Epistemic modal adverbs in Slovenian academic discourse

Jakob Lenardič* **and Darja Fišer*†**

*Department of Translation
Faculty of Arts
University of Ljubljana
Aškerčeva cesta 2, SI-1000 Ljubljana
jakob.lenardic@ff.uni-lj.si

†Department of Knowledge Technologies
Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana
darja.fiser@ff.uni-lj.si

Abstract

In this paper, we analyse Slovenian epistemic modal adverbs in the 100-million-token *KAS corpus of Slovenian PhD theses*. The focus of the analysis is a comparison of their usage in the humanities and social sciences on the one hand and natural and technical sciences on the other. Since modals are in principle ambiguous between epistemic and non-epistemic readings, we conduct a randomized concordance analysis with which we show that only those modals that are exclusively used in their epistemic sense are more frequent in the humanities and social sciences theses. We also show that the non-epistemic dispositional meaning of possibility, which is most commonly used in natural and technical sciences theses, does not constitute hedging.

1. Introduction

In this paper, we analyse Slovenian epistemic modal adverbs in the 100-million-token *KAS corpus of Slovenian PhD theses* (Erjavec et al., 2019c), comparing their use in the theses from the humanities and social sciences on the one hand and natural and technical sciences on the other.

Modals offer an interesting insight into academic discourse because they can pragmatically function as *hedges* (Lakoff, 1972; Hyland, 1996; Hyland, 1998), which are used by writers to present their claims with varying degrees of tentativeness. Hedging is a particularly important pragmatic strategy employed in academic writing, as it "enables writers to express a perspective on their statements, to present unproven claims with caution, and to enter into a dialogue with their audiences" and is therefore an "important means by which professional scientists confirm their membership in research communities" (Hyland, 1996, 251–252).

In the related work, which has been done primarily on the basis of English academic discourse, it is often shown that hedging is more characteristic of humanities and social sciences rather than natural sciences (Hyland, 1998; Takimoto, 2015), which reflects the general idea that humanities are more interpretative and less rooted in empirical research than natural sciences (Takimoto, 2015).

In this paper, we try to confirm whether this is also the case for Slovenian academic discourse. To our knowledge, this paper presents the first such comparative analysis of Slovenian academic disciplines from the perspective of hedging; in related work by e.g. Pisanski Peterlin (2010), the notion is exclusively discussed in the framework of translation theory in relation to how English hedges are translated into Slovenian and vice versa.

We present a descriptive quantitative analysis of 9 modal adverbs and then conduct a randomized concordance analysis with which we show that only those modals that are exclusively used in their epistemic sense and thus pragmatically correspond to hedging are also used more frequently in the humanities and social sciences subcorpora of the *KAS* corpus, which corresponds to similar findings for corpus-based analyses of English academic discourse reported by Hyland (1998), Takimoto (2015) and Rizomilioti (2006).

The paper is structured as follows. In Section 2., we lay out the relevant linguistic theory on modality and present the pragmatic notion of hedging, as well as discuss related work on corpus-based treatment of hedging in academic discourse. In Section 3., we present the corpus we used for our analysis from the perspective of the extra-linguistic metadata relevant for our purposes as well as discuss the selection criteria of the modal adverbs that we have analysed. In Section 4., we present the analysis. In Section 5., we discuss our findings by comparing them to those in the related literature and conclude the paper.

2. Theoretical framework

2.1. Epistemic and non-epistemic modalities

Modality has been defined in many different ways in the literature, but it is perhaps von Fintel (2006, 21) who most succinctly summarizes the notion:

> Modality is a category of linguistic meaning having to do with the expression of possibility and necessity. A modalized sentence locates an underlying or prejacent proposition in the space of possibilities [...] *Sandy might be home* says that

there is a possibility that Sandy is home. *Sandy must be home* says that in all possibilities, Sandy is home.

Modality thus evaluates a proposition from the perspective of the gradient from possibility to necessity.[1] Aside from this, modality comes in different semantic flavours that are contextually determined, and the usual linguistic distinction is made between epistemic modality on the one hand and non-epistemic modality on the other (Palmer, 2014), the latter of which is usually referred to as root modality (Coates, 1983) or circumstantial modality (Kratzer, 2012).[2]

Epistemic modality encompasses the speaker's judgement about the truth of the proposition (Palmer, 2014, 50); i.e., a modal like *mogoče* in the sentence *Ana je mogoče doma* "Ana might be home" constitutes an epistemic modal because its use asserts that the speaker is not completely certain that the prejacent i.e. unmodalised proposition *Ana je doma* "Ana is home" is true.

On the other hand, root (or circumstantial) modality also evaluates the proposition in the domain of possibility, but unlike epistemic modality does not tie the evaluation to the speaker's knowledge. An example of a non-epistemic modal is *lahko* in the sentence *Ta program se lahko namesti na Windows* "This program can be installed on Windows", where *lahko* is not used to indicate the speaker's knowledge about the truth of the expressed proposition but rather to attribute possible qualities to the subject NP *Ta program* "this program".

Modals are often ambiguous between epistemic and non-epistemic readings. For instance, *lahko* in the sentence *Ana je lahko doma, lahko pa je v šoli* is ambiguous between an epistemic reading that can be paraphrased as "'It is possible that Ana is at home or at school" and a root meaning that denotes permission that Ana is granted by someone else ("Ana is allowed to stay at home or in school").[3] Such often unpredictable ambiguity motivates the manual concordance analysis of the Slovenian modal adverbs that will be presented in Section 4.2.

Lastly, what is especially important for our purposes regarding academic discourse is that many root modal expressions display prominent meta-discursive usage, as in the case of reader-oriented meta-commentary clauses like *Kot lahko vidimo iz rezultatov* "As can be seen from the results". Such use along with the purely epistemic meaning often corresponds to the pragmatic notion of hedging (Hyland, 1996; Hyland, 1998; Grabe and Kaplan, 1997), which we introduce in the following section.

2.2. Hedging

In linguistics, Lakoff (1972, 471) was the first to use the term *hedges* to refer to those "words whose meaning implicitly involves fuzziness – words whose job is to make things fuzzier or less fuzzy". Lakoff (1972)'s basic concept is further explicated by Hyland (1996, 251), who claims that hedges are "any linguistic means used to indicate either (a) a lack of complete commitment to the truth of a proposition, or (b) a desire not to express that commitment categorically". Additionally, hedging not only involves markers of tentativeness but is typically extended to include rhetoric communicative strategies, i.e., politeness strategies by means of which the author implicitly includes the addressee in the discourse her or she is presenting (Grabe and Kaplan, 1997, 154).

Hyland (1996)'s definition of hedging overlaps quite significantly with that of epistemic modality defined in the previous section, but there is an important difference: a hedge is not a lexical property that holds of a specific category like modality, but rather corresponds to a pragmatic device that can in principle hold for any kind of lexical category given the right communicative context.

In terms of grammatical categories, hedging corresponds not only to modal verbs or adverbs, but to other lexical categories such as the use of certain reporting verbs that indicate the author's tentativeness (e.g., *We believe that*) as well as syntactic strategies such as the use of the passive rather than the active voice to syntactically omit the otherwise entailed agent of the verbal event (Rizomilioti, 2006, 56) or the use of inclusive plural pronouns to help establish rapport between the reader and the writer (Hyland, 1996).

2.3. Related work

In related work on hedging in academic discourse, researchers (Hyland, 1998; Rizomilioti, 2006; Pisanski Peterlin, 2010; Takimoto, 2015, a.o.) have generally taken into account all of the major categories that can in principle be used to hedge discourse, such as modal auxiliaries, modal and non-modal (e.g., approximators) adverbs and adjectives, and lexical verbs.

For instance, Takimoto (2015) analyses how hedges corresponding to 5 syntactic categories (adverbs, adjectives, auxiliaries, nouns, and verbs) are used across 4 different natural sciences disciplines and 4 humanities/social sciences disciplines, showing that "70% of all hedges and boosters were found in humanities and social sciences" (Takimoto, 2015, 103) and that philosophy contains "almost 5.3 times as many hedges and boosters as electrical engineering" (Takimoto, 2015, 103).[4] Similarly, Rizomilioti (2006, 64) compares the use of hedging between a

[1] Notions such as *possibility*, *likelihood*, and *necessity*, which are logically related by entailment, are also referred to as the *modal force* (Kratzer, 2012).

[2] To our knowledge, no one has yet attempted a comprehensive study of the syntactic and semantic properties of the Slovenian modal system in the context of descriptive Slovenian linguistics on par with e.g. Palmer (2014)'s work for English modal auxiliaries. Usually, researchers have studied Slovenian modals in relation to highly specialised topics in formal grammar – a good example of this is Marušič and Žaucer (2016)'s work, which proposes a generative syntactic explanation why the modal adverb *lahko* is a positive-polarity item – or have taken modals as a springboard for the study of non-linguistic topics in e.g. translation studies, rather than describing them in detail in their own right.

[3] This modal meaning involving obligation/permission is referred to as *deontic modality* by Palmer (2014).

[4] Some authors use the term *boosters* to describe those hedges that convey the author's certainty rather than tentativeness; since our analysis, presented in Section 4., does not show prominent differences between hedges and boosters, we use *hedges* as a catch-all term for expressing both tentativeness and certainty.

200,000 token corpus of journal papers in literary criticism and a comparable corpus of papers in biology, showing that there are more adverbs of uncertainty in the literary criticism corpus than in the biology corpus.

Given the high degree of lexical polysemy and the consequent likelihood that not all of the observed lexemes in the studied corpus function as hedges, a prominent strategy to filter out irrelevant data relies on the researcher having to read all the concordances that potentially correspond to hedges and then singling out only the relevant occurrences. For this to be possible, the corpora used in the related literature are usually quite small, generally consisting of 100,000–500,000 tokens and around 50–60 research articles (Thompson, 2000; Pisanski Peterlin, 2010; Hyland, 1998; Rizomilioti, 2006; Takimoto, 2015).

Nevertheless, despite such a strategy, the epistemic and non-epistemic notions of possibility seem conflated in some of the related work (Pisanski Peterlin, 2015). We therefore attempt to make our quantitative analysis of the modals more precise by making such a distinction between the modality types introduced in Section 2.1., arguing that only those instances of possibility expressed by the modals that correspond either to epistemic modality or the metadiscursive use constitute hedges, whereas non-epistemic meanings of possibility that correspond to dispositional ascriptions do not.

Our corpus, which we introduce in Section 3.1., is also significantly larger than those in the related literature, consisting of approximately 100 million tokens. Because a manual reading of such a large corpus was not a feasible option for us and because we wanted to reduce the amount of irrelevant data that in part arises from often unpredictable lexical polysemy, we conduct our analysis only on the basis of one word class, i.e., modal adverbs, which arguably constitute the most prominent category for expressing modality in Slovenian.

3. Methodology

3.1. The KAS Corpus of Academic Slovenian

The study presented in this paper has been carried out on the *KAS corpus of Slovenian PhD theses* (Erjavec et al., 2019c) (henceforth abbreviated as *KAS-dr*), which is a subset of the *KAS corpus of Slovenian academic writing* (Erjavec et al., 2019a). We have chosen the *KAS-dr* corpus because PhD theses represent the most uniform kind of Slovenian academic writing at the post-graduate level in comparison to e.g. the master's theses in the *KAS-mag 1.0* (Erjavec et al., 2019b) corpus, which, because of the Bologna reform, constitutes two types of writing referred to as "master's thesis", where the pre-Bologna variant is longer, more detailed and generally closer to PhD theses in terms of academic maturity than the Bologna variant.

The *KAS-dr* corpus consists of 1569 PhD theses, which together amount to approximately 100 million tokens – the PhD corpus thus represents roughly 6% of the entire 1.7-billion-token KAS corpus.The theses were written between 2000 and 2018 at Slovenian universities and other academic institutions.[5] The corpus is linguistically annotated but is also marked up for several extra-linguistic metadata categories that are tailored to the genre of academic theses, the most relevant for our purposes being the publisher and CERIF (Common European Research Information Format).

The Publisher information corresponds to the institution or faculty where the thesis was defended. There are a total of 44 different publisher abbreviations, 36 of which are faculties of the Universities of Ljubljana, Maribor, Nova Gorica, and Primorska. The remaining 8 are research institutes with their own study programmes or private and semi-private colleges.

The corpus represents a very diverse breadth of scientific (sub)disciplines so each thesis has been assigned to (at least) one of the five top-level CERIF[6] categories: BIO(MEDICAL SCIENCES), HUM(ANITIES), PHYS(ICAL SCIENCES), SOC(IAL SCIENCES), TECH(NOLOGICAL SCIENCES). Since the CERIF categories represent a generalised division of academic disciplines, they are particularly well-suited for comparative corpus analyses of academic genres, especially given the diverse disciplinary scope of the individual publishers included in the corpus.

The CERIF division of the theses in the *KAS-dr* corpus is given in Table 1.

CERIF	Size (in tokens and %)	
BIO	9,075,823	10%
HUM	12,911,252	14%
PHYS	9,785,950	11%
SOC	46,758,605	52%
TECH	11,993,724	13%
Σ	90,525,354	100%

Table 1: The five CERIF subcorpora of *KAS-dr*

As shown in Table 1, the five CERIF subsets of *KAS-dr* are unequal in size, with the SOC(IAL SCIENCES) subset accounting for half of the corpus. Consequently, we will provide frequency counts for our modal adverbs that are relativised to a million tokens in addition to their absolute frequencies (see Section 4.1.). Furthermore, the total token size (90,525,354) listed in Table 1 is slightly smaller than that of the entire *KAS-dr* corpus (101,473,395); this is because approximately 9% of the PhD theses are assigned to multiple CERIF categories, while the texts that we take into account include the majority of the theses with only one CERIF label.

3.2. The Modal Adverbs

The modal adverbs analysed in this paper are given in Table 2. There are 5 adverbs that denote possibility (*lahko, mogoče, možno, morda, morebiti*), 2 adverbs that denote likelihood (*najbrž, verjetno*), and 2 adverbs that denote certainty (*zagotovo, gotovo*).

The modals were selected in the following way. We first extracted all the lemmas in the *KAS-dr* corpus that are morphosyntactically tagged as either adverbs or as particles. Note that the Slovenian descriptive grammar *Slovenska*

[5] For a description of the KAS corpus, see Erjavec et al. (2020).

[6] https://www.eurocris.org/cerif/main-features-cerif.

slovnica (Toporišič, 2004), which is the basis for the MULTEXT tagset[7] used by the *KAS* corpora (Erjavec, 2012), postulates that particle is a separate word class. Toporišič (2004, 445–449), rather unusually, defines the particle class solely in terms of its semantic rather than syntactic properties, claiming that the category is distinct from e.g. adverbs in that it consists of semantically abstract clausal modifiers i.e. propositional operators rather than event modifiers such as manner or time adverbials. While most of the lexemes in Table 2 are tagged as adverbs in the corpus, *morda*, *najbrž*, and *morebiti* are tagged as particles, even though their syntactic distribution is prototypically adverbial.[8] From this extracted list of adverb and "particle" lexemes in the corpus, we selected all that semantically correspond to epistemic modals and have a minimum absolute frequency of 500 tokens in the entire *KAS-dr* corpus, which allowed us to omit very infrequent and stylistically marked modals like *nemara* "perhaps".

MODAL	AF	RF
lahko "possibly"	328,481	3628.6
verjetno "likely"	14,357	158.6
morda "possibly"	10,538	116.4
zagotovo "certainly"	3,630	40.1
gotovo "certainly"	3,458	38.2
mogoče "possibly"	2,194	24.2
možno "possibly"	1,518	16.8
najbrž "likely"	1,151	12.7
morebiti "possibly"	1,048	11.6

Table 2: The most frequent epistemic modal adverbs in the *KAS-dr* corpus, sorted by absolute frequency (AF) and relative frequency (RF)

The nine lexemes in Table 2 largely correspond to the epistemic modal adverbs identified for Slovenian by Pisanski Peterlin (2015, 31). However, in contrast to her approach, our selection criteria were stricter in that we excluded those adverbs that are frequently ambiguous between a modal and non-modal (e.g., manner) interpretation.[9] Such an ambiguous modal is *očitno* "apparently", as shown by the two possible paraphrases of example (1), where the first corresponds to a modal interpretation denoting the speaker's attitude towards the proposition while the other to a non-modal interpretation in which the adverb specifies the manner of the verbal event.

(1) Voda je očitno narasla.
"It appears that the water has risen."
"The water has risen in an obvious manner."

Discounting such ambiguous adverbs reduced the amount of irrelevant data, i.e., it helped us ensure that our comparative analysis is not hindered by the noise due to such polysemy.

4. The Analysis

4.1. Quantitative Analysis of Modal Adverbs Across Disciplines

Table 3 provides the distribution of the 9 modal adverbs in focus across the 5 CERIF subcorpora, listing both the absolute frequencies for the occurrence of each modal within a subcorpus as well as the relative frequency normalized to 1,000,000 tokens, which is needed for comparison because of the unequal sizes of the subcorpora.

We now compare the relative frequencies of the modals between the subcorpora. For easier comparison, we highlight in Table 3 the two highest relative frequencies for each modal by marking them in bold.

The modals can be divided into two groups.[10] The first group consists of 4 modals (*lahko, verjetno, mogoče, možno*), whose highest or second-highest frequency is in the natural science (BIO, PHYS) or technical (TECH) subcorpora. The modals in this group are on average 1.4 times more frequent in the BIO, PHYS and TECH subcorpora than in the HUM and SOC subcorpora; in other words, they are more frequently used in PhD theses from the natural/technical sciences than in theses within the humanities or social sciences.

The second group consists of 5 modals (*morda, zagotovo, gotovo, najbrž, morebiti*), whose highest relative frequencies are consistently observed in the HUM and SOC subcorpora. The five modals in this group are on average 2.2 times more frequent in the HUM and SOC subcorpora, meaning that in contrast to the first group they are more characteristic of PhD theses in the humanities and social sciences than theses from the natural/technical sciences.

The distribution of the first group is slightly uneven. Two modals – namely, *lahko* "possibly" and *možno* "possibly" – display both the highest and second-highest relative frequencies in either the BIO, PHYS or TECH subcorpora. For instance, the two highest relative frequencies of *možno* are 29.8 instances per million in the TECH subcorpus and 27.6 instances per million in the PHYS subcorpus, which is more than two times higher than the relative frequency of the adverb in the entire *KAS-dr* corpus (cf. Table 2) and around four times higher than the relative frequency of this adverb in the HUM corpus, which is 7.7 tokens per million.

In the case of the likelihood adverb *verjetno*, only its highest relative frequency (310.8 per million tokens) is observed in a natural sciences subcorpus i.e. BIO, while its second-highest frequency is in the HUM corpus (187.4 per million tokens). The distribution of the possibility adverb

[7] https://www.sketchengine.eu/slovene-tagset-multext-east-v5/

[8] In other words, there are no categorical differences between e.g. *verjetno*, which is tagged as an adverb, and *najbrž*, which is tagged as a particle. For simplicity's sake, we thus refer to all the 9 lexemes in Table 2 as adverbs.

[9] Admittedly, *lahko* also has a manner interpretation i.e. "easily". However, this use is incredibly rare – in our analysis of a randomized set of 250 concordance examples (cf. Section 4.2.) for this adverb, there was only 1 example, given in (i), where *lahko* is used in its comparative form *lažje* and corresponds to the non-modal manner usage.

(i) [. . .] zaradi česar *lažje* in pogosteje prihaja do sprememb v vrednostih indikatorjev.
". . . because of which changes in the values of the indicators occur more frequently and more easily."

[10] In Table 3, this division is shown by the dashed line.

	BIO		HUM		PHYS		SOC		TECH	
MODAL	AF	RF	AF	RF	AF	RF	AF	RF	AF	RF
lahko	30,405	3350.1	40,223	3115.3	46,481	**4749.8**	159,948	3420.7	51,424	**4287.6**
verjetno	2,821	**310.8**	2,420	**187.4**	1,406	143.7	6,563	140.4	1,147	95.6
mogoče	201	22.1	269	20.8	196	20.0	1,224	**26.2**	304	**25.3**
možno	121	13.3	99	7.7	270	**27.6**	671	14.4	357	**29.8**
morda	893	98.4	2,365	**183.2**	861	88.0	5,967	**127.6**	452	37.7
zagotovo	260	28.6	584	**45.2**	342	34.9	2,149	**46.0**	295	24.6
gotovo	153	16.9	917	**71.0**	297	30.5	1942	**41.5**	148	12.3
najbrž	63	6.9	324	**25.1**	108	11.0	596	**12.7**	60	5.0
morebiti	61	6.7	165	**12.8**	66	6.7	680	**14.5**	76	6.3

Table 3: Modal adverbs in *KAS-dr*; the relative frequency RF is normalized to a million tokens

mogoče is similarly uneven in that its highest frequency (26.2 per million) is in the SOC subcorpus, while its second-highest frequency (25.3 per million) is in the TECH subcorpus; we will account for this distributional property in Section 4.2.

In the second group, the highest and second highest frequencies are observed consistently in the HUM and SOC subcorpora. Furthermore, the differences between the highest and lowest relative frequencies are quite large. For instance, the highest relative frequency of *morda* is 183.2 per million tokens in the HUM subcorpus, which is around twice as many as in the BIO (98.4 per million) and PHYS (88.0 per million) subcorpora, and almost five times as many as in the TECH (37.7 per million) corpus.

Similarly, the second-highest frequency of *morebiti* is 12.8 per million tokens in the HUM corpus, which is almost twice as high as in the natural sciences/technical subcorpora, where it is 6.7 per million in the BIO and PHYS subcorpora and 6.3 in TECH subcorpus , all of which are lower than the 11.6 token-per-million frequency of *morebiti* in the entire *KAS-dr* corpus in Table 2.

4.2. Epistemic and non-epistemic usage

In order to account for the pattern presented in the previous section, which is the fact that 5 out of the 9 analysed modal adverbs occur most frequently in the humanities (HUM) and social sciences (SOC) subcorpora of *KAS-dr* while the remaining adverbs are equally or even more prominent in the three natural sciences/technical subcorpora, we have examined a randomized set of 250 concordance examples for each of the nine adverbs. We manually annotated each concordance line for the type of modality that the example expresses, given that modals are in principle ambiguous between epistemic and non-epistemic readings, as was discussed in Section 2.1.

The results of the concordance analysis are presented in Table 4.[11] The analysis has first shown that the use of the modal adverbs in the *KAS-dr* corpus can be grouped according to three major types of modality – epistemic use on the one hand and two types of circumstantial/root modality that we label as the dispositional use and the discursive use on the other.

Second, and most importantly, Table 4 shows that the distribution of epistemic and non-epistemic meanings of the adverbs generally follows the distribution of the modals in the CERIF corpora (Table 3) in the following sense: all the five modals that are used most frequently in the social sciences and humanities subcorpora (*morda*, *najbrž*, *zagotovo*, *gotovo* and *morebiti*) are also almost exclusively used to denote epistemic modality, whereas the other modals – with the exception of *verjetno*, which is also used exclusively as an epistemic modal – are to varying degrees ambiguous between the epistemic and non-epistemic readings.

We now take a closer look at the three types of meaning identified in the concordance analysis, and relate the use of modality to the notion of hedging that was introduced in Section 2.2. For instance, let's first take *morda*, which is used as an epistemic modal in 240 (96%) of the randomized concordances and only in 7 (4%) as a non-epistemic modal in the discursive sense, as being representative of the group that is almost exclusively epistemic. Sentence (2), which is taken from a thesis defended at the Faculty of Social Sciences, exemplifies this epistemic usage, while sentence (3), taken from a thesis at the Faculty of Pedagogy, exemplifies one of the few cases of the non-epistemic meta-discursive use of this modal.

(2) *Morda* je to eden od razlogov, da znanstvena skupnost ni bila uspešna pri svojem "programu" izboljšanja javnega razumevanja znanosti in znanstvene pismenosti.
"Perhaps this is one of the reasons that the scientific community wasn't successful in implementing their proposed program for improving the public understanding of science and scientific literacy."

(3) Zato lahko *morda* na tem mestu poudarim strinjanje z Banduro (1997), da je samoučinkovitost precej povezana s samouravnavanjem [. . .]
"This is why I can (perhaps) emphasise my agreement with Bandura (1997) that self-effectiveness is related to self-regulation."

Pragmatically, *morda* in its epistemic sense in example

[11] Note that, in Table 4, the number of included concordances for each modal is not always exactly 250, like 248 in the case of *možno*. The lower number in these cases is due to a few instances of incorrect part-of-speech tagging in the corpus (e.g., some syncretic premodifying adjectives, like *možno* in the accusative/instrumental NP *možno analizo* "possible analysis", are incorrectly tagged as adverbs); we have discarded such irrelevant occurrences from our analysis.

MODAL	EPISTEMIC		DISPOSITION		DISCURSIVE	
lahko	25	11%	117	47%	105	42%
mogoče	150	60%	97	39%	3	1%
verjetno	250	100%	0	0%	0	0%
možno	6	2%	233	94%	9	4%
morda	240	96%	0	0%	7	4%
najbrž	250	100%	0	0%	0	0%
zagotovo	243	98%	0	0%	4	2%
gotovo	245	98%	0	0%	5	2%
morebiti	250	100%	0	0%	0	0%

Table 4: The epistemic/root distribution of the modal adverbs in *KAS-dr*

(2) corresponds to Hyland (1996, 256–257)'s notion of an *accuracy-based hedge*, as it is used by the writer to denote his or her uncertainty about the validity of the proposition in the example; i.e., that whatever is denoted by the demonstrative *to* "this" in the main clause is indeed one of the reasons for the lack of success on part of the scientific community. All the epistemic examples with the remaining modals (which we do not exemplify here so as not to exceed the scope of the paper) also function as similar accuracy-based hedges, where the sole semantic and pragmatic difference is in the modal force of the lexeme in question; i.e., a modal like *najbrž* "likely" denotes a greater degree of the speaker's commitment to the truth of the proposition than *morda* or *morebiti* "possibly".

By contrast, *morda* in sentence (3) clearly does not denote the writer's uncertainty, and could be freely omitted from the sentence without a change in the propositional truth-commitment. It is rather used as part of a metadiscursive strategy with which the writer "acknowledge[s] the reader's role in ratifying knowledge" (Hyland, 1996, 258), in the sense that the lexical meaning of possibility which is inherently entailed by the modal "subtly hedges the universality of a writer's claim by implying that a position is an individual interpretation" (ibid.).

Such metadiscursive use is most prominent with the modal *lahko*, having been observed in 105 (42%) out of a total 250 of the randomized set of concordances. The sentence in (4), which is taken from a thesis at the Biotechnical Faculty, exemplifies this usage.

(4) Zaključimo *lahko*, da alkidni premazi na osnovi organskih topil izkazujejo nižje kontaktne kote na obeh substratih kot vodni akrilni premazi [. . .]
"We can conclude that alkyd coatings on the basis of organic solvents show smaller contact angles on both substrates than aqueous acrylic coatings. . ."

In all the 105 examples with the meta-discursive use of *lahko*, the modal adverb is used with directive verbs that are inflected for the so-called inclusive plural, like *zaključimo* "we conclude" in example (4). According to Takimoto (2015, 99), the use of "inclusive pronouns (e.g., *we*) [. . .] enables the writers to produce more interpresonal signals to the readers, which may allow the writers to share contexts with the readers and draw on their assumed belief specific to a particular field of study". In other words, the inclusive inflection emphasises the meta-discursive use of *lahko* as a hedge that is reader-oriented rather than accuracy-oriented (Hyland, 1996). Note that the remaining modals which are also used in this meta-discursive role (*mogoče, možno, morda, zagotovo, morebiti*) do not as consistently pattern with the inclusive plural inflection (cf. example (3), where the first person is used), which may possibly correlate with the fact that their use in this role is much less frequent in comparison to *lahko*, this being the de-facto modal for expressing meta-discursive commentary.

Finally, we turn to the fact that the modals *lahko*, *mogoče*, and *možno* convey, in addition to the epistemic and meta-discursive meanings, the root modal meaning that we refer to as the dispositional use. Sentence (5), which is taken from a thesis at the Faculty of Medicine, exemplifies this meaning with the modal *možno*, which is by far the most frequently used in this sense (233 or 94% examples), while sentence (6), which is from a thesis at the erstwhile Faculty of Electrical Engineering, Computer Science and Information Sciences, contains the modal *mogoče*, which is used in the dispositional sense in 97 (39%) of the concordance examples.[12]

(5) Upliniti je *možno* najrazličnejšo biomaso (les, oglje, kokosove olupke, riževe lupine).
"It is possible to gasify many kinds of biomass (wood, charcoal, coconut peels, rice husks)."

(6) Celoten grafični vmesnik je zasnovan tako, da ga je *mogoče* hitro prilagoditi potrebam metode [. . .]
"The entire GUI is designed in such a way that it can be easily tailored to the needs of the method."

In such cases, the modals are used to denote possibility in its root non-epistemic sense. This kind of modality is not concerned with the knowledge or attitude of the writer (as in the case of epistemic modals and those used in the meta-discursive sense), but is rather used to convey the characteristic properties (i.e., the disposition) on the basis of which

[12] In standard descriptive Slovenian linguistics, the lexemes *možno* and *mogoče* in sentences like (5) and (6) are usually referred to as adverbs; see e.g. the *Dictionary of Standard Slovenian* entry for *možno* (`https://hdl.handle.net/11346/IA5E`). Note, however, that in both examples *možno* and *mogoče* require that the VP be infinitival. It would therefore be more precise to analyse the two lexemes as predicative adjectives, on par with those heading extrapositional *it*-constructions in English like *It is possible to*+VP_{inf} (linden and Davidse, 2009). Conversely, adverbs in clausal adjunct positions are unable to govern the syntactic properties of other sentential constituents in such a way.

the underlying subject NP can be used in some way; for instance, example (6) says that the GUI is such that it is possible to tailor it to the needs of whatever is the method in question.

Palmer (2014, 38) claims that such subject-oriented modality is actually "not strictly a kind of modality at all, modality being essentially subjective", and that such modals are used "to make purely objective statements about the subject of the sentence" (ibid.). From the perspective of pragmatics, it does not seem that such dispositional modals actually constitute hedging of any kind given that they are used to convey objective properties of what the authors are describing in a given example. It should be noted that Hyland (1998, 5) claims that "hedges are the means by which writers can present a proposition as an opinion rather than a fact: items are only hedges in their epistemic sense, and only when they mark uncertainty". Example (5) does not involve the speaker's opinion one way or the other, hence it is not a hedge.

In relation to our observation that the non-epistemic use of modal adverbs is correlated with natural sciences and technical theses rather than those in the humanities or social sciences, it is then not surprising that *možno*, which is used in such dispositional contexts in the overwhelming majority of the randomized concordance examples, is by far the most frequent in the PHYS and TECH subcorpora (27.6 and 29.8 tokens per million, respectively), while its frequency in e.g. the HUM subcorpus, i.e., 7.7 tokens per million, is well below the 16.8 tokens-per-million frequency for the entire *KAS-dr* corpus (cf. Table 2). That is, *možno*, which is used almost exclusively as a non-attitudinal dispositional modal, is well suited for the natural sciences, which are generally objective in that they deal "with numerical data, which is more likely to generate a more precise picture of their findings" (Takimoto, 2015, 95) than e.g. the presumably more subjective and less empirical humanities.

By contrast, *mogoče*, which is lexically perfectly synonymous with *možno* in that it can in principle convey all the three observed types of modality, is used as a dispositional modal only in 39% cases, while 60% of the examples constitute the epistemic use in parallel with *morda* in example (2), so its use is more evenly distributed between natural/technical sciences on the one hand and social sciences/humanities on the other, as shown by the fact that the two highest relative frequencies are in the technical subcorpus (i.e., 25.3 per million in TECH) and in the social sciences subcorpus (i.e., 26.2 per million in SOC).

5. Discussion and conclusion

In this paper, we have analysed modal adverbs in the 100-million-token *KAS corpus of Slovenian PhD theses*, comparing their frequency and use between humanities and social sciences subcorpora on the one hand and natural sciences and technical subcorpora on the other. As one of our main contributions to the research on hedging, we have taken into account the fact that modals are in actual usage often unpredictably ambiguous between epistemic and non-epistemic readings, and argued that only those modals that either convey epistemic judgements or metadiscursive commentary also function as hedges, whereas those that express dispositional possibilities do not.

On the basis of this distinction, we have shown that 5 out of the 6 modals that are almost exclusively used in the epistemic sense (i.e., *morda, najbrž, zagotovo, gotovo, morebiti*[13]), and that thereby constitute accuracy-based hedges displaying varying degrees of the authors' tentativeness about the truth of the proposition, are used 2.2 times more frequently in Slovenian PhD theses in the humanities and social sciences rather than the natural and technical sciences.

This result is generally consistent with findings in similar studies that compare the use of adverbial hedging between humanities disciplines on the one hand and natural sciences on the other. For instance, Takimoto (2015) shows that in his corpus, the highest relative frequency of English adverbs of possibility, namely 1200 per million, is in the humanities disciplines, while the second highest is in the social sciences disciplines (800 per million) and the lowest in the natural sciences disciplines (600 per million) (Takimoto, 2015, 102). This is consistent with our findings, since we have for instance shown that the highest frequencies of the epistemic possibility adverb *morda* are also in humanities and social sciences (183.2 and 127.6 per million respectively), while the third highest, i.e. 98.4 per million is in the biology subcorpus. Similarly, Rizomilioti (2006, 64) shows that there are more adverbs of uncertainty in the literary criticism corpus (a total of 212 examples) than in her comparable biology corpus (a total of 181 examples), whereas we have shown an even greater difference – on average, the purely epistemic modals (that is, the accuracy-based hedges) in our corpus are 2.2 times more frequent in the humanities and social sciences.[14]

However, Rizomilioti (2006) shows that on the whole, i.e., when taking into account other categories such as modal auxiliaries, epistemic adjectives, and epistemic lexical verbs (e.g., *believe*, *indicate*), hedging is actually less

[13]The modal *verjetno* thus remains unaccounted for, so explaining the fact that it is the most frequent in natural sciences discourse, i.e., BIO in Table 3, despite its purely epistemic meaning is left for future work.

Nevertheless, we speculate that the difference arises because *verjetno* does not seem to be completely synonymous with *najbrž* even though both entail likelihood. *Verjetno* seems to have a stronger evidential meaning (i.e., the speaker has some empirical evidence for judging the given proposition as likely), whereas *najbrž* seems more rooted in non-evidential inference. A similar claim has been made for the distinction between the certainty modal auxiliaries in English, where the "difference between *will* and *must* is that *will* indicates what is a reasonable conclusion, while *must* indicates the only possible conclusion on the basis of the evidence available" (Palmer, 2014, 57).

If *verjetno* thus truly has a stronger evidential meaning than *najbrž*, it would then come as no surprise that it is more frequent in biomedical sciences, where empirical evidence abounds.

[14]We do note, however, that the empirical vs. non-empirical divide partially transcends the distinction between humanities/social sciences on the one hand and natural/technical sciences on the other, but is rather influenced by the methodological framework adopted by the researcher. Thus, a thesis in a humanities discipline may be more concerned with empirical data than other theses in the same discipline.

frequent in her literary corpus than it is in the biology corpus, which contradicts the findings reported by Takimoto (2015), who shows that hedging is the most frequent in humanities disciplines across all lexical categories.

The findings in the related work are thus to a degree inconsistent. Consequently, for future work we would like to extend our analysis of modals in the *KAS-dr* corpus to other word classes as well, such as adjectives (which are in many cases cognates of the adverbs, like ADJ *možen* "possible" vs. ADV *možno* "possibly") as well as lexical verbs reporting epistemic judgements. We will thereby be able to further ascertain whether expressions of epistemic modality are really more characteristic of humanities and/or social sciences disciplines across the board, as claimed by Takimoto (2015) and Hyland (1998), or whether they are a quirk of a specific word class, such as adverbs, as shown by Rizomilioti (2006).

6. Acknowledgments

We would like to thanks the anonymous reviewers for their comments and suggestions. The work described in this paper was funded by the Slovenian Research Agency within the national research programme *Slovene Language – Basic, Contrastive, and Applied Studies* (P6-0215) and within the national basic research project *Slovene Scientific Texts: Resources and Description* (J6-7094, 2014-2017).

7. References

Jennifer Coates. 1983. *The Semantics of the Modal Auxiliaries*. Croom Helm, London and Canberra.

Tomaž Erjavec, Darja Fišer, and Nikola Ljubešić. 2019a. *Corpus of Academic Slovene KAS 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1244.

Tomaž Erjavec, Darja Fišer, and Nikola Ljubešić. 2019b. *Corpus of Academic Slovene (MSc/MA theses) KAS-mag 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1266.

Tomaž Erjavec, Darja Fišer, and Nikola Ljubešić. 2019c. *Corpus of Academic Slovene (PhD theses) KAS-dr 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1265.

Tomaž Erjavec, Darja Fišer, and Nikola Ljubešić. 2020. The kas corpus of slovenian academic writing. *Language Resource and Evaluation*. Submitted/under review.

Tomaž Erjavec. 2012. Multext-east: morphosyntactic resources for central and eastern european languages. *Language Resources and Evaluation*, 46:131–142. https://doi.org/10.1007/s10579-011-9174-8.

William Grabe and Robert B. Kaplan. 1997. On the writing of science and the science of writing: Hedging in science text and elsewhere. In János S. Petöfi, editor, *Hedging and Discourse*, pages 151–167. de Gruyter, Berlin and New York.

Ken Hyland. 1996. Talking to the academy: Forms of hedging in science research articles. *Written Communication*, 13(2):251–281. https://www.doi.org/10.1177/0741088396013002004.

Ken Hyland. 1998. *Hedging in Scientific Research Articles*. John Benjamins, Amsterdam.

Angelika Kratzer. 2012. The notional category of modality. In *Modals and Conditionals: New and Revised Perspectives*, pages 27–69. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780199234684.003.0002.

George Lakoff. 1972. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic*, 2(4):458–508. https://www.jstor.org/stable/30226076.

An Van linden and Kristin Davidse. 2009. The clausal complementation of deontic-evaluative adjectives in extraposition constructions: a synchronic-diachronic approach. *Folia Linguistica*, 43(1):171–211. https://doi.org/10.1515/FLIN.2009.005.

Franc Marušič and Rok Žaucer. 2016. The modal cycle vs. negation in slovenian. In Franc Marušič and Rok Žaucer, editors, *Formal Studies in Slovenian Syntax*, pages 167–192. John Benjamins, Amsterdam. https://www.doi.org/10.1075/la.236.08mar.

F. R. Palmer. 2014. *Modality and the English modals*. Routledge, Abingdon-on-Thames.

Agnes Pisanski Peterlin. 2010. Hedging devices in slovene-english translation: A corpus-based study. *Nordic Journal of English Studies*, 9(2):171–193. http://doi.org/10.35360/njes.222.

Agnes Pisanski Peterlin. 2015. So prevedena poljudnoznanstvena besedila v slovenščini drugačna od izvirnih? korpusna študija na primeru izražanja epistemske naklonskosti. *Slavistična revija*, 63:29–44. https://srl.si/ojs/srl/article/view/COBISS_ID-57701986.

Vassiliki Rizomilioti. 2006. Exploring epistemic modality in academic discourse using corpora. In *Information Technology in Languages for Specific Purposes*, Educational Linguistics 7, pages 53–71. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-28624-2_4.

Masahiro Takimoto. 2015. A corpus-based analysis of hedges and boosters in english academic articles. *Indonesian Journal of Applied Linguistics*, 5(1):95–105. https://doi.org/10.17509/ijal.v5i1.836.

Paul Thompson. 2000. Modal verbs in academic writing. In Bernhard Kettemann and Georg Marko, editors, *Teaching and Learning by Doing Corpus Analysis – Proceedings of the Fourth International Conference on Teaching and Language Corpora*, pages 305–328.

Jože Toporišič. 2004. *Slovenska Slovnica*. Založba "Obzorja", Maribor.

Kai von Fintel. 2006. Modality and language. In Donald M. Borchert, editor, *Encyclopedia of Philosophy – Second Edition*, pages 20–27. MacMillan Reference USA, Detroit.