
Corpus Building and LLM Benchmarking for Language Understanding

Building a cross-lingual corpus of BA theses: Rationale and preliminary observations

Karla Csuros
West University of
Timisoara, Romania
karla.csuros@
e-uvvt.ro

Madalina Chitez
West University of
Timisoara, Romania
madalina.chitez@
e-uvvt.ro

Roxana Rogobete
West University of
Timisoara, Romania
roxana.rogobete@
e-uvvt.ro

Petya Osenova
Institute of Information and Communication Technologies
Bulgarian Academy of Sciences, Bulgaria
petya@bultreebank.org

Abstract

In recent years, the impact of AI has increased in all areas of our lives, including academic writing. While the writing of MA and doctoral theses had been researched extensively even prior to AI, BA theses were not a focus of analysis. However, BA theses are the first serious academic texts students produce, in which they must demonstrate their ability to logically construct academic texts, posit research questions, critically review related works, classify and analyze data, and make observations and conclusions. The growing use of these tools makes it urgent to establish empirical baselines for what student writing looks like in specific languages and genres, against which future change can be measured. In this paper, we report on the creation of an initial cross-lingual corpus of theses from Bulgarian and Romanian BA students and present some quantitative and qualitative results on the usage of hedges in both corpora. We are interested in seeing how certain or uncertain students are with respect to the various parts of academic texts.

Keywords: academic writing, BA theses, cross-lingual corpus, AI in education, corpus linguistics.

1 Introduction

Academic writing research has devoted substantial attention to MA dissertations and doctoral theses (see, for instance, Charles, 2007; Thompson, 2005), yet BA theses, as the first extended academic texts students produce, remain comparatively under-studied, particularly outside English-language contexts. This gap has acquired new urgency with the widespread availability of genera-

tive AI (genAI) tools. If we wish to understand how AI adoption is reshaping student writing, we need empirical baselines for what student writing looks like in specific languages, genres, and institutional settings.

This paper contributes to that effort by reporting on the construction of a pilot cross-lingual corpus of 36 BA theses (18 Romanian, 18 Bulgarian) from the 2025 cohort, formed of students who had access to large language models (LLMs) throughout their undergraduate studies. As a first analytical lens, we examine hedging: the linguistic devices through which writers signal uncertainty, attribute claims, and modulate the strength of their assertions. Hedging is a particularly sensitive indicator of epistemological development, as it reflects how students position themselves relative to knowledge claims in their discipline. To this end, we address two research questions:

(RQ1) What hedging profiles characterise BA writing in Romanian and Bulgarian, and how do they compare cross-linguistically?

(RQ2) What can these profiles tell us about the rhetorical strategies available to novice academic writers in each language?

2 Related work

2.1 Student BA thesis writing

Corpus-based approaches to academic writing have developed substantially since the shift from text-oriented to process-centred pedagogy in the 1970s, yet most corpus resources and pedagogical applications target MA and doctoral writing (Chitez et al., 2015; Flowerdew, 2015). Platforms such as *The-*

sis Writer provide real-time formulation support for graduate students through POS-informed pattern searches and discipline-specific phrase lists (Chitez et al., 2015), while workshop-based approaches use broad reference corpora to teach prototypical move structures at the postgraduate level (Flowerdew, 2015). These initiatives have demonstrated the value of Data-Driven Learning (DDL) for developing metadiscursive awareness, but their focus remains overwhelmingly on graduate genres. BA theses, by contrast, have received limited corpus-based attention, a gap that is particularly consequential for non-Anglophone academic traditions. Baniceru et al. (2012) offer one of the first studies of Romanian BA writing, finding that students successfully adopt descriptive rhetorical moves (introducing the topic, presenting structure) but struggle with reflective ones (summarising prior research, identifying gaps). This observation can be situated within what Doroholschi et al. (2018) explore as a broader Central and Eastern European context where academic writing cultures shaped by French-influenced, author-oriented traditions are undergoing a sometimes mechanical transition toward Anglo-Saxon reader-oriented conventions. The present study contributes to filling this gap by building a cross-lingual BA thesis corpus and applying corpus methods, specifically lexicon-based hedge identification, to a genre and pair of languages for which no comparable quantitative data has previously been reported.

More recent research has benefited from the development of the Corpus of Romanian Academic Genres (ROGER), a bilingual learner corpus containing undergraduate academic texts written in Romanian (L1) and English (L2). The ROGER corpus (see Chitez 2022) includes BA-level research papers and theses across multiple disciplines (Economics, Humanities, Political Sciences, Computer Science, etc.) and allows, together with EXPRES corpus (Corpus of Expert Writing in Romanian and English; Chitez et al. (2022), comparisons between novice student writing and expert academic discourse. Corpus-based analyses using ROGER have documented limited phraseological diversity, underuse of stance markers, and restricted metadiscursive resources in Romanian student writing, particularly at the undergraduate level (Oravițan et al. 2022, Dincă et al.2022).

2.2 Hedging in academic discourse

Hedging is a fundamental communicative strategy through which academic writers balance "conviction with caution" in what Hyland (1998) characterises as the social negotiation of knowledge. Rather than signalling weakness, hedges serve both epistemic and interpersonal functions, encoding the writer's degree of commitment to a proposition while maintaining collegial deference toward the reader (Hyland, 1998). Unhedged claims risk appearing as overstatements that threaten the writer's standing within a disciplinary community.

Several classification systems have been proposed for these devices. Prince et al. (1982) distinguish between approximators (affecting propositional content) and shields (affecting speaker commitment). Salager-Meyer (1994) and Crompton (1997) offer pragmatically grounded typologies, while Caffi (1999) proposes a tripartite model of mitigation. Hyland (1998) further distinguishes standard hedges, which qualify authorial commitment, from attribute hedges, which manage "permissible imprecision" without necessarily signalling uncertainty. Thompson and Yiyun (1991) complement this with a multi-layered taxonomy of reporting verbs, distinguishing author acts, writer acts, and stance categories, a system which is particularly relevant to the categories used in the present study.

Cross-linguistically, Vassileva (2001) demonstrates that hedging conventions differ substantially between academic traditions, while Vold (2006) shows that even typologically related languages can produce divergent hedging profiles in comparable academic genres. These findings motivate the present cross-lingual design for investigating hedge use in student BA thesis writing.

3 Corpus and methodology

3.1 TRAI cross-lingual BA thesis corpus

The pilot corpus compiled for this study belongs to the TRAI (Exploring the Triad of Academic Writing, Critical Thinking and AI Literacy: Interdisciplinary Perspectives on Undergraduate Thesis Writing) project and comprises 36 BA theses in two languages: 18 Romanian theses from the West University of Timișoara (412,654 words), and 18 Bulgarian theses from Sofia University St. Kliment Ohridski (369,721 words). The pairing of Romanian and Bulgarian is motivated first practically: both subcorpora are compiled within the TRAI

project, a partnership between the two institutions, which provides access to comparable BA theses from philology programmes in each country. Typologically, the pair also sets a Romance language against a Slavic one while holding constant their shared membership in the Balkan area, allowing hedging strategies to be compared across genetically distinct but areally linked systems. All theses were submitted by cohorts with access to generative AI tools throughout their undergraduate studies, situating the corpus in the post-LLM period. Because the design includes no pre-LLM comparison cohort, we make no claim that the profiles reported here reflect AI influence; that question is left to future work (Section 5). Table 1 summarises the composition and size of the two subcorpora alongside the lexicon specifications and aggregate hedging rates.

The two subcorpora were compiled as structured XML using a custom Python pipeline that identifies front matter, table of contents, body sections, and bibliography through heading-style detection and keyword matching. Both subcorpora share the same target structure: an XML document with elements segmenting each thesis into title page, table of contents, introduction, body, conclusion, and bibliography. For hedge analysis, only the content sections are retained; front matter and bibliography are excluded.

3.2 Cross-lingual hedge identification

Hedging devices are identified using a lexicon-based approach that draws on established hedging taxonomies (Caffi, 1999; Crompton, 1997; Hyland, 1998, 2005; Prince et al., 1982; Salager-Meyer, 1994; Thompson and Yiyun, 1991), with cross-linguistic adaptations informed by Vassileva (2001) and Vold (2006). Six categories are used for this pilot study: MODAL (epistemic modal verbs and particles), EPISTEMIC (lexical verbs encoding authorial stance), APPROXIMATOR (degree, quantity, and frequency modifiers), SHIELD (impersonal and distancing constructions), ADVERB (epistemic adverbs, concessives, and boosters), and DEFERENCE (attribution and perspectivisation markers).

For each language, a dedicated lexicon of regular expression patterns was developed: 132 patterns for Romanian and 159 for Bulgarian, with each pattern mapped to a canonical form and category. Each language's lexicon was compiled by a single expert linguist working from a shared English tax-

onomy: the English categories provided a common scaffold, which was then translated and enriched to fit the morphosyntactic resources of the target language. For instance, Romanian gerund forms (*considerând*, *observând*) and Bulgarian impersonal reflexive constructions (смята се, предполага се) have no direct English equivalents, but serve analogous hedging functions. Because the two experts worked from the same category definitions, the lexicons remain comparable across languages, and because a single annotator built each, there is no within-language disagreement to reconcile; every pattern is mapped to a canonical form and category and is released with the corpus for inspection. We note as a limitation that formal inter-annotator agreement and manual validation of the automatic annotation output were not carried out for this pilot; both, alongside the morphological analysis flagged below, are priorities for the next stage.

Pattern matching operates on the raw running text within each XML file, with overlapping matches resolved through a longest-match-first deduplication strategy. This surface-level approach prioritises transparency and reproducibility over the precision gains that POS tagging or lemmatisation could offer. Known limitations include the polysemy of high-frequency items, such as Romanian *poate* (modal verb "can" vs. adverb "perhaps") and Bulgarian *около* (approximator "approximately" vs. spatial preposition "around"), and the exclusion of the Bulgarian renarrative forms (преизказни форми), a grammaticalised evidential with no Romanian equivalent whose forms are homophonous with past active participles and therefore undetectable by regex. Both limitations are flagged for future work involving morphological analysis.

4 Preliminary results

Across the combined corpus of 36 theses ($\approx 782,000$ words), we identified 5,156 hedge instances in Romanian (mean 12.5/1000w, SD = 2.8) and 4,986 in Bulgarian (mean 14.18/1000w, SD = 3.71). Bulgarian theses thus hedge at a moderately higher overall rate, but the more revealing finding lies in how the two student populations hedge in the distribution across categories rather than in the aggregate frequency. Figure 1 visualises the distribution of hedges across the six categories, revealing markedly different profiles despite comparable overall rates.

	Romanian	Bulgarian
Institution	West Univ. Timisoara	Sofia Univ. St. Kliment Ohridski
Specialisation	Romanian Philology	Slavic Philology
Year	2025	2025
Theses	18	18
Total words	412,654	369,721
Mean words/thesis (SD)	22,925 (7,049)	20,540 (8,559)
Total hedges	5,156	4,986
Mean hedges/1000w (SD)	12.5 (2.80)	14.18 (3.71)
Lexicon patterns	132	159

Table 1: Pilot TRAI cross-lingual BA thesis corpus

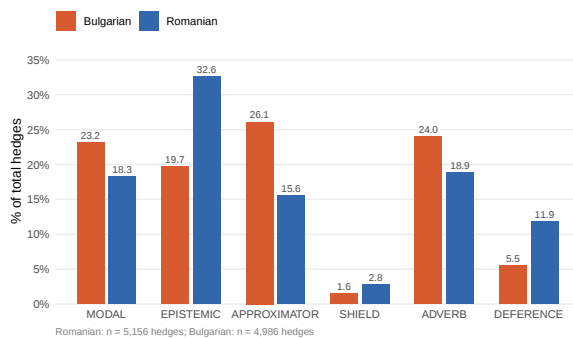


Figure 1: Hedging category profiles: Romanian vs. Bulgarian BA theses (% of total hedges, 18 theses per language).

To test whether these differences are statistically reliable, we treated each thesis as a single observation ($n = 18$ per language) and compared per-thesis hedging rates with Mann-Whitney U tests, applying Holm correction across the six categories and reporting Cliff's delta (δ) as an effect size. The aggregate hedging rate does not differ significantly between the two subcorpora (Romanian 12.5 vs. Bulgarian 14.18 per 1000 words; $U = 123$, $p = .22$), confirming that the contrast lies in the distribution rather than the overall frequency. Three categories differ significantly after correction, all with large effect sizes: Romanian shows higher rates of EPISTEMIC verbs ($p = .04$, $\delta = 0.51$) and DEFERENCE markers ($p = .002$, $\delta = 0.69$), while Bulgarian shows markedly higher use of APPROXIMATOR ($p < .001$, $\delta = 0.77$). MODAL and ADVERB use trend higher in Bulgarian but do not survive correction ($p = .08$ and $.13$), and SHIELD use is low in both subcorpora and does not differ significantly ($p = .13$).

Romanian students rely heavily on EPISTEMIC verbs, which account for 32% of all hedges (4.11/1000w). The most frequent epistemic forms (*observă, consideră, sugerează, propune*) are re-

porting verbs through which writers attribute claims to cited scholars or signal their own interpretive stance. This pattern is reinforced by the DEFERENCE category, where Romanian rates are nearly double those of Bulgarian (1.45 vs. 0.77/1000w), driven by high-frequency attribution markers such as *din punct de vedere*, *după cum*, and *conform*. Together, EPISTEMIC and DEFERENCE account for over 44% of Romanian hedging, suggesting a rhetorical strategy oriented around positioning claims relative to external authorities. For instance, a typical Romanian hedging sequence reads: "Această emoție *pare a fi* apogeul existenței" ("This emotion *appears to be* the apex of existence"), where the evidential *pare a fi* simultaneously claims and distances. Similarly, attribution to authorities is pervasive: "Pe de altă parte, *conform lui* Albert Camus, individul ce recurge..." ("On the other hand, *according to* Albert Camus, the individual who resorts to..."), combining DEFERENCE (*conform*) with a perspectival frame that defers the claim.

The Bulgarian profile is structurally different. Hedging is more evenly distributed across APPROXIMATOR (26.1%, 3.75/1000w), ADVERB (24%, 3.34/1000w), and MODAL (23.2%, 3.35/1000w). The most frequent forms (*често, обаче, може да + V, може би, почти*) are degree modifiers, possibility markers, and concessive connectors. Bulgarian students appear to hedge primarily by softening the strength or scope of their propositions rather than by delegating epistemic responsibility to cited sources. A representative Bulgarian passage illustrates this strategy: "Оптимизмът може да се разглежда като позитивна черта" ("Optimism *can be* viewed as a positive trait"), where a MODAL (*може да*) combines with a reflexive impersonal (*се разглежда*) to double-hedge a proposition that an English or Romanian

student might assert directly. Elsewhere, frequency-based softening is common: "като често пренебрегват неблагоприятните обстоятелства" ("as they *often* disregard unfavourable circumstances"), where често ("often") limits the scope of the generalisation.

SHIELD devices (impersonal constructions such as *se poate afirma* or *може да се каже*) are rare in both subcorpora (2.8% in Romanian, 1.6% in Bulgarian), and, as the tests above show, do not differ significantly between them. The few SHIELD instances that do occur tend to be formulaic: Romanian "Generația 60 *se poate spune* că a urmat..." ("The 60s generation, *it can be said*, followed...") and Bulgarian "може да се допусне, че социалното време..." ("*it can be assumed* that social time...") both use fixed constructions that suggest imitation of a template rather than productive mastery of impersonal distancing.

These observations should be interpreted with the caveats noted earlier: the lexicon-based approach cannot fully resolve polysemy, and the Bulgarian renarrative mood, a grammaticalised evidential with no Romanian equivalent, remains outside the scope of automatic detection.

5 Discussion and next steps

The divergent hedging profiles suggest distinct rhetorical orientations at the BA level: Romanian students manage uncertainty primarily through attribution, positioning claims relative to cited authorities, while Bulgarian students distribute hedging more evenly across modal, approximative, and adverbial resources, softening propositions lexically rather than delegating epistemic responsibility. SHIELD devices are rare in both subcorpora, and the tests confirm the difference between languages is not significant. We resist reading this as evidence of a developmental stage, however: that interpretation would require comparison with expert or graduate writing in the same languages, which this pilot does not include. We therefore offer it as a hypothesis for future work, in which BA profiles would be compared against MA, doctoral, or expert reference corpora (for Romanian, for instance, EXPRES). These patterns are consistent with the novice writing profiles described by Thompson and Yiyun (1991) and with the difficulties in reflective rhetorical moves observed in Romanian BA writing by Baniceru et al. (2012).

A fuller account of why the two groups differ is

beyond a pilot, but one plausible factor is writing tradition. The Romanian attribution-heavy profile in particular is consistent with the author-oriented tradition Doroholschi et al. (2018) describe for the region, in which claims are habitually anchored to named authorities. Social and pedagogical background are variables worth incorporating in future comparison.

Because the present corpus captures a post-LLM cohort, the hedging profiles reported here cannot be attributed to AI in the absence of a pre-LLM baseline; disentangling any AI effect would require comparison across cohort years, which we leave to future work. More immediate next steps within the TRAI project include adding Swiss German BA theses as a third comparison language, integrating POS tagging to resolve polysemy in high-frequency items, and expanding the corpus beyond philology programmes to test whether the patterns observed here are discipline-specific or generalisable.

Acknowledgements

This work was supported by the Swiss National Science Foundation (SNSF) under the Multilateral Academic Projects (MAPS) grant (No. IZ11Z0_230816), awarded for the project "Exploring the Triad of Academic Writing, Critical Thinking and AI Literacy: Interdisciplinary Perspectives on Undergraduate Thesis Writing" (TRAJ), co-financed by the Romanian Ministry of Education and Research through UEFISCDI (Grant No. 15ROCH /01/07/2025) and by the Bulgarian Ministry of Education and Science (Reg N D03-34/30.07.2025).

The author from ICT-BAS gratefully acknowledges the financial and institutional support provided by the Swiss-Bulgarian Research Programme in the framework of the second Swiss Contribution to selected EU member states.

References

- Cristina Ana Baniceru, Mirela Ioana Borchin, Claudia Ioana Doroholschi, and Dumitru Tucan. 2012. Academic writing in Romania: a contrastive analysis of BA thesis introductions in Romanian and English. *CICCRE*, 1.
- Claudia Caffi. 1999. [On mitigation](#). *Journal of Pragmatics*, 31(7):881–909.
- Maggie Charles. 2007. [Reconciling top-down and bottom-up approaches to graduate writing: Using](#)

- a corpus to teach rhetorical functions. *Journal of English for Academic Purposes*, 6(4):289–302.
- Mădălina Chitez. 2022. **Corpus of Romanian Academic Genres (ROGER)**. Slovenian Language Resource Repository CLARIN.SI.
- Madalina Chitez, Christian Rapp, and Otto Kruse. 2015. **Corpus-supported academic writing : how can technology help?** *Critical CALL : Proceedings of the 2015 EUROCALL Conference, Padova, Italy*.
- Mădălina Chitez, Roxana Rogobete, Valentina Mureșan, and Andreea Dincă. 2022. **Corpus of Expert Writing in Romanian and English (EXPRES)**. West University of Timișoara.
- Peter Crompton. 1997. **Hedging in academic writing: Some theoretical problems**. *English for Specific Purposes*, 16(4):271–287.
- Andreea Dincă, Mădălina Chitez, Loredana Bercuci, Alexandru Oravițan, and Roxana Rogobete. 2022. **Using bilingual novice and expert corpora to teach academic writing at the university: The case of ROGER and EXPRES**. Paper presented at the 15th Teaching and Learning Corpora Conference (TaLC 2022). Conference presentation, 13–15 July 2022.
- Claudia Ioana Doroholschi, Dumitru Tucan, Mădălina Chitez, and Otto Kruse. 2018. **Introduction: Understanding Academic Writing in the Context of Central and Eastern European Higher Education**, pages 1–12. Springer International Publishing, Cham.
- Lynne Flowerdew. 2015. **Using corpus-based research and online academic corpora to inform writing of the discussion section of a thesis**. *Journal of English for Academic Purposes*, 20:58–68.
- Ken Hyland. 1998. **Boosting, hedging and the negotiation of academic knowledge**. *Text & Talk*, 18(3):349–382.
- Ken Hyland. 2005. **Stance and engagement: a model of interaction in academic discourse**. *Discourse Stud.*, 7(2):173–192.
- Alexandru Oravițan, Mădălina Chitez, Loredana Bercuci, and Roxana Rogobete. 2022. **Using the bilingual Corpus of Romanian Academic Genres (ROGER) platform to improve students’ academic writing**. In *Intelligent CALL, granular systems and learner data: Short papers from EUROCALL 2022*, pages 315–321. Research-publishing.net.
- Ellen F Prince, Joel Frader, Charles Bosk, et al. 1982. **On hedging in physician-physician discourse**. *Linguistics and the Professions*, 8(1):83–97.
- Françoise Salager-Meyer. 1994. **Hedges and textual communicative function in medical English written discourse**. *English for Specific Purposes*, 13(2):149–170.
- Geoff Thompson and Ye Yiyun. 1991. **Evaluation in the reporting verbs used in academic papers**. *Applied Linguistics*, 12(4):365–382.
- Paul Thompson. 2005. **Points of focus and position: Intertextual reference in PhD theses**. *Journal of English for Academic Purposes*, 4(4):307–323.
- Irena Vassileva. 2001. **Commitment and detachment in English and Bulgarian academic writing**. *English for Specific Purposes*, 20(1):83–102.
- Eva Thue Vold. 2006. **Epistemic modality markers in research articles: a cross-linguistic and cross-disciplinary study**. *International Journal of Applied Linguistics*, 16(1):61–87.

Appendix A Pilot cross-lingual hedge lexicon sample

Table 2 presents the core items from each of the six pilot hedging categories used in the cross-lingual analysis.

Category	English gloss	Romanian	Bulgarian
MODAL	may/can + V perhaps could (conditional) should + V cannot exclude	<i>a putea + V</i> <i>poate (adv.)</i> <i>ar putea</i> <i>ar trebui</i> <i>nu se poate exclude</i>	може да + V може би би могъл/а/о/и трябва да не може да се изключи
EPISTEMIC	seem, appear consider assume, suppose believe suggest observe, note find, establish propose, interpret	<i>a părea</i> <i>a considera</i> <i>a presupune</i> <i>a crede</i> <i>a sugera</i> <i>a observa; a remarca</i> <i>a constata; a rezulta</i> <i>a propune; a interpreta</i>	изглеждам смятам; считам предполагам; допускам склонен съм подсказвам; насочвам към наблюдавам; отбелязвам установявам; разглеждам предлагам; тълкувам
APPROXIMATOR	approximately almost relatively to a certain extent often, sometimes rather	<i>aproximativ</i> <i>aproape</i> <i>relativ</i> <i>într-o oarecare măsură</i> <i>adesea; uneori</i> <i>mai degrabă</i>	приблизително; около почти относително в известна степен често; понякога по-скоро
SHIELD	it can be said it can be assumed it is possible we could say remains open	<i>se poate afirma/spune</i> <i>se poate presupune</i> <i>este posibil (că)</i> <i>am putea spune</i> <i>rămâne de văzut</i>	може да се каже може да се допусне възможно е (да) бихме могли да кажем остава открит/спорен
ADVERB	probably possibly apparently however still undoubtedly (B)	<i>probabil</i> <i>posibil; eventual</i> <i>aparent</i> <i>însă; totuși</i> <i>încă</i> <i>fără îndoială</i>	вероятно възможно; евентуално привидно обаче; въпреки това все още; все пак несъмнено
DEFERENCE	according to in the opinion of from the perspective of as X states in X's work	<i>conform; potrivit</i> <i>în opinia; în viziunea</i> <i>din punct de vedere</i> <i>după cum</i> <i>în lucrarea/studiul</i>	според; съгласно по мнението на от гледна точка на както посочва/отбелязва в своя/неговия труд

Table 2: Core items from pilot cross-lingual hedge lexicon by category