for the use of authentic language materials and the development of metalinguistic knowledge and learner autonomy.

In a well-known article almost 10 years ago, McEnery and Xiao argue that "while indirect uses of corpora seem to be well established, direct uses of corpora in teaching are largely confined to advanced levels such as higher education" (2010: 374). They also add that "corpus-based learning activities are almost absent in classes teaching English as a foreign language (TEFL) at lower levels, such as secondary education" (McEnery/Xiao 2010: 374), and we might add that-still today-they are decidedly absent in the teaching practices of other foreign languages.

The advent of artificial intelligence (AI) has led to a disruption of the balance in the relationship among teachers, learners, and tools, providing ready to use answers and covering the procedures that are meant to be carried out by the “sherlock-student”.

This contribution will present action-research activities that, from data and research methodologies typical of computational linguistics, have demonstrated the effectiveness of this approach, which has forced attention to the form of portions of language whose specialized features often go unnoticed, especially within the subcodes of non-linguistic disciplines. We will then problematize the role of AI within the DDL framework and discuss the pros and cons that tools such as Chat GPT bring to the landscape of applied linguistics, corpus linguistics, and foreign language teaching.

References

Boulton, A. (2009). «Corpora for All? Learning Styles and Data-Driven Learning». Mahlberg, M.; González-Díaz, V.; Smith, C. (eds), Proceedings of 5th Cor-pus Linguistics Conference (University of Liverpool, July 20-23, 2009). Lan-caster: UCREL. http://ucrel.lancs.ac.uk/publications/cl2009/.

Boulton, A. (2012). «Beyond Concordancing: Multiple Affordances of Corpora in University Language Degrees». Procedia: Social and Behavioral Sciences, 34, 33-8. https://doi.org/10.1016/j.sbspro.2012.02.008.

Boulton, A. (2020). «Data-Driven Learning for Younger Learners: Obstacles and Optimism». Crosthwaite 2020, XIV-XX.

Boulton, A.; Cobb, T. (2017). «Corpus Use in Language Learning: A Meta-Analysis». Language Learning, 67(2), 348-93. https://doi.org/10.1111/lang.12224.

Boulton, A.; Leńko-Szymańska, A. (eds) (2015). Multiple Affordances of Language Corpora for Data-Driven Learning. Amsterdam; Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/scl.69.

Braun, S. (2007). «Integrating Corpus Work into Secondary Education: from Da-taDriven Learning to Needs-Driven Corpora». ReCALL, 19(3), 307-28. htt - ps://doi.org/10.1017/S0958344007000535.

Corino, E. (2020.). «Lingua e diritto: la Costituzione a scuola». Visconti, J. (a cura di), Linguaggi settoriali e specialistici: sincronia, diacronia, traduzione, variazione = Atti del XV Congresso della Società internazionale di Linguistica e Filologia italiana. Firenze: Cesati.

Corino, E.; Longo, F.; Onesti, C.; Pallante, F.; Sobrino, G. (i.p.). Leggere la Costituzione a scuola. Tra lingua e diritto. Roma: Carocci.

Corino, E.; Onesti, C. (2019). «Data-Driven Learning: A Scaffolding Methodology for CLIL and LSP Teaching and Learning». Frontiers in Education. htt - ps://doi.org/10.3389/feduc.2019.00007.

Crosthwaite, P. (ed.) (2020). Data-Driven Learning for the Next Generation. New York: Routledge.

De Mauro, T. (2008). Il linguaggio della Costituzione. Resoconto del convegno tenutosi a Palazzo Minerva il 16 giugno 2008, Senato della Repubblica. Convegni e seminari 18. https://www.senato.it/application/xmanager/projects/leg18/file/convegni_seminari_n18.pdf.

Doughty, C.; Williams, J. (eds) (1998). Focus on Form in Classroom Second Language Acquisition. Cambridge: Cambridge University Press.

Flowerdew, L. (2015). «Data-Driven Learning and Language Learning Theories: Whither the Twain Shall Meet». Boulton, Leńko-Szymańska 2015, 15-36.

Friginal, E. (2018). Corpus Linguistics for English Teachers. New York; London: Routledge.

Hafner, C.; Candlin, C. (2007). «Corpus Tools as an Affordance to Learning in Professional Legal Education». Journal of English for Academic Purposes, 6(4), 303-18. https://doi.org/10.1016/j.jeap.2007.09.005

Korzen, I. (2010). «Lingua, cognizione e due Costituzioni». Visconti, J. (a cura di), Lingua e diritto. Livelli di analisi. Milano: LED, 163-201.

La Fauci, N. (2011). «La Costituzione della Repubblica italiana. Spunti per una riflessione linguistica». Nesi, A.; Morgana, S.; Maraschio, N. (a cura di), Storia della lingua italiana e storia dell'Italia unita. L'italiano e lo stato nazionale = Atti del IX Convegno ASLI (Firenze, 2-4 dicembre 2010). Firenze: Franco Cesati Editore, 383-94.

Leńko-Szymańska, A. (2017). «Training Teachers in Data-Driven Learning: Tack-ling the Challenge». Language Learning & Technology, 21(3), 217-41.

O'Keeffe, A.; McCarthy, M.; Carter, R. (2007). From Corpus to Classroom: Language Use and Language Teaching. Cambridge (UK): Cambridge University Press.

Poole, R. (2018). A Guide to Using Corpora for English Language Learners. Edinburgh: Edinburgh University Press. Römer, U. (2006). «Pedagogical Applications of Corpora: Some Reflections on the Current Scope and a Wish List for Future Developments». ZAA, 54(2), 121-34.

Römer, U. (2009). «Corpus Research and Practice: What Help Do Teachers Need and What can We Offer?». Aijmer, K. (ed.), Corpora and Language Teaching. Amsterdam: John Benjamins, 83-98. Studies in Corpus Linguistics 33. https://doi.org/10.1075/scl.33.09rom

Scheffer-Lacroix, E. (2020). «Barriers to Trainee Teachers' Corpus Use». Crosthwaite 2020, 47-64.Snow, C.E. (2010). «Academic Language and the Challen

Daniela Terenzi, Diego Brito (Federal Institute of Education, Science and Technology of São Paulo - IFSP), **Malila Carvalho de Almeida Prado** (BNU-HKBU United International College) <daniela.terenzi@ifsp.edu.br>, <diego.campanelli@aluno.ifsp.edu.br>, < malilaprado@uic.edu.cnq>

Study of clusters in aviation documents: corpus linguistics and international cooperation for ESP teaching

English is the official language of aviation, as it is the most globalized sector in the industry. All professionals in the field not only need but are also obligated to use English for communication in both oral and written situations. Mechanics and aeronautical engineers, whose work is based on technical publications, often face challenges related to English given that such documentation is always written in English. Additionally, 80% of all aircraft technicians worldwide are not native English speakers. In this context, it is essential for professionals, especially those in training, to familiarize themselves with aviation specific vocabulary to correctly follow instructions and avoid misunderstandings, especially in countries where English is not the native language, such as Brazil. Despite initiatives aimed at simplifying the language of manuals, there is a shortage of materials that can be used by aviation professionals as a reference for technical terms and other linguistic aspects. In this regard, a Brazilian education institution has partnered with a Chinese university in order to develop a study aiming to conduct an analysis of clusters in aviation documents through a concordancer based on corpus linguistics principles (Berber Sardinha 2000, 2004; Aluísio &amp; Almeida 2006; Tagnin 2013, 2015). Some of the research results will be displayed in a visual dictionary that will be designed based on models by Selistre and Bugueño Miranda (2010) and Zacarias (2011). The outcome of the investigation has the potential to contribute not only to the linguistic aspects of the aviation field but also to be used in English for Specific Purposes (ESP) classes (Boulton 2016; Azadnia 2023) and serve as reference material for students and professionals in the field. Clusters are also called Lexical Bundles and can be defined as combinations of three or more words identified in a corpus (Wood 2015). According to Biber and Barbieri (2007), they play a key role in the comprehension and construction of language in specific contexts and, despite their importance, clusters are not common in second language materials (Wood &amp; Appel 2014). In addition, although Prado (2015) has made an analysis of lexical-grammar patterns of aviation English vectored by Corpus Linguistics and other studies have focused on lexical bundles in academic texts from other areas (Alasmary

2019), there is no investigations concerning English used in aviation written documents. The work in progress will be presented in the poster, so research results and the idea of the visual dictionary will be shown, fostering the opportunity to discuss the study, its results and ideas for future research. The study has been supported by FAPESP (Brazil).

References

Alasmary, A. (2019), “Academic lexical bundles in graduate-level math texts: a corpus-based expert-approved list”, Language Teaching Research, 26(1), p. 99-123.

Aluísio, S. M., &amp; Almeida, G. M. de B. (2006), “O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística”. Calidoscópio, 4(3), p.155-177.

Azadnia, M. (2023), “A Corpus-based Study of the Use of Lexical Bundles in EAP Texts by Iranian EFL and ESL Learners”, Teaching English as a Second Language Quarterly, 42(3), p.1-26.

Berber Sardinha, T. (2000), “Linguística de Corpus: histórico e problemática”, D.E.L.T.A., 16(2), p. 323-367.

Berber Sardinha, T. (2004), Linguística de Corpus, Barueri: Manole.

Biber, D., &amp; Barbieri, F. (2007), “Lexical bundles in university spoken and written registers”, English for Specific Purposes, 26(3), p. 263–286.

Boulton, A. (2016), “Integrating Corpus Tools and Techniques in ESP Courses.”, ASp, 69, p. 113-137.

PRADO, M. (2015), “Levantamento dos padrões léxico-gramaticais do inglês para aviação: um estudo vetorado pela Linguística de Corpus”, University of São Paulo, São Paulo, Brasil.

SELISTRE, I. C. T.; BUGUEÑO MIRANDA, F. V. (2010), “Os diferentes tipos de dicionários e as tarefas de compreensão e produção de textos em língua inglesa.”, Travessias, 8, p. 757-767.

TAGNIN, S. E. O. (2013), O jeito que a gente diz: combinações consagradas em inglês e português, Barueri: Disal.

Wood, D. (2015), Fundamentals of formulaic language: An introduction, New York: Bloomsbury.

Wood, D. C., &amp; Appel, R. (2014), “Multiword constructions in first-year university Engineering and Business textbooks and in EAP textbooks”, Journal of English for Academic Purposes, 15(9), p.1–13.

Zacarias, R. A. S. (2011), “Dicionário bilíngue pedagógico português-inglês: um novo parâmetro para a elaboração de informações gramaticais”, Universidade Estadual de Londrina, Londrina.

Jasmina Jelčić Čolaković (Faculty of Maritime studies, University of Rijeka, Croatia) <jasmina.jelcic@uniri.hr>; **Irena Bogunović** (Faculty of Maritime studies, University of Rijeka, Croatia) <irena.bogunovic@uniri.hr>; **Bojana Ćoso** (ISPP Cambodia) <coso.bojana@gmail.com>

A Corpus Study of Verb-Noun Dyads in Croatian: Towards a Psycholinguistic Database of Conceptual Metaphors in English and Croatian

Ever since the notion of conceptual metaphors (CMs) has been introduced by Lakoff and Johnson in 1980, there has been an attempt at exploring their pedagogical implications, especially in the field of vocabulary instruction (Boers 2000; Littlemore & Low 2006; Jelčić Čolakovac 2020). The need for raising metaphorical awareness and fostering metaphoric competence in language learners has been recognized by numerous researchers (Deignan, Gabryś & Solska 1997; Vasiljevic 2001; Cameron & Deignan 2003; Littlemore et al. 2011). It is now strongly believed that metaphoric competence should be fostered in the classroom setting without a disregard for the cultural component that can be found in the background of many figurative multiword expressions (MWEs) (Goranova 2023). Neurocognitive research into figurative language processing in unbalanced Croatian (L1)—English (L2) bilinguals presents a challenge for researchers not only because of the inaccessibility of necessary equipment, but also because of the unavailability of corpus data required in the stimuli design process. The present research aims to address this gap by providing frequency data on keywords to be used in an ERP experiment. The purpose of the experiment will be to provide data on the influence of metaphoric awareness on the processing of multiword expressions motivated by conceptual metaphors found in both L1 (Croatian) and L2 (English). For this purpose, a list of shared metaphors will be compiled (*Baza hrvatsko-engleskih metafora*, HREN) that will include MWEs, i.e. noun-verb dyads, motivated by metaphors whose presence has been detected in both languages (e.g. ARGUMENT IS WAR - VERBALNI SUKOB JE RAT). In compiling the list, the following procedure will be adhered to: 1) manual search of collocational dictionaries in Croatian and search of existing collocational and metaphor databases in Croatian (e.g. MetaNet.HR) for metaphorical verb-noun dyads existing in both L1 (e.g. *napraviti pogrešku*) and L2 (e.g. *make a mistake*); 2) corpus search of L1 verb-noun dyads using the Sketch Engine tool in order to collect data on phrase frequency as well as absolute and relative frequencies for the target words appearing in each dyad found in Step 1). Finally, the list of dyads will undergo a comprehensive normative study in which the MWEs will be checked for familiarity, predictability, metaphoricity, and word concreteness by analyzing existing databases (Ćoso et al. 2019; Peti-Stantić et al. 2021). The present talk will focus on the specific problems encountered during corpora search and will attempt to provide methods of alleviating them in future research. Firstly, an analysis will be provided on the discrepancies in verb-noun dyads use by comparing dictionaries to corpus data, with specific focus being placed on MWEs appearing in corpus use, but not in the published sources. Secondly, the list of dyads will be analyzed for underlying metaphors and compared to L2 data obtained on their translational equivalents. The final results will directly contribute to stimuli design for an ERP study in the Croatian bilingual context and will emphasize the importance of combining different methodologies with corpus searches in any neurolinguistic research.

References

Boers, F. "Metaphor awareness and vocabulary retention." *Applied Linguistics* 21.4 (2000): 553-571.

Cameron, L., and A. Deignan. "Combining large and small corpora to investigate tuning devices around metaphor in spoken discourse." *Metaphor and Symbol* 18.3 (2003): 149-160.

Ćoso, B., et al. "Affective and concreteness norms for 3,022 Croatian words." *Quarterly Journal of Experimental Psychology* 72.9 (2019): 2302-2312.

Deignan, A., D. Gabryś, A. Solska. "Teaching English metaphors using cross-linguistic awareness-raising activities." *ELT Journal* 51.4 (1997): 352-360.

Goranova, S. "Simile and metaphor in Medical English.". *Факултет по хуманитарни науки* 34.1 (2023): 210-219.

Jelčić Čolakovac, J. "Where Culture and Metaphor Meet: Metaphoric Awareness in Comprehension of Culturally-Specific Idioms." *Open Journal for Studies in Linguistics* 3.2 (2020).

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago–London: University of Chicago Press.

Littlemore, J., and G. Low. "Metaphoric competence, second language learning, and communicative language ability." *Applied Linguistics* 27.2 (2006): 268-294.

Littlemore, J., et al. "Difficulties in metaphor comprehension faced by international students whose first language is not English." *Applied Linguistics* 32.4 (2011): 408-429.

Peti-Stantić, A., et al. "The Croatian psycholinguistic database: Estimates for 6000 nouns, verbs, adjectives and adverbs." *Behavior Research Methods* 53.4 (2021): 1799-1816.

Vasiljević, Z. (2011). Using conceptual metaphors and L1 definitions in teaching idioms to non-native speakers. *The Journal of Asia TEFL,* 8(3), 135-160.

Sanja Marinov Vranješ (Faculty of Economics, Business and Tourism, University of Split, Croatia) <smarinov@efst.hr>

The role of L2 proficiency in the outcomes and perceptions of data-driven learning of business and tourism students

The aim of this study is to investigate the outcomes of a data-driven learning (DDL) intervention with a group of university non-language majors and their perceptions of this experience. The intervention involved the direct use of a small, specialised corpus (Koester, 2022) compiled by business and tourism students on the impact of the COVID pandemic on business and tourism flows, and a series of six targeted language information retrieval tasks prepared in advance by their language teacher. The study utilised a pretest/posttest research design in which students (N=112) were tested on six lexical and lexico-grammatical items before and after the intervention. A brief, 60-minute introduction to the basics of corpus consultation was given prior to the intervention, as corpus use has been shown to be successful in vocabulary learning even without prior training (Lee et al., 2019). Students' perceptions of the experience were operationalised using CORPATT, a scale to measure attitudes towards corpus use (Marinov, 2019). In addition to this widely used research design and the target structures tested (Boulton & Cobb, 2017), the study introduces L2 proficiency as an individual difference between respondents to broaden understanding of its relationship to DDL scores and perceptions. L2 proficiency is operationalised using a comprehensively validated C-test (Hessel, 2017). The data set will be analysed both quantitatively and qualitatively. Preliminary results have shown a statistically significant improvement in the posttest, with a large effect size ($z = 7.904$; $p = <.001$; $r = 0.75$). Attitude towards the experience was $M = 3.600$, $SD = 0.459$, which is an average score on a 5-point Likert scale. The two highest rated items relate to the cognitive aspect of the experience. The students believe that the approach enables them to recognise and practise lexis that they have previously only mastered receptively. As far as the emotional component of attitude is concerned, it is interesting to note that positive and negative emotions seem to go hand in hand. In other words, they are almost equally pleased with the number of examples presented and uncertain about handling the task.

References

Boulton, A., & Cobb, T. (2017). Corpus Use in Language Learning: A Meta-Analysis. *Language Learning, 67*(2)*, 348–393.* doi:10.1111/lang.12224

Hessel, G. (2017). A new take on individual differences in L2 proficiency gain during study abroad. *System, 66,* 39–55. doi:10.1016/j.system.2017.03.004

Koester, A. (2022). Building small specialised corpora. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics*. Routledge.

Lee, H., Warschauer, M., & Lee, J. H. (2019). The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis. *Applied Linguistics*, *40*(5), 721–753. doi.org/10.1093/applin/amy012

Marinov, S. (2018). CORPATT: a scale for measuring attitudes towards corpus use. *Strani jezici*, *47*(4), 221-242. Available at: https://hrcak.srce.hr/225527

Artem Velychko (University of Bialystok, Poland) <av76763@student.uwb.edu.pl>

Building a Specialised Corpus on the Theme of Fishing

The aim of the study is to compile a corpus representative of a fishing sociolect that could serve for, but is not limited to, the extraction of specialised topic-related collocations. The rationale behind the creation of this specialised corpus is attributed to several factors. First, the existing studies have predominantly analysed phraseology in general language and academic genres (e.g., Altenberg and Granger 2001, Durrant and Schmitt 2009, Paquot 2019, to name just a few). In relation to the latter, Greaves and Warren (2022: 207) point out that there is ‘a disproportionate amount of interest’ in academic language as opposed to other specialised domains. Another reason constitutes the paucity of reference materials comprehensively covering terminology and phraseology employed within the field of recreational activities compared to, for instance, business English (e.g., *Oxford Dictionary of Business English* 1993), legal English (e.g., Michta and Mroczyńska 2022), and the aforementioned academic English (e.g., Simpson-Vlach and Ellis 2010, Martinez and Schmitt 2012, Gardner and Davies 2014). The above issues along with the lack of available corpora of the fishing sociolect call for the creation of an authentic, representative, and large corpus suitable for research on angling phraseology and terminology.

The current study focuses primarily on the fishing corpus compilation, describing all the stages this process has involved. Prior to uploading corpus files to Sketch Engine (Kilgarriff et al. 2014), such steps had been taken as the collection of spoken and written authentic data, converting files into .txt format, removal of noise, i.e., unnecessary elements, delineating corpus sections and subsections according to fishing-related categories and topics, ensuring equal representation of data across all the sections and subsections, and naming text files in line with an adopted format.

The resultant corpus contains 4.4 million words (369 text files), consisting of written (74.9%) and spoken (25.1%) sections. The former incorporates 32 magazines and 28 books of angling manuals and narratives, while the latter is composed of 316 YouTube transcripts organised into five segments covering bass fishing, carp fishing, trout fishing, bait, and gear, with the last section being further subdivided into subsections about fishing hooks, reels and lines, rigs, as well as rods. Furthermore, due to the selection of multiple fishing categories and topics for inclusion, the corpus represents British and American varieties of English, which are distributed in different proportions depending on the material type.

The discussion concludes with several examples of lexical collocations on the theme of fishing to demonstrate which expressions were extracted from the given corpus by employing some quantitative and qualitative methods. Importantly, only those lexical items that formed part of the so-called practical components of the fishing discourse were selected. Unlike the theoretical component, which focuses on factual, historical, or legal issues of angling, the practical component encompasses aspects essential at each stage of a fishing session. Consequently, corpus analysis of these aspects has yielded a variety of terminological and phraseological items related to such topics as angling gear, e.g., *spinner blade*, angling techniques, e.g., *flip the bait*, angler skills, e.g., as *form/make/tie a knot*, gear properties, e.g., *rod bends/hoops (over)*, fish response, e.g., *fish gain line*, and many others. To conclude, corpus research whose goal is to capture terminology and/or phraseology characteristic of the fishing sociolect should focus on the practical component of this discourse, with a corpus under scrutiny containing representative samples of the outlined themes.

References

Altenberg, B., and Granger, S. 2001. The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics*, 22(2), 173–95.

Durrant, P., and Schmitt, N. 2009. To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching*, 47(2), 157–77.

Gardner, D., and Davies, M. 2014. ‘A new academic vocabulary list’. *Applied Linguistics*, 35(3), 305–327.

Greaves, C., and Warren, M. 2022. “What can a corpus tell us about multi-word units?”. In Anne O’Keeffe and Michael J. McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics* (2nd edition), 204–220. Abingdon, New York: Routledge.

Kilgarriff, A., Baisa, V., Bušta, J. et al. 2014. “The Sketch Engine: ten years on”. *Lexicography ASIALEX* 1, 7–36.

Martinez, R., and Schmitt, N. 2012. A phrasal expressions list. *Applied Linguistics*, 33(3), 299–320.

Michta, T., and Mroczyńska, K., 2022. *Towards a Dictionary of Legal English Collocations*. Siedlce: Siedlce University of Natural Sciences and Humanities.

Oxford Dictionary of Business English. 1993. Oxford: Oxford University Press.

Paquot, M. 2019. The phraseological dimension in interlanguage complexity research. *Second Language Research*, 35(1), 121–45.

Simpson-Vlach, R., and Ellis, N. 2010. An Academic Formulas List: New methods in phraseology research. *Applied Linguistics*, 31(4), 487–512.

Jasna Potočnik Topler (Faculty of Tourism, University of Maribor, Slovenia)
<jasna.potocnik1@um.si>

Travel Writing Texts as Resources in Teaching English for Tourism

Travel writing texts that simply put various accounts of journeys to or on destinations, present a valuable and with students quite popular resource in teaching English for tourism. The inclusion of travel writing texts into the ESP curriculum can enhance language skills and cultural knowledge that is essential for students and future professionals in the tourism industry. This article focuses on the potential of travel writing within a corpus-based pedagogical framework, emphasising its benefits for developing students' linguistic proficiency, intercultural competence, translation and digital skills. The tourism industry demands a high level of language proficiency and cultural awareness from its professionals. As English remains the global lingua franca, the ability to communicate effectively in this language remains significant. Travel writing, with its diverse vocabulary and varied use of language, serves as a suitable corpus for teaching English in a tourism context. By reading, writing and analysing travel texts, students can acquire tourism-specific vocabulary, learn to navigate different registers, and gain insights into cultural differences and nuances that are essential for communicating with international tourists and clientele from various cultural backgrounds. To implement travel writing in English for Tourism classes effectively, teachers should consider several factors. One essential is choosing a diverse range of travel texts to expose tourism students to different writing styles, destinations, and cultural perspectives. Utilising digital tools and resources can enhance the corpus-based methodology, enabling students to do text analysis and corpus searches independently. Integrating technology, such as language learning apps and online translation tools, can support students' learning and enhance their digital skills relevant to the tourism industry. Travel writing texts indeed represent a useful resource for teaching English for tourism, providing authentic material that enhances linguistic proficiency, intercultural competence, translation and digital skills. By using them, English teachers can create engaging, relevant, and effective learning experiences.

Furthermore, travel writing is a form of authentic material. It provides diverse examples of narrative, descriptive, and persuasive language. These texts can be compiled into specialised corpora, allowing English teachers to design activities that promote active learning. Through corpus analysis, students can identify common linguistic patterns, idiomatic expressions, and stylistic features, typical of tourism discourse and travel literature. This hands-on exploration of the English language helps students to internalize language structures, improve both comprehension and production skills, and enhance their intercultural competence. Understanding cultural contexts is essential in tourism, where professionals must cater to tourists from very diverse cultural backgrounds. Travel texts often offer the writers' reflections and observations of experiences and interactions with different cultures, enabling students to get an insight into the customs, habits and traditions of various destinations. Reading and analysing these texts encourages students to develop cultural sensitivity and adaptability, which are indispensable personality traits in the tourism sector.

Translation skills are another key component of language acquisition for tourism. The ability to translate travel-related content accurately and sensitively is vital for providing quality services. Travel writing texts serve as excellent study and practice material also for translation exercises. Students can engage in translating travel descriptions and articles, increasing their ability to convey meaning and tone across languages. This practice not only improves their translation competency but also deepens their understanding of the target and source languages' cultural and linguistic specifics.

Integrating travel writing into ESP curricula also aligns with modern pedagogical approaches such as for example student-oriented teaching and learning. By teaching with authentic texts, students are more likely to find the material engaging and relevant, which increases motivation and active participation. Tasks such as creating travel blogs, writing travel guides, and developing itineraries based on travel texts can make learning more interactive and enjoyable. These activities also encourage collaborative learning, as students can work in pairs or groups to discuss and analyse texts, share insights, and provide feedback to each other.

Jana Kegalj (Faculty of Maritime Studies, University of Rijeka, Croatia)
<jana.kegalj@pfri.uniri.hr>

A corpus-based analysis of specialized spoken genre: a case study of maritime communications

Maritime English, a variety of English in the domain of occupational English, is considered as the official language of communication in shipping. Bocanegra-Valle (2012) defined the subvarieties of Maritime English, with English for navigation and maritime communications being one subvariety mostly achieved through speech (communication in person, via VHF or broadcasting services). The aim of this paper is to analyse the current features of maritime communications based on a corpus of contemporary communication exchanges, MARCOM. The main advantage of the corpus is that it involves speakers in real, everyday situations producing unconstrained exchanges and thus providing us with an insight into current topics and concerns. Corpus data, such as noun and verb frequencies were used in the first stage to identify some recurring features or patterns, which were then further analysed qualitatively to explore the functional and pragmatic aspects of those patterns. The corpus data provided insight into which specific features to focus on while qualitative analysis was necessary for some pragmatic features of this genre. The analysis specifically focused on exploring the generic features of maritime spoken discourse, whose primary function is to facilitate clear, unambiguous, and professional communication in the maritime domain (cf. Bocanegra-Valle 2011, Pritchard & Kalogjera 2000; Dževerdanović-Pejović 2013; Perea-Barberá and Parada Galindo 2020). In that sense it shares some features of spoken discourse, such as simpler syntax, spontaneity and contextualization (cf. Schober and Brennan 2003, Cornish 2014), while at the same time having some features of specialized discourse, such as specialized and restricted vocabulary, formal register, standardized phrases, repetitions and confirmations (cf. Franceschi 2014). The speakers rely on their shared knowledge of the world to complete the conversation with minimum effort. This kind of in-depth study of the genre can provide valuable insights into the specific traits of the genre and may be of use in the future revisions of standardized phrases as well as rules of communication in maritime operations.

References

Bocanegra-Valle, A. (2011). The language of seafaring: Standardized conventions and discursive features in speech communications. *International Journal of English Studies*, 11(1), 35-53. https://doi.org/10.6018/ijes/2011/1/137091.

Bocanegra-Valle, A. (2012). Maritime English. U C. C.A, *The Encyclopedia of Applied Linguistics* (str. 3570-3583). Oxford: Wiley-Blackwell.

Cornish, F. (2014). Understanding spoken discourse. Elsevier Ltd. *Encyclopedia of Language and Linguistics* (2nd edition), vol 13, Elsevier Ltd., pp.227-230, 2006. hal-00952132

Dževerdanović-Pejović, M. (2013). Discourse analysis of the VHF communication at sea. *Maritime English Journal*, 1, 12-24. https://doi.org/10.7708/ijtte.2013.3(4).03.

Franceschi, D. (2014). The Features of Maritime English Discourse. *International Journal of English Linguistics* 4(2). 78-87. doi:10.5539/ijel.v4n2p78.

Perea-Barberá, M. D., & Parada Galindo, D. (2020). The use of standardized maritime English: Study of a small corpus of VHF communications in southern Spain. In D. Levey (Ed.), *Strategies and analyses of language and communication in multilingual and international contexts* (pp. 131-140). Cambridge Scholars.

Pritchard, B., & Kalogjera, D. (2000). On some features of conversation in maritime VHF communication. In M. Coulthard, J. Cotterill, & F. Rock (Eds.), *Dialogue analysis VII: Working with dialogue: Selected papers from the 7th IADA conference* (pp. 185-196). Niemeyer.

Schober, M. F. and Brennan, S. E. (2003). Processes of Interactive Spoken Discourse: the Role of the Partner. In A. Graesser, M. A. Gernsbacher, S. R. Goldman, *Handbook of Discourse Processes*. New Jersey: Lawrence Erlbaum Associates.

Georgeta Bordea (University La Rochelle, France) <georgeta.bordea@univ-lr.fr>;
Larisa Grčić (University of Zadar, Croatia) <lgrcic@unizd.hr>

Corpus-based expert profiling in the sustainability domain

In this paper we take a closer look at the task of automatically building topical profiles from a corpus of specialised texts in the sustainability domain. Our research is part of the EU-CONNEXUS project *Smart Urban Sustainable Coastal Development*, which aims to bring forward current social, economic, and environmental challenges such as coastal development, sustainable maritime tourism, blue economy and marine biotechnology by drawing knowledge from different disciplines. This kind of integration requires joint efforts from multiple stakeholders, like experts from neighbouring disciplines, users, and decision makers. In our previous research (Grčić, Rajh 2023) we approached sustainability as an interdisciplinary domain in which experts from several domains participate and exchange knowledge alongside a network of governmental organisations and EU institutions. Interdisciplinary collaboration is seen as a vital ingredient for finding innovative solutions to climate change (Krohn 2010), but crossing traditional disciplinary boundaries from natural, social and human sciences to engineering is a challenge on its own.

In our initial corpus-based study we aim to identify different fields of expertise which contribute to the exchange of discourses and the integration of core disciplinary knowledge. By recognizing significant authors and creating genres oriented subcorpora we aim to present representative sources for the sustainability domain. Bearing in mind the heterogeneity of the field we consider different communication settings and put emphasis on various textual types that embody the range of social intentions concerning the subject of sustainability.

Existing approaches for automatically building teams rely mainly on similarity and show that users are highly biased towards their own social connections, resulting in insular teams that lack diversity and concentrate expertise in a small number of teams (Gómez-Zará 2019). To some extent, more experienced researchers can rely on their extensive networks, but for early stage researchers automatic solutions for expert profiling and matching are essential. Matchmaking individuals based on complementarity of skills and expertise ensures teams are composed of members with different strengths and diverse perspectives. For that purpose topic modeling and topic labelling approaches will be used, along with term and taxonomy extraction to optimise the results. Expert profiling has garnered significant attention due to its capacity for systematically organising and evaluating the expertise, experience, and contributions of individuals. But current work on expert search relies on readily available keywords that describe competencies and knowledge areas to build topical profiles (Balog, De Rijke 2007). Instead, we will investigate approaches for extracting terms from existing publications (Bordea et al. 2013) and provide a qualitative comparison of results.

References

Balog, K., De Rijke, M. "Determining Expert Profiles (With an Application to Expert Finding)." IJCAI. Vol. 7. No. 625. (2007).

Bordea, G., Bogers, T., Buitelaar, P. "Benchmarking domain-specific expert search using workshop program committees." Proceedings of the 2013 workshop on Computational scientometrics: theory & applications. (2013).

Etienne, M., Du Toit, D. R., Pollard, S. "ARDI: a co-construction method for participatory modeling in natural resources management." *Ecology and society.* 16.1. (2011).

Gómez-Zará, D., Paras, M., Twyman, M. et al. "Who would you like to work with?." *Proceedings of the 2019 CHI conference on human factors in computing systems*. (2019).

Grčić, L., Rajh, I. "Towards Integration of Perspectives in the Terminology of the Interdisciplinary Domain 'Smart Urban Coastal Sustainability'". *Rasprave Instituta za hrvatski jezik i jezikoslovlje.* 49/2. 387-417. (2023).

Jensen, M. Sustainable Development Goals Interface Ontology. In: *ICBO/BioCreative*. (2016).

Krohn, W. Interdisciplinary cases and disciplinary knowledge. In: Frodeman, R. (ed.) *The Oxford Handbook of Interdisciplinarity*. Oxford University Press. 31-39. (2010).

Marija Brkić Bakarić (Faculty of Informatics and Digital Technologies, University of Rijeka) <mbrkic@uniri.hr>; **Ivana Lalli Paćelat** (Faculty of Humanities, Juraj Dobrila University of Pula) <ivana.lalli-pacelat@unipu.hr>

Quick and Easy Term Extraction: Making Use of LLM-based Chatbots

Automatic Term Extraction (ATE) based on a corpus is considered a translation technology with a high ROI (Zhao and Wang 2022). ATE is not only used to create terminology, but can also be used to improve complex downstream tasks.

Recent advances in the field of Natural Language Processing (NLP) have impacted the task of ATE in terms of both time and effort. Early ATE systems were mostly rule-based and relied on NLP tools such as tokenizers, lemmatizers, POS taggers, etc. Another set of tools used different statistical measures or combined statistical with linguistic information. Neural architectures were used for this task either by using embeddings as inputs to the classifiers or by fine-tuning the pre-trained language models. An overview of the approaches, tools, algorithms, and methods used over the last decade can be found in (Carlos et al. 2023). A comprehensive study on deep learning-based approaches for ATE with a focus on transformer-based neural models is presented in (Tran et al. 2023).

The aim of this study is to utilise the existing language resources described in (Brkic Bakaric and Lalli Pacelat 2019) to create a terminology at a relatively low cost. With the goal of observing terminological diversity, in our previous work we use the monolingual corpora and a statistical approach to extract the terms separately for each subcorpus (Paćelat, Bakarić, and Matticchio 2020).

The research question we aim to answer with this study is whether generative large language models like chatGPT can be used as a standalone extractor or only as a helper that needs the human in the loop. A similar study can be found in (de Paiva et al. 2023) with a focus on mathematical terms. Our work differs in that we perform multilingual term extraction on parallel corpora. According to the classification made in (Foo 2012), we use the “align-extract” method as the extraction process follows the alignment procedure. We divide the study into two parts. In the first part we randomly extract 100 sentences and ask a human annotator to extract the results to create the gold standard that can be used to calculate precision and recall.

In the second part of the study, we perform a full extraction based on almost 300k sentences and randomly select 100 terms to be scored by our human annotator. We obtain promising results. ChatGPT as a low-cost, web-based program which allows quick and easy term extraction will likely be used more intensively in the future.

References

Brkic Bakaric, Marija, and Ivana Lalli Pacelat. 2019. "Parallel Corpus of Croatian-Italian Administrative Texts." In *Proceedings of the Second Workshop Human-Informed Translation and Interpreting Technology Associated with RANLP 2019*, 11–18. Incoma Ltd., Shoumen, Bulgaria. https://doi.org/10.26615/issn.2683-0078.2019_002.

Carlos, Juan, Blandón Andrade, Carlos Mario, Medina Otálvaro, Zapata Jaramillo, Alejandro Morales Ríos, Carlos Mario Medina, and Otálvaro Msc. 2023. "Approaches, Tools, Algorithms, and Methods for Automatic Term Extraction: A Systematic Literature Mapping." https://doi.org/10.21203/rs.3.rs-2465373/v1.

Foo, Jody. (2012.) "Computational Terminology: Exploring Bilingual and Monolingual Term Extraction" (PhD diss., Linköping University Press).

Paćelat, Ivana Lalli, Marija Brkić Bakarić, and Isabella Matticchio. (2020.) "The Official Bilingualism in the Istrian County: State of the Art and Perspectives." *Rasprave Instituta Za Hrvatski Jezik i Jezikoslovlje*. Institute of Croatian Language and Linguistics. https://doi.org/10.31724/RIHJJ.46.2.20.

Paiva, Valeria de, Qiyue Gao, Pavel Kovalev, and Lawrence S. Moss. (2023.) "Extracting Mathematical Concepts with Large Language Models," August. http://arxiv.org/abs/2309.00642.

Tran, Hanh Thi Hong, Matej Martinc, Jaya Caporusso, Antoine Doucet, and Senja Pollak. (2023.) "The Recent Advances in Automatic Term Extraction: A Survey," January. http://arxiv.org/abs/2301.06767.

Zhao, Qingguo, and Jun Wang. (2022.) "Construction of Chinese-English Parallel Corpus of Chinese Laws and Regulations and Term Extraction Assisted by CAD Virtual Reality

Technology." *Computer-Aided Design and Applications* 20:46–55. https://doi.org/10.14733/CADAPS.2023.S1.46-55.

Pedro Humánez-Berral (University of Cantabria, Santander, Spain),
<pedro.humanez@unican.es>

Potential vocabulary learning through children's movies for young EFL learners: A study using corpora and topic modeling

Vocabulary is essential for language use (Nation & Waring, 1997). However, formal instruction does not provide learners with enough vocabulary to make progress in foreign language learning (Malone, 2018). In this regard, recent research has investigated the potential of audiovisual content consumption for L2 vocabulary learning (Peters & Webb, 2018). Extensive exposure to L2 media has been proven to have positive effects on different linguistic areas, such as vocabulary (Suárez & Gesa, 2019) or oral comprehension (Rodgers & Webb, 2017). Nevertheless, the vocabulary that learners could learn would depend on the topics they are exposed to. On this matter, it is particularly interesting to find out which topics and, subsequently, which vocabulary appear in these contents to better know the vocabulary that young EFL learners might learn. This has already been addressed in studies investigating the links between vocabulary learning and topics in EFL textbooks (Álvarez & Catalán, 2022). However, to the best of our knowledge, there are no similar studies focusing on movies that might be useful for foreign language vocabulary learning, especially among young EFL learners. In order to find out the specific vocabulary presented, corpora are excellent sources of data for research into this issue (Anthony, 2018) .

The objective of this study is thus to investigate the topics and vocabulary present in children's movies that could be used in EFL learning. To do so, we present an approach combining a corpus-based study and topic modeling, which is a statistical method used in natural language processing (NLP) and machine learning to identify topics in a corpus. This method does that by discovering the latent semantic structure in any given corpus (Churchill & Singh, 2022), relying on generative probabilistic models for collections of discrete data (Blei et al., 2003). Topic modeling has been used in studies framed within different fields of knowledge, such as social psychology (Rodríguez-Arauz et al., 2017). However, to the best of our knowledge, this approach and its potential implications in foreign language teaching and acquisition have not been fully explored yet.

To conduct our research, we compiled a corpus made of 11 closed captions (CC) scripts of children's movies in English, which featured ca. 100,000 words. We then used the Meaning Extraction Helper, a tool "that streamlines the extraction and analysis of information from text data" (Boyd, 2018) in a way that makes it possible to subsequently analyze that data with statistical analysis software, such as amovi (The jamovi project, 2024). We then performed a principal components extraction with varimax rotation to find out the factors of our corpus, using a scree of values for the principal components (Cattell, 1966). As done in previous studies, only words with factor loadings ≥ |0.20| were retained (Ramírez-Esparza et al., 2012; Rodríguez-Arauz et al., 2017). This allowed us to find the underlying topics in the movies analyzed, which will be presented and discussed, focusing on its pedagogical implications for children's movies consumption in EFL contexts.

References

Anthony, L. (2018). Introducing English for Specific Purposes. Routledge.

Álvarez, M. I. G., & Catalán, R. M. J. (2022). Words and topics in ELT textbooks for young EFL learners. *Language Teaching for Young Learners*, *4*(2), 264–288.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research* , *3*, 993–1022.

Boyd, R. L. (2018). *MEH: Meaning Extraction Helper (Version 2.3.02)*.

Cattell, R. B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research*, *1*(2), 245–276.

Churchill, R., & Singh, L. (2022). The Evolution of Topic Modeling. *ACM Computing Surveys*, *54*(10s), 1–35.

Malone, J. (2018). Incidental vocabulary learning in SLA: Effects of frequency, aural enhancement, and working memory. *Studies in Second Language Acquisition*, *40*(3), 651-675.

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. *Vocabulary: Description, acquisition and pedagogy*, *14*(1), 6-19.

Peters, E., & Webb, S. (2018). Incidental vocabulary acquisition through viewing L2 television and factors that affect learning. *Studies in second language acquisition*, *40*(3), 551-577.

Ramírez-Esparza, N., Chung, C. K., Sierra-Otero, G., & Pennebaker, J. W. (2012). Cross-Cultural Constructions of Self-Schemas. *Journal of Cross-Cultural Psychology*, *43*(2), 233–250.

Rodgers, M. P. H., & Webb, S. (2017). The effects of captions on EFL learners' comprehension of English-language television programs. *CALICO Journal, 34*(1), 20-38.

Rodríguez-Arauz, G., Ramírez-Esparza, N., Pérez-Brena, N., & Boyd, R. L. (2017). Hablo Inglés y Español: Cultural Self-Schemas as a Function of Language. *Frontiers in Psychology*, *8*.

Suárez, M. M., & Gesa, F. (2019). Learning vocabulary with the support of sustained exposure to captioned video: do proficiency and aptitude make a difference? *The Language Learning Journal*, *47*(4), 497-517.

The jamovi project. (2024). *jamovi* (2.5).

Antonio Hermán-Carvajal (University of Granada, Spain) <aherman@ugr.es>

Questionnaires on Healthcare Workers' Mental Health: A Corpus-Based Study of Their Emotional Features for English-Spanish Translation

Questionnaires are one of the main tools for assessing the mental health of healthcare workers. However, most of them are designed in English. Their use in other languages seems to rely on translation and cultural adaptation processes (Escobar Bravo, 2004). Most of the literature on the translation and adaptation of instruments used in health care research focuses on validating the already adapted instrument (Sousa & Rojjanasrirat, 2011), without focusing much on psychometric equivalence between the source and target cultures. The objective of this research is thus to explore the potential of corpus linguistics for translating healthcare questionnaires, paying special attention to the emotional elements of said questionnaires. This might potentially allow cultural nuances to be better conveyed in adapted questionnaires, ideally improving their effectiveness in the target language (Congost-Maestre, 2010; Reis, 2009). More specifically, the aim of this presentation is to characterize the emotions of the texts included in a bilingual corpus of 90 questionnaires (63 in English and 27 in Spanish) used to assess healthcare workers' mental health. Although the main focus of our research is to analyse the questionnaires designed to measure specific emotional problems such as burnout or moral distress, we also included other questionnaires which are usually used to assess mental health of healthcare workers as a whole. To characterize the emotional elements in these questionnaires, we used LIWC-2022, a program that "consists of software and a 'dictionary' — that is, a map that connects important psychosocial constructs and theories with words, phrases, and other linguistic constructions" (Boyd et al., 2022, p. 2). This software has been extensively used within the field of psychological sciences (Ramírez-Esparza et al., 2012; Rodríguez-Arauz et al., 2017, among others) and it has proven to be valid and reliable for language analysis (Boyd et al., 2022). We also used Lingmotif (Moreno-Ortiz, 2017), a sentiment analysis tool, to gain further understanding of the emotional features of the corpus. Both LIWC-2022 and Lingmotif analyze texts in English and Spanish, which allows us to conduct a comparable analysis of the questionnaires in both languages. Our results show a clear predominance of negative emotions in these texts (in terms of words and emotional tone), although the specific emotion lemmas vary among the questionnaires analyzed. These results will be discussed addressing their possible implications for English-Spanish translations of healthcare questionnaires.

References

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. University of Texas at Austin.

Congost-Maestre, N. (2010). El lenguaje de las Ciencias de la Salud: los cuestionarios de salud y calidad de vida y su traducción del inglés al español [PhD Dissertation]. Universidad de Alicante.

Escobar Bravo, M. Á. (2004). Adaptación transcultural de instrumentos de medida relacionados con la salud. Enfermería Clínica, 14(2), 102–106. https://doi.org/10.1016/S1130-8621(04)73863-2

Moreno-Ortiz, A. (2017). Lingmotif: A User-focused Sentiment Analysis Tool. Procesamiento Del Lenguaje Natural, 58, 133–140.

Ramírez-Esparza, N., Chung, C. K., Sierra-Otero, G., & Pennebaker, J. W. (2012). CrossCultural Constructions of Self-Schemas. Journal of Cross-Cultural Psychology, 43(2), 233–250. https://doi.org/10.1177/0022022110385231

Reis, L. O. (2009). Traducción de los cuestionarios para su uso en investigación multicultural - estamos haciendo lo correcto? Actas Urológicas Españolas, 33(1), 5–7.

Rodríguez-Arauz, G., Ramírez-Esparza, N., Pérez-Brena, N., & Boyd, R. L. (2017). Hablo Inglés y Español: Cultural Self-Schemas as a Function of Language. Frontiers in Psychology, 8. https://doi.org/10.3389/fpsyg.2017.00885

Sousa, V. D., & Rojjanasrirat, W. (2011). Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline. Journal of Evaluation in Clinical Practice, 17(2), 268–274. https://doi.org/10.1111/j.1365-2753.2010.01434.x

Ivanka Rajh, Gorana Bikić-Carić (Zagreb University, Department for Romance Studies), <irajh@m.ffzg.hr>; <gbcaric@ffzg.unizg.hr>

Translators between the French passé simple and the Croatian aorist

Recent interviews with three French translators about the *passé simple* in their translations reveal that the use of this past tense, describing an action that was completed in the past with no relation to the present, depends on the language combination involved in the translation, but also on the personal assessment of the translator based on the context and intended stylistic effect. Thus, when translating from Russian, which has only one past tense and two aspects, perfective and imperfective, the translator needs to "feel" whether to translate the perfective past by *passé simple*, *passé composé* or *plus-que-parfait*. The translator will opt for the *passé simple* when she wants the text to evoke the flair of ancient times, but for the *passé composé* when the "spoken" aspect needs to be emphasized. Even in contemporary texts, the *passé simple* is used for narration, while the *passé composé* is used for direct speech. When translating from German into French, a single tense, the *preterite,* corresponds to two French tenses, the *passé simple* and the *imparfait,* which can be a challenge for translators. The interviewed French translator insists on using the *passé simple* when possible and considers the variety of past tenses to be a unique asset of the French language, which enables her to emphasize differences in meaning. Unlike in contemporary French, the *passé simple* is frequently used in contemporary Italian. Therefore, when translating Italian texts into French, translators need to be careful and avoid systematically translating all instances of the Italian *passé simple* into the French *passé simple*. The French translator considers that the *passé simple* shouldn't be used in translation of works of autofiction, where this tense sounds anachronistic in French.

In Croatian, the past tenses *aorist* and *perfekt* would be counterparts of the *passé simple* and the *passé composé* in French. However, the *aorist* is rarely used in contemporary spoken language. It persists in literature, in narrative types of texts, as well as in conversational style, where its use was boosted by the new communication platforms such as SMS. Even rarer is the *imperfekt*, which is regularly replaced by the *perfekt,* and is only used in literary style. The aim of this paper is to see how the Croatian *aorist* and *imperfekt* translate into French, and vice versa, how the French *passé simple* and *imparfait* translate into Croatian. The research is going to be conducted on the corpus of translated literary texts making up the RomCro parallel corpus. While the above testimonials may suggest some general tendencies when translating into French, we expect that the corpus data shall reveal particularities of the French-Croatian language pair, which may have implications for translation pedagogy.

References

Bikić-Carić, G., Mikelenić, B. & Bezlaj, M. (2023). Construcción del RomCro, un corpus paralelo multilingüe. Procesamiento del Lenguaje Natural, 70. Sociedad Española para el Procesamiento del Lenguaje Natural, 99-110.

Benech, S. (2023, September 15). *Où en sommes-nous avec le passé simple ? Réponses de traductrices*. Contreligne. https://www.contreligne.eu/2023/07/sophie-benech-francoise-retif-ou-en-sommes-nous-avec-le-passe-simple-reponses-de-traductrices/

Grevisse, M., & Goosse, A. (1995). *Nouvelle grammaire française* (3. éd). De Boeck.

Rétif, F. (2023, September 15). *Où en sommes-nous avec le passé simple ? Réponses de traductrices*. Contreligne. https://www.contreligne.eu/2023/07/sophie-benech-francoise-retif-ou-en-sommes-nous-avec-le-passe-simple-reponses-de-traductrices/

Pisetta, J.P. (2023, September 15). *Où en sommes-nous avec le passé simple ? Réponses de traductrices*. Contreligne. https://www.contreligne.eu/2023/08/ou-en-sommes-nous-avec-le-passe-simple-suite-le-cas-italien/

Silić, J. & Pranjković, I. (2005). Gramatika hrvatskoga jezika : za gimnazije i visoka učilišta. Zagreb: Školska knjiga

Žic Fuchs, M. & Tuđman Vuković, N. (2008). Communication technologies and their influence on language: Reshuffling tenses in Croatian SMS text messaging. Jezikoslovlje, 9 (1-2), 109-122. Retrieved from https://hrcak.srce.hr/30681

Frane Malenica, Emilija Mustapić Malenica (Department of English Studies, University of Zadar) <fmalenica@unizd.hr>, <emustapic@unizd.hr>, **Jelena Gugić** (Faculty of Educational Sciences, University of Pula) <jelena.gugic@unipu.hr>, **Mojca Kompara Lukančić** (Faculty of Criminal Justice and Security, University of Maribor) <mojca.kompara@um.si>, **Jakov Proroković** (Department of Teachers and Preschool Teachers Education) <jprorokov@unizd.hr>

When Corpus-Driven Investigation Meets Experimental Design: The Role of Frequency Measures in Compound Processing

Compounds have garnered significant scholarly focus in language processing research, with an emerging number of studies done on non-native speakers in recent years (for example, see Pham, 2022; Du et al. 2023; Hamada, 2024; Puimege et al., 2024 etc.). Comparing compound processing across languages is particularly interesting once the significant typological differences existing between them are taken into account (Scalise & Bisetto, 2009; Semenza & Luzatti, 2014), and these only seem to intensify when comparing native and non-native speakers' processing. This study aims to present the initial findings derived from the identification and extraction of specific compounds in the preliminary phase of the research, serving as a foundation for subsequent experimental investigations that delve into the effects of morphological relatedness, frequency and schematicity of morphological family/pattern and L2 proficiency on native and non-native processing of nominal compounds in Croatian and Slovene as L1 and English as L2. We intend to reach this aim by answering the following research questions:

1. Does morphological decomposition occur with nominal compounds in Croatian and Slovene as L1 and English as L2?
2. Do frequency measures (frequency of individual constituents, schematicity) affect processing in L1 and L2 and to what extent?
3. Is the effect of L2 proficiency modulated by frequency measures, i.e. are speakers of different L2 proficiency affected by frequency measures in a different manner?

The initial phase of the research involved the identification and extraction of compounds essential for the experiment, with around 50 000 word forms which had to be filtered after having been identified in the corpora based on the morphologically-conditioned extraction criteria. The selection criteria for the corpora encompassed factors such as the target language, corpus size, relevance, recency, and the nature of texts included. Finally, approximately 2000 Croatian and Slovenian compound examples were retained after having been extracted and filtered out from the CLASSLA web corpora (CLASSLA-web.hr and CLASSLA-web.si), while English compound examples were sourced from the ukWaC English corpus (only 10 000 most frequent ones were retained). Based on the corpus data, we determined the frequency of individual compound constituents and the size of pattern/series employed in the experiment, while the data on L2 proficiency was collected via adapted proficiency test. The presented findings connect said corpus research and data obtained from native speakers of English and Slovene with moderate to high proficiency in English as a second language (L2) via experiments that employed masked priming lexical decision tasks (Forster & Davis, 1984). The first experiment investigated the effects of morphological relatedness in nominal compounds in Croatian/Slovene as L1 and in English as L2, while the second experiment delved into the range of existing patterns (or schematicity) and the impact this potentially has on compound processing, i.e., specifically whether compounds with right constituents that combine with a greater number of left constituents tend to be processed more quickly in both L1 and L2.

References

Hamada, M. (2024). Multiword Expressions: Understanding the Meanings of Noun-Noun Compounds in L2. *English Teaching & Learning,* 1-19.

Du, L., Elgort, I., & Siyanova-Chanturia, A. (2023). Cross-language influences in the processing of L2 multi-word expressions. *Cross-Language Influences in Bilingual Processing and Second Language Acquisition*, *16,* 187.

Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(4), 680–698.

Pham, H. (2022). English Compound Meaning Predictability: An Exploratory Cross-Linguistic Study. *3L: Southeast Asian Journal of English Language Studies*, *28*(3).

Puimege, E., Montero Perez, M., & Peters, E. (2024). The effects of typographic enhancement on L2 collocation processing and learning from reading: An eye-tracking study. *Applied Linguistics*, *45*(1), 88-110.

Scalise, S., & Bisetto, A. (2009). The classification of compounds. In R. Lieber & P. Štekauer (Eds.), *The Oxford Handbook of Compounding* (pp. 34–53). Oxford University Press.

Semenza, C., & Luzzatti, C. (2014). Combining words in the brain: The processing of compound words. Introduction to the special issue. *Cognitive Neuropsychology, 31(1–2),* 1–7.

Sandrine Peraldi (University College Dublin, Ireland) <sandrine.peraldi@ucd.ie>

Examining online interactional commonalities and differences in the workplace: a multimodal corpus analysis of work meetings

As Bennett (1997: 16) states, ‘to avoid becoming a fluent fool, we need to understand more completely the cultural dimension of language’. This cultural dimension refers, of course, to the obvious and less obvious differences arising between various national cultures or subcultures in terms of worldviews, communication preferences/practices and behaviours, which progressively led to the introduction of intercultural modules or components both in second language teaching and translation programmes. However, this cultural dimension can (and should) also encompass and examine more local discourse communities (Swales 1990, 2016) and, in particular, that of specific professional communities. Indeed, in an academic context where both universities and students seek to enhance employability skills and rates, it is vital to train future graduates to fully understand and adapt to the communication practises and subtleties that will characterise their profession. This is all the more important that the recent Covid 19 pandemic has brought a whole new dimension to professional communication. Online communication via video platforms such as Zoom and Microsoft Teams have now become a ubiquitous component of workplace interaction, leading to a widespread adoption of virtual and hybrid meetings as a primary means of communicating with colleagues.

This paper describes research from the ‘Interactional Variation Online’ (IVO) project, a 3- year UK/Irish funded project, aimed at examining virtual workplace communication through an innovative corpus-based multi-modal analysis. This project, which focuses particularly on the examination of gestures, gaze and non-verbal behaviours as well preferred communication strategies, was designed to gain depth of insight into the potential barriers and carriers in effective dyadic and multiparty online talk.

A multimodal corpus of virtual workplace meetings emanating from 5 different private and public companies/institutions has been built and analysed in order to demonstrate how virtual meetings differ from face-to-face meetings, but also differ from one discourse community to another and how this analysis can inform improved virtual workplace interaction. This paper will first present the construction of a multi-modal corpus of virtual workplace meetings with examples of how this type of data can be captured and analysed. Indeed, a key objective of this project was to develop appropriate technical protocols for capturing and investigating multimodal interaction. This includes the development of next generation analytical frameworks for analysing multimodal discourse, creating a standardised set of protocols for use by researchers and end-user collaborative communities. Examples of recordings of virtual workplace meetings will be shown to exemplify the challenges faced when attempting to create a multi-modal corpus using this type of data. The stages of corpus construction will be demonstrated from orthographic transcription to gestural annotation, presenting the affordances of various tools which aid this process. The paper will then delve into some of the findings of the analysis, with a specific focus on the interactional commonalities and differences found from one discourse community to another. Do different companies adapt similarly or differently when they move from a face-to-face to an online environment? Do they reproduce the typical register and idiolect that usually characterise their specific profession? Are meetings structured and chaired in similar manners from one institution to another? And to which extent each individual participant compensates both verbally and nonverbally for interacting in what can be considered as a cue-filtered environment (focus on the upper body of the participant, multiplicity of participants, technical problems, to name a few difficulties, that impact our ability to interpret correctly people's reactions and emotions).

In summary, this paper is aimed at demonstrating the need for a better understanding of what has become, and is likely to remain, a normalised way of communicating in the workplace, whilst highlighting avenues for classroom/workplace applications and further investigation.

References

Bennett, M. J. (1997). How not to be a fluent fool: Understanding the cultural dimensions of language. In A. E. Fantini, (Vol. Ed.) & J. C. Richards (Series Ed.). (1997). New ways in teaching culture. New ways in TESOL series II: Innovative classroom techniques (pp. 16– 21). Alexandria, VA: TESOL.

Swales, J. M. (1990). Genre Analysis: English in Academic and Research Settings. Cambridge [England] ; New York: Cambridge University Press.

Swales, J. M. (2016). 'Reflections on the concept of discourse community ', ASp [Online], 69 | 2016, Online since 01 March 2017. URL : http://journals.openedition.org/ asp/4774 ; DOI : https://doi.org/10.4000/asp.4774

Lucia Načinović Prskalo (University of Rijeka, Faculty of Informatics and Digital Technologies) <lnacinovic@uniri.hr>

Corpus Analysis in Addressing Gender Bias, Discrimination and Gender-Based Violence in Language: An Overview of the Possibilities

Language is a fundamental cultural component that shapes social norms and values and influences the way individuals perceive and interact with their society. Gendered language reflects and reinforces traditional gender roles, which contribute significantly to gender inequality. Analyzing these inequalities through corpus analysis gives insight into how to develop automatic methods for promoting equality in communication. This paper considers the relationship between language and gender equality, focusing on modern Natural Language Processing (NLP) techniques for detecting and mitigating gender bias.

Gender equality is more than just equal rights, opportunities, and access to resources irrespective of one's gender; it cuts across various sectors of life such as education, employment, and politics among others. A fairer and more inclusive society can be achieved by eliminating barriers that only favor certain genders at the expense of others. In this respect, language becomes important because it reflects and perpetuates social norms and biases (Lakoff, 1975; Talbot, 1998).

In many languages, including Croatian, gender-specific word forms often convey traditional gender roles. For instance, in professions like "doctor" or "driver", masculine forms are usually employed while feminine forms are associated with caring professions such as "nurse" (Mills, 2008; Mihaljević, 2021). Such language dichotomy sustains stereotypes and at the same time can impact societal understanding of gender roles. Analyzing language from a gender equality perspective involves recognizing and addressing these biases to promote more neutral and inclusive expressions (Ehrlich et al., 2014).

An important aspect of research in this area is the identification of gender stereotypes in language. In Croatian for instance, 'weak' often connotes femininity while 'strong' indicates masculinity. These connections support negative myths which can be identified and combated through corpus analysis (Widodo & Elyas, 2020). By examining the frequency and context of these terms, researchers can suggest neutral alternatives that do not reinforce traditional gender roles (Lindvall-Östling, et al., 2020).

Additionally, it is possible to consider the effect of language on career choices as well as educational opportunities. Linguistic stereotypes about suitable occupations for different genders can influence career paths and access to education. Language barriers can be determined by reviewing job adverts and learning materials. This aspect of research highlights the importance of language in shaping career aspirations and educational outcomes (Hu et al., 2022).

Furthermore, within the language, investigations can also be expanded to cover violence against genders. Derogatory terms directed at women are prevalent in various texts and contribute to a culture of tolerance of violence against women. Identifying the forms of linguistic violence is crucial in developing strategies against gender bias and discrimination. Textual analysis of emotional and verbal abuse helps to reveal subtle forms of violence that are often overlooked and highlight the need for comprehensive language reform (Sunderland, 2006).

Another aspect is the comparative approach between different languages. This method aims at determining how a gender-sensitive language can be translated or adapted properly across different linguistic contexts (Prewitt-Freilino et al., 2012). For example, translating gender-neutral terms from English into Croatian, where gender-specific forms are more prevalent, presents a particular challenge (Buikema et al., 2017).

In machine translation, translating sentences from gender-neutral languages into gender-specific languages often leads to assumptions based on stereotypes. As an example, “Taylor is a great doctor.” translated from English to Croatian would require a specification for Taylor’s gender thus sometimes reinforcing stereotypes. Addressing these issues in machine translation is crucial for the development of more accurate and inclusive translation systems (Savoldi et al., 2021).

Modern NLP tools and Artificial Intelligence (AI) enable the efficient processing of large text corpora and the identification of patterns of gender bias and stereotypes that would be difficult to detect manually (Sun et al., 2019). Additionally, they can also detect the frequency or type of words used in a document that includes emotions hence proposing alternative phrases that will help in promoting gender equality. The creation of algorithms capable of recognizing such gender bias-related issues has wide-ranging applications, from the advertising/media industry, and education up to human resources (Zhao et al., 2018).

To summarize, language plays a crucial role in gender equality. NLP tools together with AI using corpus analysis can identify and reduce the gender bias observed in language leading to more inclusive communication. These efforts can provide actionable insights for policy-making, education, and technology development, aimed at promoting a fairer and more inclusive society.

References

Buikema, R., Plate, L., & Thiele, K. (2017). *Doing Gender in Media, Art and Culture*. Routledge.

Ehrlich, S., et al., eds. (2014). *The Handbook of Language, Gender, and Sexuality*. Wiley-Blackwell.

Hu, S., et al. (2022). Balancing Gender Bias in Job Advertisements with Text-Level Bias Mitigation. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1, 326-340.

Lakoff, R. (1975). *Language and Woman's Place*. Harper & Row.

Lindvall-Östling, M., Deutschmann, M., & Steinvall, A. (2020). An Exploratory Study on Linguistic Gender Stereotypes and their Effects on Perception. *Journal of Language and Discrimination*, 3(1), 51-76.

Mills, S. (2008). *Language and Sexism*. Cambridge University Press.

Mihaljević, M. (2021). Male and Female in Croatian Language and Lexicography – Stereotypes and Language Discrimination. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, 47(2), 655-68.

Prewitt-Freilino, J. L., Caswell, T. A., & Laakso, E. K. (2012). The Gendering of Language: A Comparison of Gender Equality in Countries with Gendered, Natural Gender, and Genderless Languages. *Sex Roles: A Journal of Research*, 66(3-4), 268-281.

Savoldi, B., et al. (2021). Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, 8, 854-870.

Sun, T., et al. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1630-1640.

Sunderland, J. (2006). *Language and Gender: An Advanced Resource Book*. Routledge.

Talbot, M. (1998). *Gender and Language: An Introduction*. Polity Press.

Widodo, H. P., & Elyas, T. (2020). Introduction to Gender in Language Education. *Gender and Language*, 14(1), 1-20.

Zhao, J., et al. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2, 15-20.