L2P-2023: A book of abstracts

Organizer talks

Swedish Lexical profile

David Alfter, david.alfter@gu.se
University of Gothenburg, Gothenburg Research Infrastructure for Digital Humanities, Sweden

Vocabulary plays an important role in language learning, as “while without grammar very little can be conveyed, without vocabulary nothing can be conveyed” (Wilkins 1972: pp. 111-112). Recent advances in technology have made it possible to move from manually crafted vocabulary lists to automatically derived word lists. The Swedish Lexical profile groups together recent efforts that focus on vocabulary for learners using computational approaches.

The Swedish Lexical Profile is based on previous work for Swedish vocabulary lists, namely SVALex (François et al. 2016) and SweLLex (Volodina et al. 2016), two resources automatically extracted from textbooks and from learner essays respectively. These resources also contain information about the CEFR levels at which words occur. However, these resources have a few drawbacks: first of all, distributions across CEFR levels do not allow us to draw conclusions as to which level the word should be considered criterial for. Second, these lists do not distinguish word senses.

In this talk, we will describe how we automatically linked vocabulary to CEFR levels based on CEFR distributions, as well as a machine learning algorithm that can predict the CEFR level for any given word. We then detail the creation of a semi-manually curated sense-based list, Sen*Lex, regrouping SVALex and SweLLex but with sense distinctions. We finish by an overview demo of the resources.

References

François, T., Volodina, E., Pilán, I., & Tack, A. (2016). SVALex: a CEFR-graded lexical resource for Swedish foreign and second language learners. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 213-219).

Volodina, E., Pilán, I., Llozhi, L., Degryse, B., & François, T. (2016). SweLLex: second language learners’ productive vocabulary. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition* (pp. 76-84).

Wilkins, D. A. (1972). *Linguistics in language teaching* (Vol. 111). London: Edward Arnold.

Swedish Grammar profile

Therese Lindström Tiedemann, therese.lindstromtiedemann@helsinki.fi
University of Helsinki, Finland

The Swedish L2 grammar profile provides a usage-based comparison of various aspects of Swedish grammar that we know are of particular interest in relation to the development of Swedish as a second language. The profile is based on automatic annotation (part-of-speech tagging and morpho-syntactic descriptors) as well as certain lexical and syntactic characteristics which can be used well in corpus queries.

The Swedish L2 Grammar profile contains two main features: 1) Nominal patterns which focus on the category of definiteness and how this is expressed in Swedish and 2) verbal patterns which focus on tense, mood and voice in Swedish. Through these patterns the profile facilitates the study of these features in L2 Swedish course books which can be seen to represent receptive competences at a given CEFR-level and L2 Swedish learner essays which represent productive competences at a given CEFR-level. Furthermore, links to the corpus search tool (Korp) enable us not only to see the actual data in the corpora which the profile is based on, but also to rerun the same searches in reference corpora which is incredibly useful for research, teaching and material design.

Other aspects of the Swedish L2 profile enable us to compare other grammatical features such as: gender, verbal conjugations, nominal declinations, the use of particular parts of speech (e.g. prepositions or conjunctions).

Swedish Morphological profile

Elena Volodina, elena.volodina@svenska.gu.se
University of Gothenburg, Språkbanken Text, Sweden

Our knowledge of what learners know - or should know - is formed from observations of their linguistic behavior collected from different sources. It is operationalized in different ways, one of the most influential ones being the CAF model (Michel 2017), which describes **C**omplexity, **A**ccuracy and **F**luency of the learner language, as evidenced in both written and spoken language. The picture is very complex since it involves various aspects of language, therefore linguists need to subpartition linguistic constructs into manageable parts, such as lexical complexity and grammatical complexity. While lexical and grammatical aspects of learner language have been given a fair share of attention, morphological complexity suffers from very little attention.

In this talk, we will focus on the morphological profile for L2 Swedish, which organizes all L2-related vocabulary into morpheme families based on word-building morphemes. For example, we can see all words that have suffix *-ing* and see their appearance at different levels of proficiency, or all words that share the same root *stud-* (‘study’). We can study frequency patterns of appearance of certain morphemes and growth of their use over time.

We will demo the tool, which is freely available online, describe the process of its creation, and potentially offer a small teaser exercise for the audience.

References

Michel, M. (2017). Complexity, accuracy, and fluency in L2 production. In *The Routledge handbook of instructed second language acquisition* (pp. 50-68). Routledge.

Invited talks

Aligning leaners' dictionaries with the CEFR: the case of Estonian Vocabulary and Grammar Profiles.

Jelena Kallas, jelena.kallas@eki.ee
Institute of the Estonian language (Estonia)

In the talk, we will introduce the Estonian Vocabulary Profile and the Estonian Grammar Profile, which are designed to support the CEFR illustrative descriptors scales of linguistic competence with language-specific descriptions. We will focus on the methodology and corpora that were used for the development, trial and validation of the Estonian Grammar Profile. Currently, the profile provides descriptions of grammar competence on the morphology, derivation, phrase and sentence levels, from the pre-A1 level up to the B2 level for young learners, and from A1 to C1 for adult learners. All descriptions are equipped with example sentences compiled either by experts or taken from the coursebook and learner’s corpora.

In addition, we will address the issues related to our attempt to combine this resource with the Estonian learners’ dictionary Sõnaveeb for Learners. The dictionary is compiled in the Dictionary Writing System Ekilex, whose long-term goal is to have a single data source that provides consistent and comprehensive information about Estonian, including CEFR labels. We will report on the work in progress from the point of view of data modelling. Given a construction-based and usage-based understanding of L2 acquisition, we assume that linguistic knowledge at a particular proficiency level is not best described as a set of words and a set of grammatical structures, as is the current practice, but rather as a set of combinations of particular word meanings and forms with particular schematic constructions. This means that the lexicographic resource must include descriptions of grammatical constructions, and that the language proficiency level should be attributed not to lemmas and constructions, but to particular word meanings in particular forms and in particular constructions.

Building on insights from the English Grammar Profile: From *really good* to *painfully obvious*

Geraldine Mark, germark@icloud.com
Cardiff University, Wales

The English Grammar Profile (EGP) Project was a four-year quasi-longitudinal study investigating learner grammar from the Cambridge Learner Corpus (CLC). The main output of the research is the EGP, a free educational online database, which provides a profile of over 1,200 corpus-based grammar competency statements about learner grammar use across the six CEFR levels. In the first part of this talk I'll describe the methodology that we developed to build the EGP, discuss the key insights from the study and show how the investigation has enhanced our understanding of the developmental nature of grammar acquisition and use. I'll then look at further ways to explore the data taking a usage-based (UB) approach. UB studies have shown that language users are sensitive to the statistics of repeated patterns in language and that we figure out 'structural regularities' in language as we subconsciously tune into mappings of form and meaning (Ellis et al. 2016). Using this large scale proficiency-levelled data I will look at how we can use corpus tools to investigate if and how structural regularities develop in L2 English and how this might offer further insight into learner language development.

References

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of construction grammar. Oxford: Wiley.

Profiling complexity: methodological issues and applications to L2 morphology

Gabriele Pallotti, gabriele.pallotti@unimore.it
University of Modena and Reggio Emilia, Italy

In this talk I will first discuss how linguistic complexity should be theoretically defined and practically operationalized, in a wider context of interlanguage analysis and linguistic profiling. In particular, I will argue that it needs to be kept apart from other constructs such as processing difficulty or developmental timing. Then, I will present an approach to empirically measuring morphological complexity, its conceptual and methodological challenges and how they were addressed in the development of an online morphological complexity analyzer.

We need to know more about relative complexity and learnability

Aleksandrs Berdicevskis, aleksandrs.berdicevskis@gu.se
University of Gothenburg, Språkbanken Text, Sweden

For the last two decades, language typology and related fields have witnessed a hot debate about language complexity. Several influential theories have emerged that claim that languages are not equally complex, and that the distribution of complexity depends on social factors, such as number of speakers, degree of language contact and number of non-native speakers. I will briefly review some recent evidence in favour and against those theories and argue that whether the theories are correct or not, they make interesting non-trivial hypotheses about mechanisms of language learning and language change. I will then make my main point, which is that these hypotheses cannot be properly addressed without a deep understanding of second language acquisition, most crucially, the concepts of relative complexity ("what is difficult for whom") and learnability. The talk will mostly focus on morphological complexity.

Talks: Lexical profiling

Linking CEFR-based learner profiles to lexicographic data

Kris Heylen[1], Ilan Kernerman[2], Carole Tiberius[1]

[1] Dutch Language Institute, The Netherlands kris.heylen@ivdnt.org

[2] Lexicala by K Dictionaries, Israel ilan@lexicala.com

Dictionaries have a long tradition of supporting second language learning. Their highly structured and thoroughly edited lexical knowledge bases provide learners, teachers and course developers with concise but clear and reliable information about form, meaning, grammatical properties and use of vocabulary items. However, apart from a handful of exceptions[1], most dictionaries do not contain information about the proficiency level at which a lexical item is (or should be) learned, making it difficult to integrate them directly into language learning programmes based on the Common European Framework of Reference for Languages (CEFR). On the other hand, CEFR-graded word lists have been developed for a number of languages through collaborations within the Second Language Acquisition (SLA) and the Natural Language Processing (NLP) communities[2]. Yet, these resources are often compiled (semi-)automatically, using different methodologies, with limited (post-)editing and no or a limited regard for sense distinctions. In this presentation, we first discuss the challenges of linking existing CEFR-graded lexical resources and lexicographic knowledge bases. Next, we look at the potential benefits of integrating both types of resources within a linked data infrastructure. Our case studies involve, on the one hand, the CEFR-graded resources compiled in different CEFRLex projects (François and colleagues, ongoing) and, on the other hand, the lexicographic databases of K Dictionaries and the Dutch Language Institute. Specific challenges include (a) evaluating the probabilistic CEFR-gradings and linking these to the appropriate word senses in dictionaries, (b) ensuring that additional lexicographic components (examples, definitions etc.) are also on the appropriate CEFR level, and (c) creating CEFR lists for additional languages that can be consistently linked with multilingual lexicographic data. We argue that the potential benefits of linking CEFR-based and lexicographic resources include (a) easier access to combined resources for developers of language learning programmes, (b) the linking of CEFR-based resources to the existing lexicographic research infrastructure[3] and to the Linguistic Linked Open Data cloud[4] to enable new types of research and applications, and (c) the faster development of CEFR-based resources for additional (under-resourced) languages. To conclude, we discuss avenues for collaboration between partners from the SLA, NLP and lexicographic communities in a future project.

Keywords

language learning; proficiency levels; CEFR lists; lexicographic resources, LLOD cloud

References

[1] E.g., Cambridge and Oxford advanced learner’s dictionaries for English, Sõnaveeb for Estonian

[2] a.o. the Kelly and CEFRLex projects

[3] ELEXIS (EU Horizon-2020 Research Infrastructure project, 2018-2022), https://elex.is/

[4] https://linguistic-lod.org/

Using Learner language models for lexical profile

Bernardo Stearns, bernardo.stearns@insight-centre.org
University of Galway, Insight-centre for Data Analytics, Ireland

This talk will describe work in progress with using probability outputs of trained learner language models and word embeddings to investigate lexical competence. BERT has been fine-tuned with the EFCAMDAT according to levels and assumed mother tongues. We will discuss several strategies based on what language models do best : predicting the next token.

We explore how can masking error-prone sentences be used with learner language models to estimate lexical profiles and how does this offer a window unto learner lexical competence. We have identified several masking strategies to use probes of the learner language model to simulate lexical competence as instantiated in next token prediction tasks.

We discuss three of them being explored :

1) adapting the beam search method typically used In Neural Machine Translation to survey the alternative words that were eventually discarded by the model (in decreasing probability). We will explore this virtual paradigm provided by alternative next tokens of decreasing probability and what it reveals of lexical knowledge. The set of alternative tokens can be analysed in terms of surprisal and of distance to the tokens predicted by BERT.

2) Masking hypernyms (things). Learners are more likely to use hypermyms (thing) used as metawords. Masking different types of synsets (thing, do, nice) could allow the lexical network of alternative token to be surveyed using Wordnet (Fellbaum, 2005)

3) masking tokens marked as errors where we investigate the error-annotated corpora to try to predict a given type of error.

Using a learner corpus to design a phraseological syllabus of Italian collocations

Francesca La Russa and Maria Roccaforte, franlarussa3@gmail.com
Sapienza Università di Roma, Italy

Lexical combinations are central to language learning because they can be processed quickly (Siyanova-Chanturia, 2015) and their use gives the idea of fluency in production (Nattinger & DeCarrico, 1992). However, the acquisition of L2 phraseological competence is often difficult for learners. This is particularly true for collocations, "sequences of words which tend to occur in stable and privileged combinations" (Simone, 1990: 440). The semantic transparency of collocations facilitates their understanding and makes them difficult to notice. Since collocations are often not highlighted in language courses, learning them is even more difficult because students do not notice and assimilate them as complex lexemes (Bini et al., 2007). As a matter of fact, in Italian L2 syllabuses, vocabulary is often presented as a list of single words and the phraseological dimension is usually absent. To fill this gap, we designed a syllabus of Italian collocations.

Following the model of the English Vocabulary Profile , a descriptive rather than a prescriptive approach was chosen. Thus, collocations were extracted from the CELI learner corpus (Spina et al. 2022) which collects 3041 written texts produced by learners of Italian L2 who passed the CELI exams (levels B1, B2, C1, C2) and provides reliable data on learners' authentic use of the language. To assign each collocation to the appropriate proficiency level, the following criteria were adopted:

- frequency of the collocation in the Perugia Corpus (PEC) (Spina, 2014) - which collects written and oral texts produced by native speakers;
- number of occurrences of the collocation in the four CELI subcorpora;
- presence of the collocates in the lexical lists of the Profilo della lingua italiana (Spinelli & Parizzi, 2010);
- topic.

The result is a syllabus in which Italian collocations are organized according to the proficiency level they should be taught and the topic they refer to.

References

Bini, Milena / Pernas, Almudena / Pernas, Paloma. 2007. “Apprendimento e insegnamento collocazioni dell’italiano. Con i NUNC più facile”. In Corpora e linguistica in rete, edited by Manuel Barbera / Elisa Corino/ Cristina Onesti, 323–333. Perugia: Guerra Edizioni.

Nattinger, James / DeCarrico, Jeannette. 1992. Lexical phrases and language teaching. Oxford University Press.

Simone, Raffaele. 1990. Fondamenti di linguistica. Bari: Laterza.

Siyanova-Chanturia, Anna. 2015. “On the ‘holistic’nature of formulaic language”. Corpus Linguistics and Linguistic Theory, 11(2): 285-301.

Spina Stefania / Fioravanti Irene / Forti Luciana / Santucci Valentino / Scerra Angela / Zanda Fabio. 2022. “Il corpus CELI: una nuova risorsa per studiare l’acquisizione dell’italiano L2”. Italiano LinguaDue, 14(1):116-138.

Spinelli, Barbara /Parizzi, Francesca. 2010. Profilo della lingua italiana. Livelli di riferimento del QCER A1, A2, B1, B2. Milano : La Nuova Italia.

Lexical features in adolescents' writing: Insights from the trilingual parallel corpus SWIKO

Nina Hicks, nina.hicks@unifr.ch
University of Fribourg, Switzerland

We investigate the influence of task characteristics on lexical features among adolescents' writing based on the multilingual parallel corpus SWIKO (Karges et al. 2019, 2022). Corpus-based language acquisition research usually relies on learner language elicited through a single or range of tasks. However, the characteristics of these tasks influence the extent to which learners can demonstrate their language competences. Despite considerable research on task effects on language performance focusing predominantly on English (e.g., Alexopoulou et al. 2017), little is known about the combination of different characteristics and the cross-linguistic generalizability of these findings. To address these challenges, Swiss secondary school students’ trilingual productions in SWIKO are based on eight different tasks, which are systematically varied by rhetorical type, topic, and structuredness. We analyze how these characteristics relate to four lexical richness features (i.e., lexical density, diversity, sophistication, and errors; Read 2000) in the resulting productions across three languages (German, French, and English) and two acquisitional types (language of schooling L1 and foreign languages L2). Preliminary results suggest that rhetorical mode had the largest impact, affecting density and sophistication across all three languages in both acquisitional types as well as diversity in L2. Topic did not differentiate between any L1 productions and structuredness mostly influenced sophistication. Across all task characteristics and acquisitional types, German productions were affected the most and French productions the least. Overall, our findings highlight the importance of careful task selection in both first and second language education and research.

References

Alexopoulou, T., Michel, M., Murakami, A. & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. Language Learning, 67(S1), 180-208.

Karges, K., Studer, T. & Wiedenkeller, E. (2019). On the way to a new multilingual learner corpus of foreign language learning in school: Observations about task variations. In A. Abel, A. Glaznieks, V. Lyding & L. Nicolas (Eds.). Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference., Corpora and Language in Use – Proceedings 5, Louvain-La-Neuve: Presses Universitaires de Louvain, 137-165.

Karges, K., Studer, T. & Hicks, N. S. (2022). Lernersprache, Aufgabe und Modalität: Beobachtungen zu Texten aus dem Schweizer Lernerkorpus SWIKO. Zeitschrift für germanistische Linguistik, 50(1), 104-130.

Read, J. (2000). Assessing Vocabulary. Cambridge: Cambridge University Press.

Talks: Grammar profiling

Profiling learner Finnish and Estonian: interaction of frequency and accuracy as an indicator of language skills

Annekatrin Kaivapalu, annekatrin.kaivapalu@helsinki.fi
University of Helsinki, Finland

This presentation introduces the following language resources of written learner Finnish and Estonian aligned with language proficiency levels of the Common European Framework of Reference for Languages (CEFR):

1) the data collected for the projects Linguistic Basis of the Common European Framework for L2 English and L2 Finnish (Cefling) and Paths in Second language Acquisition (Topling);
2) Corpus of Advanced Finnish Learners
3) International Corpus of Learner Finnish;
4) two subcorpora of Estonian Interlanguage Corpus: the subcorpus of Estonian language proficiency examinations and the ScriptLog subcorpus
The aim of the presentation is to discuss the relationship between frequency and accuracy of using various constructions across functionally determined CEFR levels. The discussion is based on a large number of studies (see Eslon et al 2021, Martin 2022) conducted by applying the mentioned resources. The results show that an increase of use of the given constructions is often followed by an improvement in accuracy at the next CEFR level. Thus, the interaction of frequency and accuracy of particular constructions can function as indicator of language skills.

References:

Eslon, Pille; Kaivapalu, Annekatrin; Õim, Katre; Kitsnik, Mare; Allkivi-Metsoja, Kais; Gaitšenja, Olga 2021. Eesti keele oskuse arenemine ja arendamine. Kirjalik õppijakeel [Development and devel-oping of Estonian language proficiency. Written learner language]. Annekatrin Kaivapalu, Pille Eslon (eds..). Tallinn: Eesti Keele Sihtasutus.

Martin, Maisa 2022. Käyttötaajuus ja tarkkuus toisen kielen kehityksessä [Frequency of use and accu-racy in second language development]. – Lotta Aarikka, Katri Priikki, Ilmari Ivaska (eds.) Soveltavan kielitietelijan sormenjälkiä etsimässä. In search of the fingerprints of an applied linguist. AfinLA-teema No. 14, 81–102

French Verb profile

David Alfter, david.alfter@gu.se
University of Gothenburg, Sweden

Morphological inflection is known to be difficult to master for L2 learners. In this presentation, we examine the state of the use of inflection in the verbal tense system among learners of French, and contrast it with the use in FFL textbooks. The objectives of our study are threefold: 1) To establish the distribution of verbal tenses on French textbooks in an automatic way, in order to obtain the first fully empirical and extensive resource on French verbal tenses; 2) To objectively describe the use of verbal tenses by learners of different CEFR levels; 3) To identify the tenses that learners struggle with. Through the description of the use of the tenses in the learners, we found that they had difficulty with the past perfect indicative, even at advanced levels. The proposed Verb Profile summarizes which tenses should be understood at which level, and as such can guide teachers and learners, as well as help pinpoint tenses that learners are underperforming on.

The FineDesc learner corpus: Making the CEFR/CV more user-friendly: fine-tuning descriptors with Learner Corpus Research results

María Belén Díez-Bedmar, belendb@ujaen.es
University of Jaén, Spain

The CEFR and, later, the CV, established a common metalanguage that encompasses the main aspects related to language teaching, learning, and assessment. Two further aims underpin the CEFR (North, 2007, p. 659) and, consequently, the CV: a) to promote reflection on learners' needs, set objectives and identify ways to follow up and check their progress; and b) to establish a series of levels which considers the learners' use of the language from a communicative point of view. These aims are encapsulated in the illustrative descriptors.

However, the overriding philosophy in the CEFR/CV, i.e., providing users with a document which may trigger reflection on the learning, teaching, and assessment of any language as well as providing a common standard for the levels, makes the use of the descriptors by CEFR/CV end-user problematic. The descriptors are perceived as too impressionistic and global in nature to provide a linguistic description of the type and quality of language when engaging in language activities in the different competences (Hawkins & Filipović, 2012; Hulstijn, 2007; North, 2007; Díez-Bedmar & Luque-Agulló, under review). This situation makes the use of the CEFR and the CV limited (Díez-Bedmar & Byram, 2018).

The main aim of the FineDesc project (http://web.ujaen.es/investiga/finedesc/index.php), funded by the Spanish Ministry of Science and Innovation Project (PID2020-117041GA-I00 funded by MCIN/AEI/10.13039/501100011033) is to provide CEFR/CV users with fine-tuned descriptors for L1 Spanish users of the CEFR/CV by complementing the information in the descriptors with information on the type and quality of the language produced by Spanish learners when they develop the communicative language competences at B1, B2 and C1 levels. To do so, the FineDesc learner corpus is being compiled with the texts produced by L1 Spanish learners of English in the CertAcles exam suite at those levels in ten University Language Centres in Spain. The CertAcles exam suite is accredited by ACLES (https://www.acles.es/index.php/en/what-is-acles/what-is-acles2), which is part of CerCles. Five main variables in the analysis of learner language are being considered to inform the descriptors, namely, CEFR level (B1, B2 and C1), the learner's L1, considering bilingual speakers (Spanish, Galician, Basque, Valencian, Catalan), status of English (EFL1, EFL2, ESL, etc.) and learner's gender.

Thanks to the FineDesc learner corpus analyses are now being conducted so that the results can inform the fine-tuning of the descriptors and make them more transparent to CEFR/CV users, thus fostering the use of the CEFR/CV in the learning, teaching and assessment of English as a Foreign Language by L1 Spanish learners.

This presentation will show the compilation criteria of the FineDesc learner corpus, an overview of the texts available at the time, as well as the results obtained so far in this project regarding the linguistic competence, specially regarding **NP complexity**.

Grammatical profiling with UD annotation (WiP)

Nicolas Ballier, nicolas.ballier@gmail.com
Joint Research with Cyriel Mallart and Thomas Gaillat (Rennes)

Université Paris Cité, France

This lightening talk will present the benefits of using Universal Dependency annotation for the investigation of learner grammatical profiles. The basic properties of the ConnL-format produced by UD annotation maximise the possibilities of queries based on syntactic (dependency relations), grammatical (pos) and morphological properties, and a combination of these features can be used to detect candidates for errors.

With GREW match SQL and the app GRE, one can extract error examples and potential detection rules from annotated Treebanks. These tools are used in the ANR Autogram project for typological linguistic descriptions based on rules extractions from Treebanks (investigations) and can be applied to CEFR-based learner groups within two SLA-oriented projects.

We have added an extra component to the query system. We have automated the UD annotation process, the querying of the CONLL-U data with GREW and the generation of a data respresentation all in one pipeline. The output dataset hinges on specifically-selected learner forms.

We will showcase some of the queries we have implemented to capture learner profiles. Two strategies have been explored : a kitchen sink method that makes the most of annotated properties and a preliminary grammar of queries likely to capture different profiles.

We will show we can capture linguistic micro-systems (Gaillat et al. 2021) or translate our micro-system analysis into queries for error detection. Profiling can be obtained by comparing the different sequences of UD/upos/morphological tags.

References

Gaillat, T., Simpkin, A., Ballier, N., Stearns, B., Sousa, A., Bouyé, M., & Zarrouk, M. (2021). Predicting CEFR levels in learners of English: The use of microsystem criterial features in a machine learning approach. ReCALL, 1-17. doi:10.1017/S095834402100029X
GREW syntax
http://universal.grew.fr/?corpus=UD_English-GUM@2.10
the autogramm project
https://autogramm.github.io/
GRE Grammatical rules extraction
https://github.com/santiagohy/grammar-rules-extraction

Prepositions in L2 Russian

Ekaterina Vlasova, ekaterina.vlasova@helsinki.fi
University of Helsinki, Finland

Online

This paper focuses on a quantitative analysis of the incorrect prepositional phrases produced by advanced Russian learners from Finland. The Russian prepositional phrase is a complex morphosyntactic phenomenon, as each preposition demands a certain case form, for example, dlja + Genitive case 'for', o + Prepositional case 'about'. The academic Russian grammars list about 35-64 such preposition-case combinations, which international learners often mixture. In my study, based on the error-annotated Russian learner corpus of 315, 000 tokens, I perform the linear regression analysis and explore the correlation between a total frequency of the preposition and a number of phrases with incorrect case government. The learner corpus effect assumes that the more frequent the preposition is, the more often it occurs with an incorrect case. The results show that a few prepositions, such as o, mezhdu and na, decline from the regression line. The paper discusses how statistic methods can be used for detecting the difficulties specific for language learners with different background, including heritage speakers and L2 learners.

Talks: Morphology profiling and word families

A Digital Dictionary of Romance Word Families

Christina Lindqvist, christina.lindqvist@sprak.gu.se
Mårten Ramnäs, marten.ramnas@sprak.gu.se

University of Gothenburg, Sweden

In this contribution, we will present a project in which we intend to create a multilingual digital dictionary based on Italian, French and Spanish word families. Although there are some word family based dictionaries for Romance languages (e.g. Colombo & D'Achille 2019 for Italian), there is no such dictionary that gathers several Romance languages. Also, the ones that exist are not available online. Many vocabulary acquisition researchers suggest that the concept of word families has pedagogical and learning advantages (Nation 2021, Webb 2021). Thus, the Digital dictionary of Romance Word Families will be an important asset to learners, teachers and researchers alike. The main research questions of the project are:

- To what extent is it possible to relate one word to one another on the basis of their roots?
- How can we build a word family using coherent criteria valid for French, Italian and Spanish at the same time?
- What is the best lexicographic solution for presenting a word family to the reader?
- Which are the word families that are most important for language learners?

These questions will guide our presentation at the workshop.

Talks: Starting new projects based on L2 data

From corpus to profiles: Icelandic L2 corpus

Isidora Glišić, isg14@hi.is
University of Iceland, Iceland

The presentation will introduce the Icelandic L2 Error Corpus, first learner corpus for Icelandic, created at the Language and Technology Lab at the University of Iceland. It is accessible on CLARIN: https://repository.clarin.is/repository/xmlui/handle/20.500.12537/280 and used to examine the potential of creating an automated skill level detection software based on textual input from language learners.

By extracting features from the corpus such as type and frequency of errors, along with lexical and syntactic characteristics (mean sentence and word length, most common words, and unique lemma count), interlanguage development can be mapped out. This information will help formalize the CEFR scale for Icelandic.

The texts in the corpus were labeled according to CEFR level and pre-annotated for errors. Preliminary analysis revealed a consistent decrease in the frequency of errors between skill levels, as well as variations in specific error categories. This and the result of analyzing other relevant lexical and syntactic features linked to skill level will be presented at the workshop.

A cross-section of linguistic competence of South Slavic university students learning Slovene as L2

Mojca Stritar Kučuk, mojca.stritarkucuk@ff.uni-lj.si
University of Ljubljana, Slovenia

The University of Ljubljana offers a free Slovene as a second language course, which is completed each year by around 250 regularly enrolled foreign students. At the end of the academic year 2021/22, as a part of a wider survey on the use of machine translation, all students who took the written exam wrote a short text (150-250 words) on one topic, describing and commenting on their first year at the University of Ljubljana. Thus, 210 texts were collected. They were mainly written by South Slavic speakers (i.e. speakers of Serbian, Bosnian, Croatian, Montenegrin and Macedonian) who had learnt Slovene language and studied in Slovene language for two semesters. The texts were digitised, anonymised and evaluated according to predefined criteria (content and coherence, vocabulary, grammar), which were based on the criteria used in the Slovene as a foreign language exams (level B2 according to CEFR). After resolving certain inconsistencies between the evaluators, we have so far carried out some preliminary analyses on the basis of the data collected. These have shown that speakers of Serbian got the highest average scores (7.5 out of 10), while speakers of Macedonian got the lowest (7.03). The most interesting finding is that there was no significant difference in the written production of students who were placed in beginner groups (average score 7.037) and students who had already been learning Slovene before their arrival to Slovenia and were placed in advanced groups (average score 7.039). In the future, we intend to conduct a more in-depth analysis of the data, comparing it with the data from Slovene learner corpus KOST (https://www.cjvt.si/korpus-kost/) and the list of core vocabulary for Slovene as a L2 (http://hdl.handle.net/11356/1697) and, of course, including our main findings in our teaching process and material.