

Automatic Extraction of the Romanian Academic Word List: Data and Methods

Ana-Maria Bucur^{1,2}, Andreea Dincă², Mădălina Chitez², Roxana Rogobete²

¹ Interdisciplinary School of Doctoral Studies, University of Bucharest

² West University of Timișoara, Romania

ana-maria.bucur@drd.unibuc.ro

{madalina.chitez, andreea.dinca, roxana.rogobete}@e-uvt.ro

Abstract

This paper presents the methodology and data used for the automatic extraction of the Romanian Academic Word List (Ro-AWL). Academic Word Lists are useful in both L2 and L1 teaching contexts. For the Romanian language, no such resource exists so far. Ro-AWL has been generated by combining methods from corpus and computational linguistics with L2 academic writing approaches. We use two types of data: (a) existing data, such as the Romanian Frequency List based on the ROMBAC corpus, and (b) self-compiled data, such as the expert academic writing corpus EXPRES. For constructing the academic word list, we follow the methodology for building the Academic Vocabulary List for the English language. The distribution of Ro-AWL features (general distribution, POS distribution) into four disciplinary datasets is in line with previous research. Ro-AWL is freely available and can be used for teaching, research and NLP applications.

1 Introduction

Since academic language differs from everyday social language and is an essential acquisition target in current education, extracting salient features contributes to linguistic, register, genre and disciplinary feature identification that can benefit students, teachers and educational app developers alike. Compiling an Academic Word List (AWL) is an effective solution to support both language teaching and NLP tasks. From the didactic perspective, AWLs reflecting either the L1 (i.e. mother tongue) or the L2 (i.e. foreign language) academic vocabulary can be used to offer linguistic support to novice academic writers in the form of discipline-specific and general lexical prompts. Teachers of all disciplines can integrate AWLs into teaching materials to help students write better (see, for example, Wangdi and Shimray (2022)).

NLP studies can exploit AWL datasets on topics such as text classification (Zampieri, 2012) and topic modelling (Murakami et al., 2017). For example, field-specific academic lists can be used to automatically classify texts into disciplinary areas. The same can be applied for the automatic distribution of texts in academic versus non-academic batches. In machine learning methods for language modelling tasks, AWLs are essential in training models to generate accurate academic writing samples. By combining NLP tasks with linguistic approaches in relation to AWLs, important advances can be achieved in the frame of lexical and syntactic analyses that evaluate the use of collocations and phraseology specific to the academic varieties. For the Romanian language, there have been few attempts to extract a valid Romanian Word List (Szabo, 2015) and only one study has extracted and analysed multiword units using academic writing corpora (Muresan et al., 2022).

In recent years, researchers have worked to create several academic writing corpora. EXPRES – Corpus of Expert Writing in Romanian and English (Chitez et al., 2022b) is one of them. It is the only bilingual multidisciplinary corpus capturing the Romanian academic writing context. By combining datasets representing the Romanian Frequency List (Szabo, 2015) based on the ROMBAC Corpus (Ion et al., 2012), and EXPRES disciplinary datasets (Chitez et al., 2022b), we were able to generate an empirically based Romanian Academic Word List. Ro-AWL is made publicly available¹ and can be used for teaching, text classification and language modelling.

2 Related Work

Most academic vocabulary lists have been developed in the context of English for Academic Pur-

¹<https://github.com/bucuram/Ro-AWL>

poses (EAP). On the whole, two categories of lists exist. One list type aims to identify academic words commonly used in EAP across disciplines, which students could be made aware of. The studies aiming to provide cross-disciplinary academic word lists usually use large corpora containing expert academic writing from various disciplines. The widely used lists of this type are the Academic Word List (AWL) (Coxhead, 2000) and the Academic Vocabulary List (AVL) (Gardner and Davies, 2014). The second type of list seeks to identify discipline or field-specific words worth teaching. Various specialised lists have been developed for fields such as veterinary medicine (Ohashi et al., 2020) or nursing (Yang, 2015).

While there is a growing interest in building cross-disciplinary academic word lists for languages other than English, these academic word lists remain few. See, for example studies conducted for French (Cobb and Horst, 2004), Persian (Rezvani et al., 2016), Portuguese (Baptista et al., 2010), Swedish (Carlund et al., 2012), and Norwegian (Johannessen et al., 2016). An explanation for this scarcity might be that academic language data sets are rare and often not freely available due to copyright. This can be especially true for low-resource languages, such as Romanian. Access to a representative corpus is crucial, as the validity and reliability of an academic word list highly depend on the quality of the data set.

Apart from the limited availability of academic writing corpora, an additional challenge may be that there is no standard procedure for extracting academic word lists. Scholars are still exploring and testing various methodologies. For example, some studies build on the methods used for the AWL or the AVL (Johannessen et al., 2016; Rezvani et al., 2016). One study uses the translated version of the AVL in Portuguese as a starting point for its investigation (Baptista et al., 2010). Another study proposes a new word list extraction method different from previous ones (Carlund et al., 2012).

In the case of Romanian, no previous studies have compiled specialised or general academic word lists. Although in the last 10-15 years, several research institutions and projects have been involved in developing corpus resources in Romanian, relatively few have focused exclusively on general academic writing. Some of the most significant corpora recently compiled, such as ROMBAC (Romanian Balanced Annotated Corpus, see Ion

et al. (2012)), with more than 30 million words, CoRoLa (Corpus of Contemporary Romanian Language, see Mititelu et al. (2014)), or The Balanced Romanian Corpus (BRC, see Midrigan-Ciochina et al. (2020)) cover only few disciplines or subsets: 5 sections for ROMBAC (journalism, literature, medical texts, legal texts, biographies), uneven and unfiltered distribution of resources in CoRoLa (the collection of academic writing texts is based on agreements with publishing houses and journals, without filtering of the content on quality criteria) and BRC (literary text samples, research articles, news, spoken data). The ROMBAC corpus (excluding the medical subcorpus) was already used to develop the Romanian Word List (RWL, see Szabo (2015)), targeted at Romanian L2 learners (e.g. from the Hungarian minority in Romania). The list is a general list of words, not focused on academic language. As far as discipline-specific corpora are concerned, smaller corpora such as SiMoNERo (medical corpus, Mitrofan et al. (2019)), BioRo (Mitrofan and Tufiş, 2018), PARSEME-Ro (news articles), LegalNERo (legal, Păiş et al. (2021)), MARCELL (legal, multilingual, see Váradi et al. (2020)), CURLICAT (multilingual, containing several domains: Economics, Education, Health, Sciences, etc., see Váradi et al. (2022)) have been compiled. However, apart from compiling the datasets and conducting a series of descriptive studies, no special attention is given to the lexical level.

In this context, the EXPRES corpus (Corpus of Expert Writing in Romanian and English) is the first corpus of discipline-specific academic writing in the Romanian context (academic writing in Romanian L1 and academic writing in English L2 produced by Romanians) (Bucur et al., 2022; Chitez et al., 2022a). Covering four disciplines – Linguistics, Economics, Political Sciences, Information Technology –, the Romanian subset contains 200 open-access research articles from each domain, published in the past 5-10 years in peer-reviewed journals (see Chitez et al. (2022b)). The rigorous selection criteria (Rogobete et al., 2021) contribute to the representativeness of the corpus, making it a suitable candidate for testing a possible Romanian Word List and narrowing it down to an Academic Word List. Furthermore, the EXPRES corpus is the first Romanian expert academic corpus available on an open-access query platform. Unlike other Romanian corpora, which offer limited access to third parties and poor resources for downloading search

results or statistics, the EXPRES corpus support platform has been implemented as a cross-platform distributed web application (Chitez et al., 2022b).

3 Data

This work uses two different corpora: the academic corpus EXPRES and the Romanian Academic Word List (Szabo, 2015) compiled from the general corpus ROMBAC. The Romanian language sub-corpus of EXPRES² (Chitez et al., 2022b) consists of 800 research articles, 200 articles for each of the four fields: Linguistics (LG), Economics (EC), Information Technology (IT) and Political Sciences (PS). The articles from the corpus were manually processed to preserve the anonymity of the authors (e.g., the name of the authors were replaced with AUTHOR_NAME) and the beginning and end of the title, abstract and sections are annotated with corresponding XML tags (e.g., <TITLE>, </TITLE>) (Chitez et al., 2022b). Table 1 shows the distribution of words in EXPRES, without counting the manually added tags. The corpus contains more than 3 million words, with more than 200 thousand unique words.

Domain	Tokens	Types
EC	1,092,846	48,807
LG	674,277	73,667
IT	750,236	40,494
PS	963,061	62,096
Total	3,480,420	225,064

Table 1: EXPRES Statistics

The Romanian Academic Word List (Szabo, 2015) contains a frequency list for all the words in the Romanian Balanced Annotated Corpus (ROMBAC) (Ion et al., 2012). ROMBAC (Ion et al., 2012) is a large general collection of texts from the Romanian language. It contains texts from five domains: news, medical, legal, biographies and fiction. The texts from ROMBAC are tokenized and lemmatized. The version we use in this paper contains more than 25 million lemmas, of which 1 million are unique (Table 2). The dataset was previously used to derive other linguistic resources, such as the Romanian Word List and Romanian Vocabulary Levels Test (Szabo, 2015). We use the ROMBAC corpus in our work because it is the largest corpus available in Romanian that was not web-scraped, and it is a reference corpus for the contemporary Romanian language (Ion et al.,

²<https://expres-corpus.org/>

2012). Even if another larger corpus for the contemporary Romanian language exists, namely CoRoLa (Mititelu et al., 2014), it is not publicly available and cannot be downloaded; it can only be queried online³. The other reference corpus recently compiled, BRC (Midrigan-Ciochina et al., 2020), was not an option either, since its size is smaller than ROMBAC and lacks disciplinary variation.

Domain	Tokens	Types
News	1,922,109	50,945
Medical	6,783,005	362,782
Legal	6,269,543	248,354
Biographies	3,716,031	223,592
Fiction	6,950,371	105,346
Total	25,641,059	991,019

Table 2: ROMBAC Statistics

4 Methodology

Data preprocessing. The Romanian Academic Word List, with words from the ROMBAC corpus, provides the lowercase lemma for each word in the corpus and its frequency. Therefore, no preprocessing step was done on this data. Even if we use the word frequencies from the Romanian Academic Word List, we will refer to this data as the ROMBAC corpus, given that the list contains all the words from ROMBAC.

The EXPRES corpus is organised in multiple .txt files, one for each article from the four domains LG, IT, PS, and EC. For each document, we removed specific tags used for article anonymisation, such as JOURNAL_TITLE, AUTHOR_NAME, etc., and the specific XML tags used to mark the beginning or end of the title (<TITLE>, </TITLE>), abstract (<ABS_INT>, </ABS_INT>), or different sections of the article (<INTROD>, </INTROD>), etc. The EXPRES corpus statistics regarding the words and word types in the corpora are shown in Table 1. For preprocessing the text, we used Stanza (Qi et al., 2020) for lemmatising and extracting part-of-speech tags. All the lemmas from the texts are transformed into lowercase. The Stanza toolkit was chosen for its good performance for the Romanian language, compared to other NLP tools (Paiş et al., 2021). However, we performed a manual analysis of the extracted lemmas and observed that some of them are incorrect: “sociales” instead of

³<https://korap.racai.ro/>

“social” (En: “social”), “europes” instead of “european” (En: “European”), and others. Even if previous works have shown a good performance of the Stanza toolkit for lemmatisation in the Romanian language (Paiş et al., 2021), we chose to use the lemmas from the ROMBAC corpora for the words that appear in ROMBAC. We used Stanza only for extracting the lemma of words that were not part of ROMBAC. This way, the noise of lemmatisation was diminished, as the lemmas provided in the ROMBAC corpus were accurate and have been previously validated (Ion et al., 2012).

Building the academic word list. For constructing the academic word list, we follow the methodology for building the Academic Vocabulary List for the English language (Gardner and Davies, 2014), comprising different frequency measures for lemmas. We chose to use the methodology from Gardner and Davies (2014) instead of the procedure from Coxhead (2000) because the former method provides an academic list with almost twice the latter’s coverage. The approach from Coxhead (2000) is based on word families, while the method from Gardner and Davies (2014) relies on lemmas. A word family is represented by the base word from which other words are derived with suffixes and prefixes. This can be problematic in the case of academic words, as the base of a word family can be an academic word, but their derivations might not be academic (Gardner and Davies, 2014).

The methodology is based on four measures: ratio, range, dispersion and discipline measure. The ratio is used to exclude general high-frequency words from the corpus, while the other three metrics exclude technical or discipline-specific terms. We further expand on each metric below.

Ratio. Similar to Gardner and Davies (2014), general high-frequency words (in our case, lemmas) are removed from the academic word list. The ratio is computed to keep in the list words with a higher frequency in the academic corpus than in the general non-academic corpus. We computed the normalised frequency per million words of each word in the two corpora, EXPRES and ROMBAC. The ratio is calculated by dividing the academic corpus’s normalised frequency by the general corpus’s normalised frequency for each word. Gardner and Davies (2014) use the frequency ratio of 1.5 in their method, but mention that the measure is not a gold standard. We experimented with values between 1.2 and 2.0 for ratio, and, in our case, the

1.2 ratio was a suitable value, to not have important academic words excluded from our list, such as “metodologic” (En: “methodological”), “clasificare” (En: “classification”), “activitate” (En: “activity”), “distinge” (En: “distinguish”), “sugera” (En: “suggest”), which are found in the original AVL for the English language.

Range. The range measure allows for selecting words that only occur in multiple disciplines, and filtering out discipline-specific words. Gardner and Davies (2014) proposed that a word should have at least 20% of the expected frequency in 78% of the sub-corpora (i.e. 7 out of 9 domains). For computing the expected frequency, we first calculated each word’s frequency in relation to the corpus by dividing the word count by the total number of words in EXPRES. Afterwards, the frequency in relation to the corpus is multiplied by the number of words in a given sub-corpora to get the expected frequency in each sub-corpora. In our case, EXPRES has only four domains, and we chose words that had at least 20% of the expected frequency in at least three out of four fields, corresponding to 75% of sub-corpora.

Dispersion. The measure used for dispersion is Julliland’s D (Juilland and Chang-Rodríguez, 1964), which shows how evenly a word appears in a corpus. The formula is as follows:

$$Juilland's D = 1 - \frac{\sigma}{\bar{x}} \times \frac{1}{\sqrt{n-1}}$$

where σ represents the standard deviation and \bar{x} represents the mean frequency of a word. n is the number of sub-corpora.

The values of dispersion range from 0.01 (corresponding to words that appear in a small part of the corpus) to 1.00 (meaning that a word is spread evenly in the corpus). Unlike the range measure, which estimates if a word has the expected frequency in the four domains, the dispersion measure ensures that a given word is distributed uniformly in the four sub-corpora. Gardner and Davies (2014) chose 0.80 dispersion, while, in other works, the dispersion measure varies between 0.30 to 0.60 (Oakes and Farrow, 2006; Johannessen et al., 2016; Lei and Liu, 2016). We decided to use a dispersion value of 0.50 in our work.

Discipline measure. This measure is used for filtering out words with a very high frequency in a given domain, which may be technical discipline-specific words. Gardner and Davies (2014) proposed that a word cannot have more than three

times the expected frequency in any domain. Following a similar approach, we remove words with more than three times the expected frequency in any of the four domains.

As an additional measure, we excluded words with low frequency in the academic corpus, because the metrics mentioned above do not filter them out. Inspired by [Coxhead \(2000\)](#) and [Lei and Liu \(2016\)](#), we remove from the final academic list the words that have a minimum frequency of 28.57 per million words, corresponding to the minimum frequency originally proposed by [Coxhead \(2000\)](#) of 100 times in the 3.5 million words corpus they used in their work. We also performed a manual analysis of the academic word list and removed the noise, such as proper nouns (e.g., “București”, En: “Bucharest”), some numerals, and some words that were not academic and that were not filtered out by the measures mentioned above.

5 Results

The final Romanian academic word list consists of 673 lemmas with their corresponding part of speech tags. The list comprises 332 nouns, 167 adjectives, 157 verbs, and 17 adverbs. We automatically translated into Romanian the words from the AVL developed for English ([Gardner and Davies, 2014](#)) that contains 3015 words. We found that 381 words in our list are in the original AVL. There are some cases of academic words found in our Romanian academic list and in the AVL for English for which the automatic translation fails to provide the correct match. For example, the noun “adoption” from AVL was translated as “adopție”, which is not in the Ro-AWL, but the word “adoptare” is an academic word from Ro-AWL with the same meaning. The fact that we found more than half of the Ro-AWL in the original AVL, even though in some cases the translation fails to capture the correct meaning of the words, makes us confident that the measures used are reliable for building a Romanian academic word list.

In line with previous works ([Gardner and Davies, 2014](#); [Coxhead, 2000](#); [Carlund et al., 2012](#)), to demonstrate the viability of the newly developed academic word list, we measured the coverage of the Ro-AWL in two corpora: the academic corpus EXPRES and in the general corpora ROMBAC. The academic words from our list cover 15.25% of the EXPRES corpus and 6.73% of ROMBAC. In line with the English AVL results, Ro-AWL has

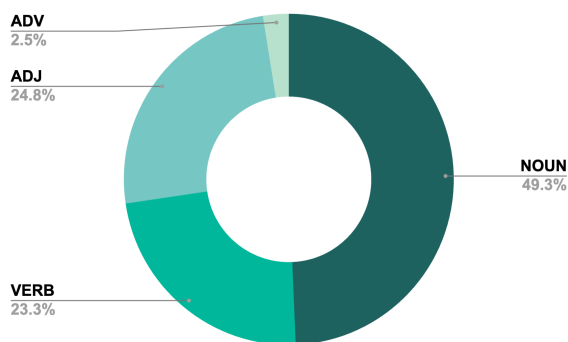


Figure 1: The distribution of the words in terms of part-of-speech from Ro-AWL

a higher coverage in the academic corpus and a lower coverage in the general corpus. Regarding the coverage in EXPRES, we show the coverage of academic words categorised by their part-of-speech tags in Table 3. The coverage of the Romanian academic word list varies in the four domains. The coverage is 17.75% for the Economics sub-corpora, 11.82% for Linguistics, 17.03% for Information Technology and 13.17% for Political Sciences.

	EC	LG	IT	PS
VERB	4.98%	3.95%	5.33%	3.95%
NOUN	9.74%	6.02%	9.20%	6.82%
ADJ	0.33%	0.27%	0.24%	0.16%
ADV	2.70%	1.59%	2.26%	2.24%
Total	17.75%	11.82%	17.03%	13.17%

Table 3: Coverage of Ro-AWL in the EXPRES corpus

6 Discussion

A first observation concerns the different coverages of Ro-AWL in the EXPRES corpus (see Table 3). The lower percentages in Linguistics and Political Sciences (with a total coverage ranging between 11% and 14%) and the higher ones in Economics and IT confirm that “The SSH community is characterised by the embedment of research in the local context and by linguistic diversity in producing and disseminating knowledge” ([Kancewicz-Hoffman and Pölönen, 2020](#)). Researchers in the Romanian context in SSH (Social Sciences and Humanities) tend to favour a more “creative” dimension of the language used in academic writing, using figurative language in constructing rhetorical structures. Although in English language academic writing “the dichotomy of soft and hard sciences is rather fluid and as such insignificant” ([Stotesbury, 2003](#)), discipline-specific peer-review practice in the Ro-

manian setting seems to influence the academic writing style. This is particularly visible in the EXPRES subset of Political Sciences and Linguistics. Romanian academic writing in SSH seems rather unfocused, descriptive and rich in rhetorical structures. In contrast, research articles in Economics and Information Technology contain many statistics, tables, and formulas, making the writing in the discipline less descriptive.

Secondly, although our extraction measures were successful in filtering most of the technical vocabulary, small amount of technical language remains in the Ro-AWL (terms such as “dauna”, En: “damage” - in contexts related to insurances; “institutional”, “security”, “electronic” etc.). Nevertheless, the majority of the Ro-AWL components are discipline neutral, thus contributing to academic discourse cohesion and coherence.

Thirdly, a technical challenge regarding the functionality and accuracy of the Romanian POS tagger should be mentioned. An overview of the assigned tags revealed the difficulty of the tagger to distinguish between adjectives and adverbs (for instance: “important”, “social”, “european” were assigned as adverbs, but the contexts prove their prevalent use as adjectives). It also confused past participles ending with “-t” (e.g. “accentuat”, En: “emphasised”). This technical difficulty can be observed in Table 3, with the coverage of adverbs being higher than the one of adjectives, because most of the adjectives had the part-of-speech mislabeled by the POS tagger. These errors of the POS tagger are due to the homonymy between the two POS, most adverbs being homonymous to their adjective counterparts (Vasile and Croitor, 2017).

A technical advantage of the Romanian POS tagger, however, is its capacity to recognise nouns with a definite article while being a part of prepositional phrases (“în pofida”, En: “despite”, “în jurul”, En: “around”). This also explains the increased percentage levels of nouns, adverbs and verbs and the lower values for adjectives (see Figure 1).

Despite some of the technical challenges, the extraction of the Romanian AWL using the EXPRES corpus resulted in successfully identifying the recurrent discourse conventions used by Romanian researchers. During the process and alongside the extraction procedure per se, translating the Academic Vocabulary List (AVL) (Gardner and Davies, 2014) was a helpful procedure, as it is well accepted that academic writing, irrespective

of the language, contains a large number of words of Greek and Latin origin (see e.g., Rasinski et al. (2008); Green (2015)).

7 Conclusions and Future Work

This study reports the extraction of the first Romanian Academic Word List (Ro-AWL), which can be used to check the degree of academic vocabulary coverage in discipline-specific and general language samples. Ro-AWL consists of 673 lemmas, distributed among the main part-of-speech categories (nouns, verbs, adverbs, adjectives). Our methodology adopted measures used for the Academic Vocabulary List for the English language, such as ratio, range, dispersion and discipline measures. The percentages calculated by testing Ro-AWL on the disciplinary datasets in the EXPRES corpus (Chitez et al., 2022b), indicate a lower coverage for Linguistics (11.82%) and Political Sciences (13.17%) and a higher coverage for Information Technology (17.03%) and Economics (17.75%). Also, the academic vocabulary coverage in ROMBAC, a general language reference corpus, is 6.73%, while the coverage is much higher (15.25%) in EXPRES, an expert academic writing corpus. This aligns with previous research, since Ro-AWL coverage is similar to thresholds for academic vocabulary (Nation, 2001).

Despite several computation constraints (e.g. Romanian POS tagger not being able to distinguish between adjectives and adverbs), our study provides important insights into the academic writing vocabulary in Romanian by proposing a validated Romanian Academic Word List. Our findings also have pedagogical implications, as the list can be used to support academic writing teaching activities and NLP tasks focusing on Romanian. For example, the Ro-AWL can be paired up with the freely available EXPRES corpus platform to develop corpus-assisted learning activities commonly known as Data-Driven Learning (DDL) (see e.g., Bennett (2010)). However, even if the coverage test results in the EXPRES are encouraging, further research is needed to test the validity of the Ro-AWL on corpora containing academic writing from more disciplines. Future work can be conducted in at least two directions: refining the lists from a contrastive perspective, by developing a discipline-specific AWL, or, on the contrary, by searching for highly frequent academic words present in an extended corpus containing more disciplines.

Acknowledgements

We would like to thank Csaba Z Szabo for giving us access to the Romanian Frequency List and Ștefan Daniel Dumitrescu for his valuable insights into the data and methods used in this paper. This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI – UEFISCDI, project number 158/2021, within PNCDI III, awarded to Dr Habil Madalina Chitez (PI), from the West University of Timisoara, Romania, for the project DACRE (*Discipline-specific expert academic writing in Romanian and English: corpus-based contrastive analysis models*, 2021-2022).

References

- Jorge Baptista, Neuza Costa, Joaquim Guerra, Marcos Zampieri, Maria Cabral, and Nuno Mamede. 2010. P-AWL: academic word list for Portuguese. In *Computational Processing of the Portuguese Language: 9th International Conference, PROPOR 2010, Porto Alegre, RS, Brazil, April 27-30, 2010. Proceedings 9*, pages 120–123. Springer.
- Gena R Bennett. 2010. Using corpora in the language learning classroom: Corpus linguistics for teachers. *University of Michigan*.
- Ana-Maria Bucur, Mădălina Chitez, Valentina Muresan, Andreea Dincă, and Roxana Rogobete. 2022. Express corpus for a field-specific automated exploratory study of 12 english expert scientific writing. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4739–4746.
- Carina Carlund, Håkan Jansson, Sofie Johansson Kokkinakis, Julia Prentice, and Judy Ribbeck. 2012. An academic word list for Swedish—a support for language learners in higher education. In *Proceedings of the SLTC 2012 workshop on NLP for CALL; Lund; 25th October; 2012*, 080, pages 20–27. Linköping University Electronic Press.
- Madalina Chitez, Valentina Mureșan, and Roxana Rogobete. 2022a. How to write good academic papers: using the EXPRES corpus to extract expert writing linguistic patterns. In *Conference Proceedings. The Future of Education 2022*.
- Mădălina Chitez, Roxana Rogobete, Valentina Muresan, and Andreea Dincă. 2022b. Corpus of Expert Writing in Romanian and English (EXPRES). In *West University of Timisoara*. <https://expres-corpus.org/>.
- Tom Cobb and Marlise Horst. 2004. Is there room for an academic word list in French. *Vocabulary in a second language*, 23:15–38.
- Averil Coxhead. 2000. A new academic word list. *TESOL quarterly*, 34(2):213–238.
- Dee Gardner and Mark Davies. 2014. A new academic vocabulary list. *Applied linguistics*, 35(3):305–327.
- Tamara M Green. 2015. *The Greek & Latin Roots of English*. Rowman & Littlefield Publishers.
- Radu Ion, Elena Irimia, Dan Stefanescu, and Dan Tufis. 2012. ROMBAC: The Romanian Balanced Annotated Corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 339–344.
- Janne M Johannessen, Arash Saidi, and Kristin Hagen. 2016. Constructing a Norwegian academic wordlist. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1457–1462.
- Alphonse Juilland and Eugenio Chang-Rodríguez. 1964. *Frequency dictionary of Spanish words*. De Gruyter Mouton.
- Nina Kancewicz-Hoffman and Janne Pölönen. 2020. Does excellence have to be in English? Language diversity and internationalisation in SSH research evaluation. *Overview of Peer Review Practices in the SSH*, pages 32–41.
- Lei Lei and Dilin Liu. 2016. A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for academic purposes*, 22:42–53.
- Ludmila Midrigan-Ciochina, Victoria Boyd, Lucila Sanchez-Ortega, Diana Malancea Malac, Doina Midrigan, and David P Corina. 2020. Resources in Underrepresented Languages: Building a Representative Romanian Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3291–3296.
- Verginica Barbu Mititelu, Elena Irimia, and Dan Tufis. 2014. CoRoLa—The Reference Corpus of Contemporary Romanian Language. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1235–1239.
- Maria Mitrofan, Verginica Barbu Mititelu, and Grigoriina Mitrofan. 2019. Monero: a biomedical gold standard corpus for the Romanian language. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 71–79.
- Maria Mitrofan and Dan Tufiș. 2018. Bioro: The biomedical corpus for the romanian language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Akira Murakami, Paul Thompson, Susan Hunston, and Dominik Vajn. 2017. ‘What is this corpus about?’: using topic modelling to explore a specialised corpus. *Corpora*, 12(2):243–277.

- Valentina Muresan, Roxana Rogobete, Ana-Maria Bucur, Madalina Chitez, and Andreea Dincă. 2022. *Phraseology in Romanian Academic Writing: Corpus Based Explorations into Field-Specific Multiword Units*. Peter Lang. D. Anca, M. Chitez, L. Dinu and M. Dobre (Eds.), *Recent Advances in Digital Humanities*. Romance Language Applications. Peter Lang Verlag.
- Ian SP Nation. 2001. *Learning vocabulary in another language*, volume 10. Cambridge University Press Cambridge.
- Michael P Oakes and Malcolm Farrow. 2006. Use of the chi-squared test to examine vocabulary differences in English language corpora representing seven different countries. *Literary and linguistic computing*, 22(1):85–99.
- Yukiko Ohashi, Noriaki Katagiri, Katsutoshi Oka, and Michiko Hanada. 2020. ESP corpus design: compilation of the Veterinary Nursing Medical Chart Corpus and the Veterinary Nursing Wordlist. *Corpora*, 15(2):125–140.
- Vasile Pais, Radu Ion, Andrei-Marius Avram, Maria Mitrofan, and Dan Tufis. 2021. In-depth evaluation of Romanian natural language processing pipelines. *Romanian Journal of Information Science and Technology (ROMJIST)*, 24(4):384–401.
- Vasile Păiș, Maria Mitrofan, Carol Luca Gasan, Vlad Coneschi, and Alexandru Ianov. 2021. Named entity recognition in the Romanian legal domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 9–18.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.
- Timothy Rasinski, Nancy Padak, Rick M Newton, and Evangeline Newton. 2008. *Greek and Latin roots: Keys to building vocabulary*. Shell Education.
- Reza Rezvani, Abbas Gholtash, and Gerannaz Zamani. 2016. The First Corpus-Based Persian Academic Word List: Development and Pedagogical Implications. *Journal of Teaching Persian to Speakers of Other Languages*.
- Roxana Rogobete, Mădălina Chitez, Valentina Mureșan, Bogdan Damian, Adrian Duciuc, Claudiu Gherasim, and Ana-Maria Bucur. 2021. Challenges in compiling expert corpora for academic writing support. In *Conference Proceedings. The Future of Education 2021*.
- Hilkka Stotesbury. 2003. Evaluation in research article abstracts in the narrative and hard sciences. *Journal of English for Academic Purposes*, 2(4):327–341.
- Cz Szabo. 2015. Introducing a Romanian frequency list and the Romanian vocabulary levels test. *Current Issues in Linguistic Variation: The 14th international conference of the Department of Linguistics*, 2.
- Tamás Váradi, Svetla Koeva, Martin Yamalov, Marko Tadić, Bálint Sass, Bartłomiej Nitoń, Maciej Ogrodniczuk, Piotr Pezik, Verginica Barbu Mititelu, Radu Ion, et al. 2020. The MARCELL legislative corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3761–3768.
- Tamás Váradi, Bence Nyéki, Svetla Koeva, Marko Tadić, Vanja Štefanec, Maciej Ogrodniczuk, Bartłomiej Nitoń, Piotr Pezik, Verginica Barbu Mititelu, Elena Irimia, et al. 2022. Introducing the CURLICAT corpora: seven-language domain specific annotated corpora from curated sources. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 100–108.
- Carmen Mîrzea Vasile and Blanca Croitor. 2017. Properties of romanian adverbs and adjectives from a categorial status perspective. *Adjective adverb interfaces in Romance*, 242:227.
- Thinley Wangdi and Ringphami Shimray. 2022. Investigating the Significance of Coxhead’s Academic Word List for Self-Access Learners. *Studies in Self-Access Learning Journal*, 13(3).
- Ming-Nuan Yang. 2015. A nursing academic word list. *English for specific purposes*, 37:27–38.
- Marcos Zampieri. 2012. Evaluating knowledge-poor and knowledge-rich features in automatic classification: A case study in WSD. In *2012 IEEE 13th International Symposium on Computational Intelligence and Informatics (CINTI)*, pages 359–363. IEEE.