Workshop on Corpus-linguistic Applications
ABSTRACTS

The discourse presentation of autism in the UK press: A case of critical corpus lexicography

Costas Gabrielatos, *Edge Hill University, UK*

This presentation will discuss the findings of the recent British Academy funded project, *Implicit Attitudes towards Autism in the British Press* (Maden-Weinberger et al., 2021; Karaminis et al., 2022). At the same time, it will show that, in many crucial respects, the methodological approach of discourse-oriented corpus studies (DOCS) is akin to critical corpus lexicography (Gabrielatos, 2011, 2021, 2022).

The overwhelming proportion of the information that dictionaries provide relate to a lexeme's senses, connotations, and sense relations, with reference to the lexicogrammatical patterns of the lexeme. Similarly, a core analytical technique in DOCS is the examination of the lexicogrammatical patterning of lexemes referring to particular *social actors* (van Leeuwen, 1996) in order to uncover their attributes, and the events and processes associated with them in discourse. In this light, the core aim of DOCS seems akin to creating dictionary entries on particular social actors, as emerging from the context-informed critical interpretation of the lexicogrammatical patterns of related lexemes in a corpus (e.g. collocates). It must be noted that *critical* is used here in the sense of adopting a questioning, evidence-based approach, rather than adhering to particular sociopolitical or critical theories (Chilton, 2012; Fjørtoft, 2013; Peters et al., 2006).

The results discussed here relate to the collocation analysis of the nodes *autism** and *autistic**, and the attendant examination of their semantic preferences and discourse prosodies (Stubbs, 2001). The results indicated that discourse presentations of autism draw mainly from the disability-based model, as evidenced by frequent associations with neurodevelopmental, physical and mental health conditions. Autism and autistic individuals are routinely discoursally connected with issues of abuse, lack of social support, and discrimination. All the above indicate that autistic individuals are rarely portrayed as living successful and independent lives. Also, autistic individuals are typically presented as being children (particularly boys), are predominantly presented as lacking agency, and are discussed from the perspective of their parents. Emergent representations of autism are also frequently differentiated by gender. Autistic girls are discussed in contrast with autistic boys, as they are presented as a newly-identified group, and are discussed more often as a group, whereas boys are more frequently discussed as individuals. Gender-related biases also extended to autistic children's families, with mothers, rather than fathers, presented as the carer.

References

Chilton, P.A. (2013) “Critical” in Critical Discourse Analysis. In Chapelle, C.A. (ed.) *The Encyclopedia of Applied Linguistics*. Wiley. 1-7.

Fjørtoft, M.R. (2013) The critical element of Critical Discourse Analysis. *SYNAPS - A Journal of Professional Communication*, 28. 67-75.

Gabrielatos, C. (2011). Collocational approaches to critical discourse studies: A case of critical corpus lexicography. Invited lecture: *International Symposium On The Sociology Of Words: Lexical Meaning, Combinatorial Potential and Computational Implementation*. LACELL, University of Murcia, Spain, 1-2 December 2011.

Gabrielatos, C. (2021). Discourse-oriented corpus studies as critical lexicography. Invited presentation: *PHRASALEX II: Phraseological Approaches to Learner’s Lexicography*. Hildesheim University, Germany. 22-23 July 2021.

Gabrielatos, C. (2022). Corpus lexicography and critical discourse studies. Invited presentation: *Critical Discourse Studies and Social Change. Where are we at?* University of Granada, Spain, 16-17 February 2022.

Karaminis, T., Gabrielatos, C., Maden-Weinberger, U., & Beattie, G. (2022) Portrayals of autism in the British press: A corpus-based study. *Autism*, Online First https://doi.org/10.1177/13623613221131752

Maden-Weinberger, U., Gabrielatos, C., Beattie, G. & Karaminis, T. (2021) Implicit attitudes towards autism in the British Press. *Autistica Research Festival*, 12-16 July 2021.

Peters, P., Tent, J. & Fernandez, T. (2006) Critical Lexicography. *Euralex 2006*. 561-565.

Stubbs, M. (2001) *Words and Phrases: Corpus Studies of Lexical Semantics.* Blackwell.

van Leeuwen, T. (1996) The representation of social actors. In Caldas-Coulthard, C.R. & Coulthard, M. (eds.) *Texts and Practices: Readings in Critical Discourse Analysis*. Routledge. 32-71.

Annotation schema and template for spoken and written corpus

Athanasios Karasimos, *Aristotle University of Thessaloniki*

Corpus annotation is the practice of adding interpretative linguistic information to a corpus. In particular, one common type of annotation is the addition of tags indicating the word class, the thematic role or morpheme class to which words in a text belong or even the errors, phonetic/phonological values or register changes. Some researchers prefer the unannotated corpus, as it is the 'pure' corpus to be investigated (an annotated corpus possibly reflecting the predilections, or even the errors, of the annotator). However, annotation is a means to make a corpus much more useful —an enrichment of the original raw corpus. From this perspective, adding annotation to a corpus is giving 'added value', which can be used for research by the individual or a team that carried out the annotation, but which can also be passed on to others who may find it useful for their own purposes. The advantages of an annotated corpus help the researchers and educators to think about the standards of good practice these corpora require,

such as manual examination and/or automatic analysis of a corpus, the crucial re-usability and multi-functionality of annotations.

Towards good practices in corpus annotations, it is important to recommend a set of standards of good practice to be observed by annotators wherever possible. It is suggested that annotations should be separable, detailed and explicit documentation should be provided, the annotation practices should be linguistically consensual and respect emergent *de facto* standards. These standards should be a broad guidelines schema to build a proper and well-encoded annotations schema. Based on these standards we are going to present, evaluate and modify different annotation schemata depending on the data nature, i.e., spoken or written corpora. Annotations can be tiered, which can be hierarchically interconnected, while at the same time they can be synchronized with the audio signal or audio-visual file or raw text input or aligned with other existing annotations. For this purpose, we will present different annotation typologies and templates in several tools (ELAN, Praat, CATMA), and we will properly build an annotation template encoded in XML, following some 'provisional standards' of best practice for different linguistics levels.

Using corpora for cross-linguistic research

Thomi Dalpanagioti, *Aristotle University of Thessaloniki*

This workshop aims to demonstrate how we can search and interpret corpus data in light of cognitive linguistic theories for the purposes of contrastive analysis. Considering pending questions about the kind of empirical data and the nature of the “tertium comparationis” that can yield robust cross-linguistic insights, we draw on different lines of research in the field of contrastive linguistics. More precisely, we combine comparable with parallel corpora as sources of data and we explore the potential of semantic frames to serve as a common ground for cross-linguistic comparison (Enghels, Defrancq & Jansegers, 2020: 5-7). Another strand of research we build on is contrastive phraseology which, inspired by Sinclair’s model of extended units of meaning, focuses on corpus-based contrastive analysis of “patterns”, i.e. recurrent multi-word combinations that function as semantic units (Ebeling & Oksefjell Ebeling, 2013: 1).

The focus of attention will be on the fluidic motion uses of the highly polysemous verbs *run* and *τρέχω* in English and Modern Greek respectively. First, we will investigate cross-linguistic correspondence on the basis of monolingual corpus data and then we will use parallel corpus data to identify shifts in translation. In both cases data will be drawn from Sketch Engine (enTenTen20, elTenTen19 and OPUS2) and interpreted using Frame Semantics and Conceptual Metaphor and Metonymy Theory. The comparable monolingual analysis of formally similar items will reveal similarities and differences in the conceptual and phraseological patterning of the specific predefined items, while the subsequent parallel corpus investigation is more exploratory and will provide insights into formally dissimilar, yet functionally similar, correspondences.

References

Ebeling, J. & Oksefjell Ebeling, S. (2013). *Patterns in contrast*. Amsterdam/ Philadelphia: John Benjamins.

Enghels, R., Defrancq, B. & Jansegers, M. (2020). Reflections on the use of data and methods in contrastive linguistics. In *R. Enghels*, *B. Defrancq* & *M. Jansegers* (Eds.), New *approaches* to *contrastive linguistics*. *Empirical and methodological challenges*. Berlin/ New York: Mouton de Gruyter, 221-264.

Applying a narratological approach to the design and analysis of literary corpora

Filio Chasioti, *City College, University of York, Europe Campus*

The aim of this workshop is to introduce a literary-inspired analytical methodology for literary corpora, one that is informed by literary/theoretical notions as it is by linguistic terminology. It seeks to bring forth the centrality of the content of a literary piece, and its narrative elements, as majorly influential in all the steps of the corpus compilation and analysis procedure. It seeks to inspire the contextualization of findings in terms of literary theoretical notions and criteria.

The workshop will focus on the design and analyses of literary corpora for the purposes of a lexico-narratological approach. It will consist of both a theoretical and a practical component. During the theoretical component, narratological elements, such as those of overt and covert narrator (Genette, 1980), of focalization (Genette, 1972; 1980; Rimmon-Kennan, 2002), of the narrative articulation of time (Genette, 1972; 1980), as well as of the element of character in storytelling will be discussed. The purpose of the brief focus on these notions is the creation of an interdisciplinary connection between structural elements of fiction and the influence they may exert on the design and analyses of literary corpora. During the theoretical part, the excerpt (fiction) that will be analyzed in the second part will also be discussed in terms of content, genre, and the relevant narratological elements.

The second part will consist in problematizing the corpus design process following narratological criteria, as these are instantiated in the excerpt in question. The Sketch Engine software (Kilgarriff et al. 2014) will be used for the corpus compilation process. A set of analyses will be performed (Keyword lists; n-grams; Word Sketch analysis) aimed at providing lexical evidence at a discourse level which will then be engaged with in terms of the narratological notions, and what this data might reveal in reference to the narrative processes of the text.

References

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engline: Ten years on. *Lexicography, 1*, 7-36. Retrieved from http://www.sketchengine.eu

Genette, G. (1972, 1980). *Narrative discourse: An essay in method.* (J. E. Lewin, Trans.) Ithaka, New York: Cornell University Press.

Rimmon-Kenan, S. (2002). *Narrative fiction: Contemporary poetics.* London, New York: Routledge.

Using Word Sketches in discourse-oriented corpus studies

Costas Gabrielatos, *Edge Hill University, UK*

This workshop will focus on using the *word sketch* function in Sketch Engine to examine the discourse presentation of social actors. The workshop ties in with the view of DOCS as a form of critical lexicography discussed in the lecture, as Sketch Engine was first designed as a lexicographical corpus tool.

Work sketches provide co-occurrences of a *node* (the lexical item in focus) with other lexical items within particular syntactic frames (e.g. when the node is the subject). Word sketches have close affinity with two theoretical constructs within the Neo-Firthian paradigm and Construction Grammar, respectively: *colligation*, the co-occurrence of lexis and grammatical frames (Hoey, 1997: 8; Stubbs, 2002: 238), and *collostruction*, the attraction/repellence of lexis to particular slots within a construction (Gries & Stefanowitsch, 2004; Stefanowitsch & Gries, 2003).

In lexicography, word sketches help derive not only senses, but also lexicogrammatical patterns of use. In DOCS, word sketches help construct discoursal definitions of social actors, as they reveal not only the events and processes they are associated with, but also their role in them as actors or recipients (e.g. Baker et al., 2013; Balfour, 2019), as well as their implicit associations with events and processes, and other actors, in lists or contrasts (e.g. Karaminis et al., 2022; McEnery, 2006).

The workshop aims to help participants:

- Select a corpus
- Derive a word sketch
- Decide on statistical cut-off values
- Carry out a semantic preference analysis (a.k.a. identifying frequent topics)
- Derive random samples for discourse prosody analysis

An equally salient aim is to dispel sanitised conceptions of corpus linguistic processes. More specifically, the workshop also aims to help participants understand that

- the process is not necessarily linear or smooth or fully automated;
- there are implicit theoretical assumptions behind ‘default’ techniques and settings;

- there are methodological decisions that need to be made at every step - decisions that are usually hidden by 'default' techniques and settings;
- the output of any corpus tool needs to be approached critically.

References

Baker, P., Gabrielatos, C. & McEnery, T. (2013) Sketching Muslims: A corpus-driven analysis of representations around the word "Muslim" in the British press 1998-2009. *Applied Linguistics*, 34(3), 255–278.

Balfour, J. (2019) 'The mythological marauding violent schizophrenic': Using the word sketch tool to examine representations of schizophrenic people as violent in the British press. *Journal of Corpora and Discourse Studies*, 2, 40–64. https://doi.org/10.18573/jcads.10

Gries, S.Th. & Stefanowitsch, A. (2004) Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics,* 9(1), 97–129.

Hoey, M. (1997) From concordance to text structure: New uses for computer corpora. In Melia, J. & Lewandoska, B. (eds.), *PALC `97: Practical Applications in Language Corpora*. Łódź: Łódź University Press. 2-23.

Karaminis, T., Gabrielatos, C., Maden-Weinberger, U., & Beattie, G. (2022) Portrayals of autism in the British press: A corpus-based study. *Autism*, Online First: https://doi.org/10.1177/13623613221131752

McEnery, T. (2006) *Swearing in English: Bad language, purity and power from 1586 to the present*. Routledge.

Stefanowitsch, A. & Gries, S.Th. (2003) Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8(2), 209-243.

Stubbs, M. (2002) Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics,* 7(2), 215-244.

Corpus-based techniques for practitioners of EFL: how to use DIY and free online corpora in classroom

Vasiliki Papaioannou, *3rd High School of Orestiada*

The use of corpora in the language classroom represents an opportunity to teach authentic language while allowing learners to become 'language detectives' and autonomously study grammatical and lexical patterns. The aim of this presentation is to provide suggestions and practical advice on the following: a) What are the advantages of using corpora in an EFL classroom? and, b) How can corpora be used by practitioners of EFL? We propose a "blended" corpus based approach to the teaching of English modal verbs, which involves students' conducting online queries and working with handouts of printed concordances from various sources: a) a DIY English course book corpus (ECCo), b) free-to-access materials from the Web, and c) the Corpus of Contemporary American English (COCA). We present the step-by-step application of training sessions and corpus-based lessons on modal verbs in a Greek high school

class, with a particular focus on the adaptation of existing textbook materials, as well as our students' opinions of the corpus-based EFL lessons.

Introduction to ICLE version 3

Thomas Zapounidis, *3rd Experimental Primary School of Evosmos*

The aim of this workshop is to present the third version of the ICLE (International Corpus of Learner English) project and provide participants with the opportunity to familiarize themselves with the interface of the new platform. ICLE is a corpus of written productions by young adults, undergraduate students of various academic institutions in Europe. All of them were English as a foreign language learners at an upper-intermediate to advanced level of proficiency. Workshop participants will be presented with a brief overview of this international project and its development over the years, as well as the design criteria of the corpus, i.e. learner variables, task variables, Markup and linguistic annotation. Finally, participants will have the opportunity to familiarize themselves with the interface of the new platform through various practical examples. By the end of the workshop participants will have gained an overview of the ICLE v3 platform as well as practical experience in extracting data from the project's resources.

Speakers' bionotes

Costas Gabrielatos (Gabrielc@edgehill.ac.uk, https://ehu.ac.uk/gabrielatos) has been working in corpus linguistics since 2001. He came to linguistics via English language teaching (1984-1993) and language teacher education (1992-2001). His current research focuses on corpus linguistics, lexicogrammar, and discourse studies. He has also published on TESOL methodology, language teacher education, and the connection between linguistics and TESOL. He teaches modules on research methodology, English lexicogrammar, and discourse studies. He supervises postgraduate research students employing corpus-based approaches to theoretical lexicogrammar and critical discourse analysis. He is the Editor of the *Journal of Corpora and Discourse Studies* and his current research projects include *Portrayals of Greta Thunberg in the British Press.*, *Implicit Attitudes towards Autism in the British Press.*

Athanasios Karasimos (akarasimos@gmail.com) is a graduate of the Department of Philology, University of Patras. He holds two European Masters in Speech and Language Processing (one of them at the University of Edinburgh) and his doctoral dissertation is in Computational Morphology. He is an Assistant Professor in Computational Linguistics, at the School of English, Aristotle University of Thessaloniki. He participated in several research projects on Modern Greek dialects, corpora, aphasic speech, Digital Humanities and training of English language teachers. He was a postdoctoral research fellow funded by IKY. He was a researcher in the national infrastructure for Digital Humanities DARIAH-GR / DYAS (Academy of Athens). His research interests focus on computational linguistics and machine learning, the use of corpora, education technology and integrating video and board games into language teaching and learning.

Thomi Dalpanagioti (thomdalp@enl.auth.gr) is a Laboratory Teaching Fellow at the School of English, Aristotle University of Thessaloniki. She holds an MA and a PhD in Linguistics-Lexicography from the National and Kapodistrian University of Athens. Her research interests lie in the areas of lexicology, lexicography and vocabulary acquisition. In particular, her research focuses on the application of Cognitive Linguistics (CMT, Frame Semantics) and Corpus Linguistics in the study of polysemy and phraseology. She has been a recipient of A.S. Hornby Dictionary Research Awards 2022.

Filio Chasioti (filiochasioti@gmail.com) holds an interdisciplinary PhD in corpus linguistics and contemporary north American literature. The works of Margaret Atwood constitute her primary area of study, while her interests lie in the field of text linguistics, narratology, the dystopian genre, and contemporary North American fiction. She currently works as a lecturer at City College, University of York Europe Campus, where she teaches literary modules and corpora in applied linguistics.

Vasiliki Papaioannou (vass_papa@yahoo.gr) is a graduate of the school of English Language and Literature (AUTh). She holds a PhD in Applied Linguistics (AUTh), an MA in Language and Communication Studies (AUTh) and an MSc in Machine Translation (UMIST, UK). She has been teaching in Secondary Education for 20 years and she also holds a Teaching Position at the Hellenic Open University (Education and Technologies in Distance Teaching and Learning Systems – Educational Sciences). Her teaching experience includes teaching in Higher Education (Democritus University of Thrace 2006-2008; 2009- 2010; 2021-2022). She is a teacher trainer in the use of ICT in education and in Innovation in Education, participating in national training courses organized by the Institute of Educational Policy (IEP). Her research interests include Corpus Linguistics, use of ICT in education and Distance Learning.

Thomas Zapounidis (thomaszapounidis@gmail.com) holds a Ph.D. in Applied Linguistics and an MA in the Teaching of English as a Second Language from the Aristotle University of Thessaloniki. He is currently conducting his postdoctoral research compiling and analyzing written learner corpora. His research interests lie in the area of Corpus Linguistics and more specifically, in learner and pedagogic corpora. Other research interests include the instruction of English as a foreign language to young learners and CEFR vocabulary analyses.