Identification of metadiscourse markers in bachelor's degree theses in Spanish

Introduction of a text mining tool

Carla Míguez-Álvarez, Luis G. Varela
& Miguel Cuevas-Alonso
Universidade de Vigo

This article introduces a text mining tool that can automatically extract information using Hyland's analysis model as a theoretical framework to analyse the use and characteristics of metadiscourse in large quantities of academic texts. To verify its validity, we present the results obtained using this tool on various bachelor's degree theses with a particular focus on the field of engineering. Our results on a 6.9 million-word corpus extracted from 680 bachelor's theses available online show that interactive metadiscourse markers are prevalent in engineering bachelor's theses as well as in the authors' metadiscourse patterns. In addition, we compared our results with previous research on metadiscourse markers. Our study can be used to identify the usage of various types of metadiscourse markers during the production of texts and for the development of software applications involving quantitative linguistic methods for the production of academic texts.

Keywords: metadiscourse, metadiscourse markers, corpus analysis, interactive markers, interactional markers, software program, text mining, bachelor thesis

1. Introduction

The last few decades have been key for the development of the study and analysis of discourse, with a particular emphasis on academic texts. Early research focused mainly on generic structures, lexical and grammatical features, and the rhetorical strategies that authors used in academic writing (Kuteeva & Mauranen, 2018). Nowadays, the analysis of the characteristics of the language used in creating documents, that is, the study of discourse, goes beyond its initial function as a mere tool for conveying information. As language is used in a social context, it

https://doi.org/10.1075/resla.20055.mig
Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics
ISSN 0213-2028 | E-ISSN 2254-6774

requires the utilization of a series of metadiscourse markers that show how the author organizes the discourse, their attitude toward the text, and how the audience comprehends and interacts with the document (Hyland, 2005; Saidian & Jalilifar, 2016). In other words, metadiscursive elements are rooted in the interpersonal realm (evaluative attitude vs. propositional content) or the textual realm (strategies used to achieve coherence and cohesiveness), but not in the ideational realm (Vande Kopple, 1985, p. 83). In addition to that, and as shown by Annelie Ädel (2006), there is a diference between a broad concept of metadiscourse in which all the linguistic resources that are used to organize the text and to express the author's opinions are taken into account, and a more narrow view that focuses only on the linguistic elements that fulfil a certain function in a text (Vande Kopple, 2002; Wei et al., 2016).

Metadiscourse markers are recognized as being particularly important in the analysis of both oral and written communication (Zhang, 2019) in a plethora of texts and genres with or without academic orientations (Jiang & Hyland, 2017; McGrath & Kuteeva, 2012; Saidian & Jalilifar, 2016). Several authors have classified metadiscourse markers into different main categories. Avon Crismore (1983) divided metadiscourse markers into two main groups: informational markers, which help improve the reader's understanding of a text, and attitudinal markers, which show the attitude of the author toward the content or structure of a text as well as toward the reader. For its part, William J. Vande Kopple's model (1985) divides metadiscourse into two vague categories: textual metadiscourse and interpersonal metadiscourse. Both categories overlap, forming seven types of metadiscourse markers: four textual (text conectives, narrators, code glosses and illocution markers) and three interpersonal (validity markers – which include hedges, emphatics and attributors – commentaries and attitude markers). Crismore & Rodney Farnsworth (1990) added to this model the category of scientific commentaries and divided metadiscourse into textual metadiscourse (text markers and interpretive markers) and interpersonal metadiscourse (hedges, certainty markers, attributors, attitude markers and commentary). Focusing on academic texts, Ken Hyland (1999, 2000) modified Crismore & Farnsworth's (1990) classification and, in 2005, he rejected the dichotomy of textual and interpersonal functions and created a model that differentiated between two macro-categories: interactive and interactional. These general categories are divided, in turn, into five subcategories each. Interactive markers include those subcategories that help guide the reader throughout the text, while interactional markers are used to involve the reader in the text (see Table 1 for a comparison between some of these models).

Table 1. Comparison of Crismore (1983); Hyland (1999) and Hyland (2005) models of metadiscursive markers

Crismore (1983)	**Hyland (1999)**	**Hyland (2005)**
Informational	*Textual metadiscourse*	*Interactive*
Goals	Logical connectives	Transitions
Pre-plans	Frame markers	Frame markers
Post-plans	Endophoric markers	Endophoric markers
Topicalizers	Evidentials	Evidentials
	Code glosses	Code glosses
Attitudinal	*Interpersonal metadiscourse*	*Interactional*
Saliency	Hedges	Hedges
Emphatics	Emphatics	Boosters
Hedges	Attitude markers	Attitude markers
Evaluative	Relational markers	Self-mentions
	Person markers	Engagement markers

Several authors have adapted Hyland's (2005) model to suit the text genre they were studying (Breeze, 2016; Hyland, 1998; Lin & Evans, 2012; Mur Dueñas, 2007, 2011; Suau Jiménez, 2015). However, there are few studies that adopt the interactional approach (Livingstone & Lunt, 2013; Mauranen, 2010), probably due to the fact that "interest in the interpersonal dimension of writing has, in fact, always been central to both systemic functional and social constructionist frameworks, which share the view that all language use is related to specific social, cultural and institutional contexts" (Hyland, 2005, p. 174).

In addition, the use of metadiscourse markers varies according to the type of text. It seems that academic texts engage in persuasion and argument more, that is, they include more interactive markers compared to interactional ones, due to the fact that references to information and data that appear in other parts of the text are necessary (Behnam & Mollanaghizadeh, 2015). Conversely, interactional markers are preferred in non-academic genres such as newspapers, internet texts and digital communications due to the fact that these markers are normally used to convey brief pieces of information or to express personal opinions and persuasion (Hyland, 1998; Kuhi & Mojood, 2014; Le, 2004; Shokouhi et al., 2015).

Nowadays, it is difficult to think about applied linguistic research, including the study of academic discourse, without the support of information and communication technologies (Kilgarriff & Grefenstette, 2003) and without discussing concepts such as 'big data' and 'data mining'. In fact, the constant creation and development of large amounts of text data has made it impossible to process data in ways that are not computer-based. In this sense, text mining techniques have been successfully used in a growing number of disciplines, thus resulting in multiple tools and applications (Ananiadou & McNaught, 2005; Cohen & Hersh, 2005; De la Calle et al., 2012; Orgeira-Crespo et al., 2020).

Qualitative analyses are difficult to perform when the data sample is large. In contrast, quantitative analyses are based on natural language processing, which is based on the idea that features of language or word usage can be counted and analysed statistically. The analysis on which this work is based follows two different approaches: on the one hand, the development of a simplified methodology for handling large volumes of data, and, on the other, its application to academic texts, focusing on metadiscourse markers.

Currently, several tools based on text analysis are available, such as MonoConc Pro (Barlow, 2000), WordSmith Tools (Scott, 2012), AntConc (Anthony, 2012) and Hyung Yoon & Ute Römer (2020). These tools cover a wide range of tasks including text data importing and structuring, coding text automatically or manually (following a dictionary or a categorization scheme), searching and retrieving text segments as well as coded text, testing hypotheses concerning the text material analysed and the categorization used, exporting the coding to other software, and generating reports on the coding performed (Alexa & Zuell, 2000). However, these tools need to be developed in greater depth, and none of them focus specifically on the analysis of metadiscourse markers in Spanish in a simple and user-friendly way.

2. Hypothesis and objectives

A traditional limitation of discourse studies is that large samples are required to draw general conclusions about how language is used in different communicative situations. In fact, it seems that the study of metadiscourse markers is usually restricted to very specific parts of documents (such as abstracts or conclusions) and is usually performed on a relatively small number of documents, both in terms of volume and length. Therefore, we set out to analyse the entire document as a whole and try to compare the results obtained to the data published about specific parts of texts in Spanish and in other languages.

We were not able to find a tool that analyses metadiscourse markers in Spanish using a simplified methodology for handling large volumes of data that will allow us to obtain a simpler and concise analysis compared to the analysis provided by other tools introduced in previous research.

Thus, the aim of this article is to present a new method based on text mining techniques that can automatically extract information from large databases or a series of documents. The software tool we present, named Metadiscourse Analyzer, uses Hyland's analysis model (2005) as a theoretical framework and is made for specific use within the field of linguistics to analyse the use of metadiscourse markers and their particular characteristics (including which types are used more in different fields) in a more simple and concise way. To verify its validity and as an example of one of its many possible uses, we present the results obtained using this tool on various bachelor's degree theses with a particular focus on the field of engineering.

In view of these research gaps, the research questions that have guided our analysis are the following: (1) is the data mining technique appropriate for the analysis of metadiscourse markers?; (2) in the case of the Spanish language, are the results comparable to the data obtained in other languages?

3. Method

The proposed work methodology (typical of text mining) can be divided into six distinct stages. Each stage defines the tasks that have been identified to cover the entire metadiscourse analysis procedure in a corpus. Stages 1 and 2 deal with the identification and selection of appropriate sources of information, as well as the transformation and filtering of bulk data. During these stages, the files are preprocessed manually (arrangement of files in folders and removal of elements that are not computable for the purposes of the analysis, such as figures) to enable subsequent automatic analysis. In Stage 3, the relevant information underlying the data obtained from information sources is extracted and classified automatically using specialized software. Stage 4 presents the output of the data processing and, during Stage 5, the extracted information is post-processed both manually and automatically to eliminate possible errors and to guarantee the quality of the information. Finally, Stage 6 focuses on the application of statistical techniques on the data to delve deeper into the information they contain. As such, the data are processed manually, although some of it comes directly from the program output. The program we present in the next section falls within the formal procedure described in Figure 1, stretching from Stages 2 to 4, thus forming part of a broader process.

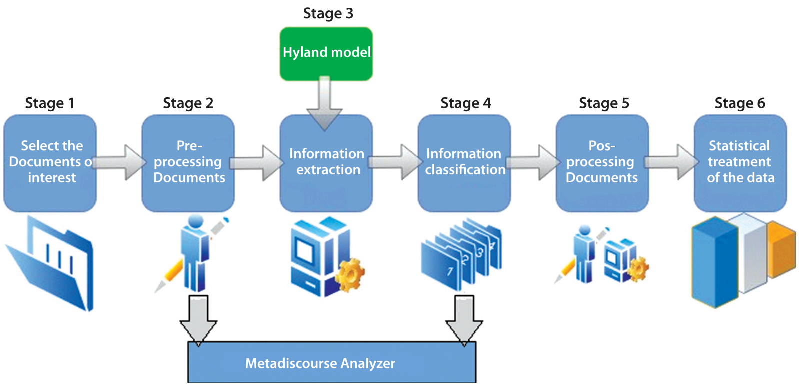

Figure 1. Formal procedure diagram

Going deeper into this part of the process, automatic data processing is essentially subdivided into four activities:

1. Determining objectives and scope. Establishing the objectives of the research by defining the extent of analysis and by pinpointing the limits and items of the search.
2. Gathering information for processing and pre-processing, which includes selection, analysis and pre-processing of the texts to be analysed. This stage is a slow process that combines both automatic and manual tasks.
3. Extraction of the information needed by the researcher.
4. Statistical treatment of the data extracted to find relevant associations, patterns and trends.

3.1 Description of the program

Metadiscourse Analyzer is a piece of software designed to retrieve occurrences of a series of words grouped under formal categories in a tree data structure where level dependencies are relevant and partially affect the results obtained. This software uses the model proposed by Hyland (2005) as its starting point.

In the present case study, the items that make up the various categories were used in response to an attempt to combine the classifications published by some authors (Hyland, 1999; Sorahi & Shabani, 2016; Tajeddin & Alemi, 2012). Special attention was paid to the publications that analyse English and Spanish from a comparative perspective (Ivorra Pérez, 2014; Lee & Casal, 2014; Lee & Deakin, 2016) and, particularly, to the classification provided by Pilar Mur Dueñas (2011), as shown in Appendix 1.

We chose to merge these classifications, despite the fact that they specialize in different domains, because our main goal was to create a tool that analyses all kinds of texts, regardless of their genre. In English, Hyland's (1999) research analysed extracts from 21 textbooks in microbiology, marketing and applied linguistics, while Mohammadamin Sorahi & Mansour Shabani (2016) focused on linguistics research articles. In Spanish, Mur Dueñas's (2011) list is based on Spanish business management articles, Francisco M. Ivorra Pérez (2014) analyses Spanish business websites, Joseph J. Lee & J. Elliot Casal (2014) study Spanish engineering theses, and Lee & Lydia Deakin (2016) analyse Chinese students' essays. However, it should be mentioned that we could only analyse the classifications that appear on papers that explicitly include a complete list of their search terms, a real rarity, and, thus, it prevented us from comparing a larger number of papers.

This software iteratively runs through a tree of directories analysing the documents in them (which can be in PDF, DOC, DOCX and TXT formats, with the latter being the recommended for its lack of text styles). Next, it extracts the useful information from each document both separately and in groups, that is, it merges the subtotals of all of the files contained in each (sub)group into a single one, as well as creating a single file with all the relevant info from all (sub)groups. The program then exports this information to XLS or XLSX files. Thus, in an ascending way from the deepest leaf of the tree, multilevel results are generated, which are the results of each document (data tree leaf) grouped by directories (tree nodes). This simple but effective process is a general solution that groups all of the information from the data tree.

A previous preparation of texts can be carried out on a regular basis to make them as comparable as possible. The pre-processing of texts has two purposes: on the one hand, to eliminate elements that are alien to the authors' metadiscourse, such as the titles of tables, graphics and figures that are commonly used in written documents, and, on the other hand, because some of the items can cause conflicts, such as images or graphics that are not formally part of the texts, it simplifies the computational processing of files during the automatic process. With this method of multilevel exploration, the organizational structure in folders and files is itself part of the software for the purpose of sorting the information to be processed.

One of the most relevant features of the software is its data traceability, in other words, the generation of multiple levels of results (per file, per folder tree directory). These ordered outputs allow us to trace the patterns found from the source (each of the specific texts) through the various categories defined for their study (gender, social origin, residence, nationality, level of studies, etc.), or to directly resort to the total results without any type of intermediate category. These possibilities not only make it possible to approach the same data from various perspectives (groupings) but also allow these perspectives to be applied simultaneously.

The use of the standard directory tree found in all available operating systems (OSs) satisfies both the need for multilevel analysis and how user-friendly this OS functionality is for basic personal computer users to sort and classify information using directories. Therefore, Metadiscourse Analyzer works as a text mining module for the quick and orderly extraction of repeating trends or patterns in large corpora made up of many text documents that may or may not be ordered in a directory tree based on the user's demand. See Appendix 2 for further information about the mathematical calculations that serve as the basis of this software. In addition, this software is easy enough for users to operate in a visual, intuitive and straightforward manner without the need for training.

3.2 Validity

To verify the validity of the software, we carried out a practical test using a set composed of bachelor's theses from various repositories of Spanish universities (usually in an open PDF format). The selected texts are accessible with the University of Vigo library resources from the Spanish Foundation for Science and Technology (FECYT) of the Spanish Ministry of Education, Culture and Sport.

To ensure comparability across disciplines, we selected bachelor's theses that meet the following criteria: (1) they were written by students from Spanish universities, (2) they have been compiled from the public repositories of Spanish universities, (3) they were submitted to the university during the period of 2013–2016, (4) they were organized using 'introduction-review-methodology-results-conclusion' as their typical structure (the acknowledgements and annex sections were discarded), (5) they were written by a single author, (6) all of the works are available to download as PDF and (7) the number of pages ranges from 60 to 100. Six hundred and eighty bachelor's theses from five educational fields were used: social sciences, human sciences, health sciences, scientific and engineering ($N > 666$, the threshold necessary for the data to be representative with a 99% confidence level). The data include theses from 34 different degrees (see Table 2 for more details).

These texts were selected to be analysed as representatives of both the technical and the formal use of language by undergraduate students preparing to graduate from different university disciplines. We evaluated the entire text, that is, the complete set of words included in the documents (after the pre-processing step described above). This is especially relevant because in the analysis of short texts or abstracts, authors tend to link short phrases that are more or less disconnected from each other to analyse the whole text in a small amount of space. This spe-

Table 2. Data detail of the bachelor's degree theses analysed according to educational field

Field	Number of theses	Different degrees	Number of words
Social sciences	140	7	1,591,308
Human sciences	120	6	1,033,891
Health sciences	120	6	929,570
Scientific	100	5	740,922
Engineering	200	10	2,652,809
Total	**680**	**34**	**6,948,500**

cific use of language, although it may be interesting for study, is not the authors' usual use of language in a lengthier context. It is necessary to process long texts to deepen the analysis of authors' metadiscourse, as long texts allow authors to put their rhetorical and linguistic resources into practice.

For the purpose of guaranteeing that the size of the set (sample size) was representative, we conducted a statistical study based on a series of basic principles (Azorín & Sánchez-Crespo, 1986; Cheaffer et al., 2007; Thompson, 1992). The estimator used is defined by the following equation:

$$n = \frac{N.k^2.(p\ .q)^2}{N.e^2 + .k^2.(p\ .q)^2} = \frac{\left[\frac{k^2.(p\ .q)^2}{e^2}\right]}{\left[1 + \frac{k^2.(p\ .q)^2}{N.e^2}\right]} \qquad \text{Equation 1}$$

Where N is the size of the population, p is the proportion of individuals in the population who meet a certain characteristic, and q is those who do not. K is the deviation that corresponds to confidence coefficient Pk, and e is the maximum permissible error. In this study, we considered that the size of the population progressively tends to infinity, as the number of final degree papers or bachelor's theses constantly grows. In this way, the expression could be simplified, resulting in the following:

$$n = \frac{k^2.(p\ .q)^2}{e^2} \quad \text{as} \lim_{n \to \infty} \left[1 + \frac{k^2.(p\ .q)^2}{N.e^2}\right]^{\infty} = 1 \qquad \text{Equation 2}$$

A 99% confidence level and a maximum error of 5% were considered. In these cases, the most conservative criterion was used as an option, making $p = q = 0.5$. Thus, a sample size could now be estimated using the above formula, where 2.58 is the value corresponding to a confidence level of 99% in a standardized normal distribution, assuming that the population size is not known or not large enough:

$$n = \frac{Z^2.(p\ .q)}{error^2} = \frac{2.58^2.(\ 0.5\ .0.5)}{0.5^2} = 666 \qquad \text{Equation 3}$$

4. Results

The output returned by Metadiscourse Analyzer provides some direct data that may be self-assessed, without applying additional statistical techniques. As expected, the first and most evident is that the amount of information in the documents analysed is considerably greater than the information extracted from them. Out of 6,948,500 words, 414,611 were items that could be analysed as discursive markers (Figure 2).

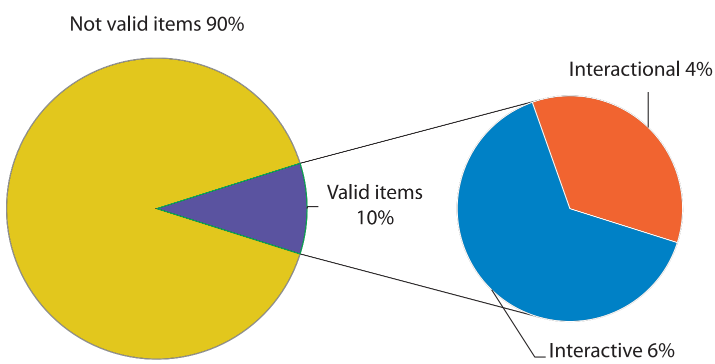

Figure 2. Ratio between the total number of words in the texts and the recognizable items (valid items) used in the study (automatically rounded percentages)

Regarding the distribution of markers in both categories (interactional and interactive) by scope, Figure 3 shows the dispersion of the data disaggregated by scientific branch on two axes. These data establish that the distribution of the markers varies depending on the field analysed: interactive markers are more prominent in scientific (80/40 per 1,000 words), engineering (75/35 per 1,000 words) and health sciences (68/34 per 1,000 words) bachelor's theses, whereas interactional markers are more common in human (68/52 per 1,000 words) and social sciences (62/40 per 1,000 words) bachelor's theses. In addition, it seems that bachelor's theses in engineering and social sciences are the ones that use the lowest number of discursive markers. Nevertheless, the assessment of the data obtained and their in-depth analysis, as well as the variations that exist between disciplines (Xiao & Sun, 2020), will be the subject of further work in the area of linguistic research.

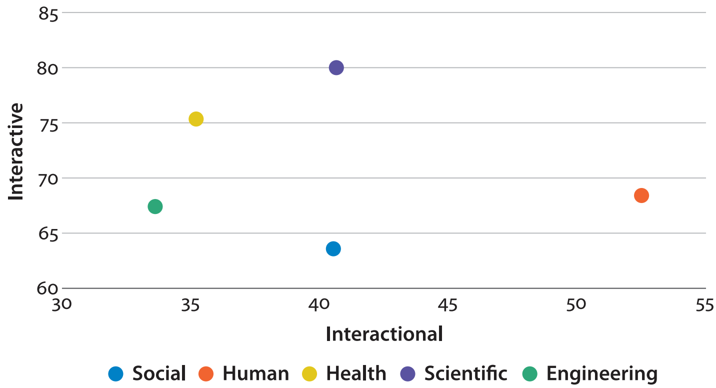

Figure 3. Distribution of interactive/interactional markers (mean per 1,000 words)

As a continuation of the demonstration of the results of the developed software, Figure 4 shows in detail the relationship between the two types of markers (interactional and interactive) within the field of engineering. As we can see, the presence of markers inherent to interactive categories (aimed at organizing information to make it comprehensible) far surpasses the markers of interactional categories (aimed at engaging the reader). This was to be expected, as research has shown a higher average of interactive markers compared with interactional markers in academic writing (Behnam & Mollanaghizadeh, 2015).

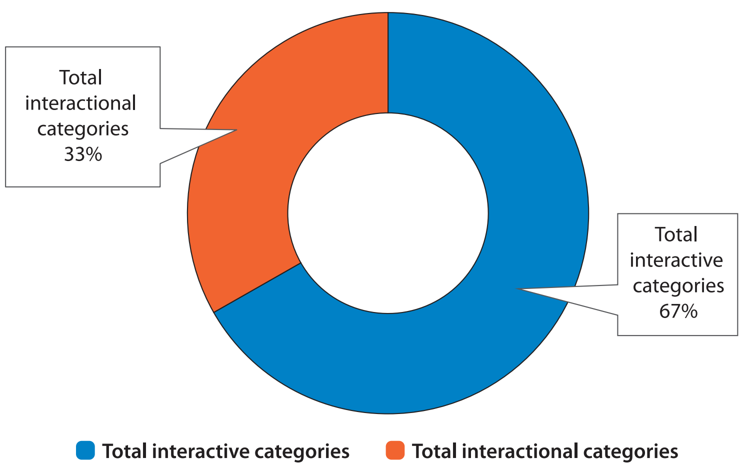

Figure 4. Distribution (interactive/interactional) in engineering theses

If we analyse the distribution of each type of interactional and interactive marker, we can see that the occurrences of the items in each subcategory within the field of engineering have a different distribution (see Table 3).

Table 3. Interactive and interactional occurrences in engineering bachelor's theses

	Occurrences	Occurrences per 1,000 words	Percentage within subcategory
Interactive categories	183,637	67.44	100%
Transitions/logical markers	72,890	26.77	40%
Code glosses	78,158	28.70	43%
Frame markers	8,202	3.01	4%
Endophoric markers	19,842	7.29	11%
Evidentials	4,545	1.67	2%
Interactional categories	91,422	33.57	100%
Hedges	17,002	6.24	19%
Boosters	10,411	3.82	11%
Attitude markers	17,840	6.55	20%
Engagement markers	33,981	12.48	37%
Self-mentions	12,188	4.476	13%

Going into detail, we can point out that the two most frequent interactive markers are code glosses and transitions/logical markers (43% and 40%, respectively). Code glosses help readers to understand the meaning of the information presented though extended examples and restatements, whereas transitions establish the semantic relationship that exists between clauses (Hyland, 2010; Saki, 2019). Transitions (mainly connectives) are, by far, the most used interactive subcategory in all research fields (Hyland, 2010). In fact, our research supports other investigations that show that both transitions and code glosses are the main interactive connectors in the field of engineering (Jin & Shang, 2016; Lee & Casal, 2014). However, contrary to Hyland's (2010) research – in which the author claimed that endophorics are highly used in engineering due to the constant reference to tables, figures, pictures, examples and so on – our results did not reveal a great use of these types of markers probably due to the differences in the authors' expertise (Hyland analysed PhD theses, while we focused on bachelor's theses). Nevertheless, we can see that, in this case, code glosses had a similar function to endophorics, providing additional information through restating, rewording, elaborating or explaining in detail some of the core concepts of the text.

In the case of interactional markers, the most frequent types are engagement markers (37%), followed by attitude markers (20%) and hedges (19%). Attitude markers convey a certain affective perception of the text, whereas hedges allow the author to acknowledge other points of view while avoiding full commitment and conveying a certain vagueness (Hyland, 2005; Salager-Meyer, 1994). Both subcategories are the most common interactional markers in research articles, which shows the need to assess the facts that appear in the text in both persuasive and acceptable ways (Akbarpour & Sadeghoghli, 2015; Hyland, 2010; Lee & Casal, 2014). Conversely, engagement markers are the discursive elements of the text that allow writers to include the reader in the text and direct them toward the intended interpretation. Although these markers are not common in the social, human and health sciences, and scientific fields, they are prominent in research that focuses on mathematics, physics and computation, as in our case (Akbarpour & Sadeghoghli, 2015; Hyland, 2010).

5. Discussion

Nowadays, thanks to the development of computer-based methods in the field of linguistics, there are different types of text analysis tools available, such as MonoConc Pro (Barlow, 2000), WordSmith Tools (Scott, 2012), AntConc (Anthony, 2012) and Yoon & Römer (2020). However, none of them focus specifically on the analysis of metadiscourse markers in Spanish in a simple and user-friendly way. Thus, we propose a new software tool (Metadiscourse Analyzer) that is easy to use and does not require any type of training. In addition, our software also sorts the information quickly under different criteria while reusing the same discretization tree. Finally, it shows the search terms used in the analysis, a feature that allows the user to make comparisons with other research.

To verify its validity and as an example of one of its many possible uses, we presented the results obtained using this tool on various bachelor's degree theses with a particular focus on the field of engineering. It should be mentioned that, for this particular work, we did not carry out any type of statistical analyses. In further research we will present the statistical results of our data (a sample of 6.9 million words from academic texts). The research questions were the following:

(1) *Is the data mining technique used appropriate for the analysis of metadiscourse markers?*

In the case of the analysis of bachelor's theses in engineering using the Metadiscourse Analyzer software, we found that interactive metadiscourse markers are prevalent in the writing of engineering bachelor's theses available online. In terms of metadiscursive markers, there are similarities in the use of the subcategories frame markers, endophoric markers and hedges. As previously stated, our research is in line with previously published studies, and the software tool we used allowed us to search for data and classify it in a very easy way. Therefore, we can establish that the software we developed can properly analyse the metadiscourse markers of Spanish bachelor's and master's thesis and, therefore, could possibly be used in the analysis of other types of documents.

(2) *In the case of the Spanish language, are the results comparable to the data obtained in other languages?*

It should be mentioned that comparing our software tool with other research was difficult not only because of the few studies of metadiscourse markers that are available in Spanish, but also because research on other languages does not explicitly include a complete list of the search terms used, which makes comparisons unfeasible.

Thus, we could only compare our research with those that explicitly include a complete list of their search terms: Hyland & Polly Tse (2004); Lee & Casal (2014) and Xin Jin & Yan Shang (2016) (see Table 4).

Lee & Casal (2014) analysed a corpus of 200 master thesis in engineering. Similarly to our own findings and following the analyses performed in other research fields, their results showed that interactive markers were more numerous than interactional ones. However, in our research, both code glosses and engagement markers had more prominent roles compared to in Lee & Casal's (2014) research. This is probably because we studied the whole document, whereas Lee & Casal (2014) focused only on the results and discussion sections, that is, they only focused on specific rhetorical sections. Therefore, the number of metadiscourse markers was higher. In addition, their research analysed master's theses, while ours focused on bachelor's degree theses. The difference in expertise between bachelor's and master's degree students accounted for the disparity in usage: bachelor's degree students need to use more restatements and examples (that is, code glosses) in order to explain their work compared to the more proficient master's degree and PhD students.

Although other papers studied the incidence of metadiscourse markers in bachelor's theses in engineering (see Hyland & Tse, 2004; and Jin & Shang, 2016, for further analysis), these studies focused specifically on the subfields of electronic engineering, computer science and material science, and, therefore, their findings cannot be generalized. In light of this, further research is needed to generalize the findings of this study.

This study has several limitations: First, the comparison with other research was difficult because there are few studies of metadiscourse markers in Spanish and, regarding studies on other languages, we only included those that explicitly feature a complete list of their search terms, which is truly exceptional and prevents us from comparing many texts. Second, in our research we could have added another series of markers that we considered appropriate, but for the sake of consistency in the analysis of the results we decided not to do so. Third,

Table 4. Interactive and interactional occurrences in engineering bachelor's/master's theses, comparison between studies (occurrences per 1,000 words)

	Hyland & Tse (2004)	Hyland & Tse (2004)	Lee & Casal (2014)	Lee & Casal (2014)	Jin & Shang (2016)	Jin & Shang (2016)	Present study (Metadiscourse Analyzer)
Documents analysed	**20 master's and 20 PhD theses in computer science**	**20 master's and 20 PhD theses in electronic engineering**	**100 master's theses in engineering**	**100 master's theses in engineering**	**20 bachelor's theses in material science**	**20 bachelor's theses in electronic engineering**	**200 bachelor's theses in engineering**
Section analysed	**All**	**All**	**Results and discussions**	**Results and discussions**	**Abstracts**	**Abstracts**	**All**
Language	**English**	**English**	**English**	**Spanish**	**Spanish**	**Spanish**	**Spanish**
Interactive categories	19.9	19.55	38.06	30.906	43.29	45.63	67.44
Transitions/ logical markers	7.43	7.69	19.22	14.5	39.14	37.41	26.77
Frame markers	3.54	2.47	2.89	2.97	2.84	5.48	3.01
Endophoric markers	2.59	4.31	10.62	8.806	0	0.78	7.29
Code glosses	3.23	3.07	3.24	3.32	1.31	1.96	28.70
Evidentials	1.31	2.01	2.09	1.31	0	0	1.67
Interactional categories	45.36	15.09	25.20	20.582	12.89	6.08	33.57
Hedges	5.58	6.15	11.06	6.73	2.4	1.57	6.24
Boosters	2.94	2.8	9.38	7.25	2.4	1.18	3.82
Attitude markers	1.62	1.06	2	1.322	0	0	6.55
Self-mentions	29.3	1.81	0.52	0.73	0.66	0	4.48
Engagement markers	5.92	3.27	2.2403	4.55	7.43	3.33	12.48
Total	**65.26**	**34.64**	**63.2603**	**51.488**	**56.18**	**51.71**	**101.01**

the findings made by a particular researcher can oftentimes not be replicated in another study. The reason for this is that the linguistic feature that under investigation is not being counted the same way (Anthony, 2013). We can palliate this issue by including the search terms used in our analysis. Finally, another important limitation is that we do not analyse the function of the metadiscourse marker within the paragraph where it is integrated, which would imply a greater complexity in the development of the software. There are other programs available for this purpose, but their use requires complex and professional training, which is not the objective of Metadiscourse Analyzer.

6. Conclusions

Current trends in linguistics research show that future studies in discourse linguistics will rely heavily on the analysis of large corpora that will result in the creation of more advanced software with an accessible interface. However, there is a noticeable lack of programs that are able to analyse and categorize metadiscourse markers in an easy and accessible way that does not require training.

This article has proved the usefulness of a generic model based on text mining techniques that allows for the automatic extraction of information from repository databases (in this study, as an example, we showed how to search for metadiscourse markers in bachelor's theses from Spanish universities). The results showed that the distribution of the markers varied depending on the field analysed, while highlighting the prevalence of interactive metadiscourse markers compared to interactional ones. The results apply to the writing of engineering bachelor's theses available online as well as to the metadiscourse pattern of the authors.

One of the great advantages of this procedure is that the designed software returns the occurrences of a series of words grouped under formal categories (as in Appendix 1). This is done in a tree data structure in which the level dependencies are relevant. The use of the standard directory tree found in all available OSs satisfies both the need for multilevel analysis and the ease of use of this OS functionality so that basic personal computer users can sort and classify information in a simple and user-friendly way.

This work opens up other possible future lines of research, including a new model of research in the stage of data extraction, as well as new possible pedagogical applications of research carried out with the software tool.

Funding

This work was supported by the Consellería de Cultura, Educación e Ordenación Universitaria (Xunta de Galicia, Spain) under Grants ED431C 2017/50, ED481A-2018/275 and ED431C 2021/52.

Acknowledgements

The authors would like to thank the reviewers for their corrections, which have contributed to improving the quality of this publication.

References

Ädel, A. (2006). *Metadiscourse in L1 and L2 English*. John Benjamins.

Akbarpour, M., & Sadeghoghli, H. (2015). The study on Ken Hyland's interactional model in OUP publications. *International Journal of Language and Linguistics*, *3*(4), 266–270.

Alexa, M., & Zuell, C. (2000). Text analysis software: Commonalities, differences and limitations: The results of a review. *Quality and Quantity*, *34*(3), 299–321.

Ananiadou, S., & McNaught, J. (2005). *Text mining for biology and biomedicine*. Artech House.

Anthony, L. (2012). *AntConc (Version 3.3.5)* [computer software]. Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Anthony. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research* *30*(2), 141–161.

Azorín, F., & Sánchez-Crespo, J. L. (1986). *Métodos y aplicaciones del muestreo* [sampling methods and applications]. Alianza.

Barlow, M. (2000). *MonoConc Pro* (Version 2.2) [computer software]. Available from http://www.athel.com/mono.html

Behnam, B., & Mollanaghizadeh, N. (2015). A comparative study of metadiscourse markers in some selected news programs on VOA: The case of regular English programs vs. special English programs. *Modern Journal of Language Teaching Methods*, *5*(1), 242–255.

Breeze, R. (2016). Negotiating alignment in newspaper editorials: The role of concur-counter patterns. *Pragmatics*, *26*(1), 1–19.

Cheaffer, R. L., Mendenhall, W., & Ott, L. (2007). *Elementos de muestreo* (6th ed.). [elements of sampling]. Grupo Editorial Iberoamérica.

Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, *6*(1), 57–71.

Crismore, A. (1983). *Metadiscourse: What it is and how it is used in school and non-school social science texts*. Retrieved from https://books.google.es/books?id=NfHmnAAACAAJ

Crismore, A., & Farnsworth, R. (1990). Metadiscourse in popular and professional science discourse. In W. Nash (Ed.), *The writing scholar: Studies in the language and conventions of academic discourse* (pp. 188–136). Sage.

De la Calle, G., García-Remesal, M., Nkumu-Mbomio, N., Kulikowski, C., & Maojo, V. M. (2012). e-MIR^2: A public online inventory of medical informatics resources. *BMC Medical Informatics and Decision Making*, *12*(1), 82–83.

Hyland, K. (1998). Persuasion and context: The pragmatics of academic metadiscourse. *Journal of Pragmatics*, *30*(4), 437–455.

Hyland, K. (1999). Talking to students: metadiscourse in introductory course books. *English for Specific Purposes*, *18*(1), 3–26.

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. Bloomsbury Publishing.

doi Hyland, K. (2010). Metadiscourse: Mapping interactions in academic writing. *Nordic Journal of English Studies*, *9*(2), 125–143.

doi Hyland, K., & Tse, P. (2004). Metadiscourse in academic writing: A reappraisal. *Applied Linguistics*, *25*(2), 156–177.

Ivorra Pérez, F.M. (2014). Cultural values and their correlation with interactional metadiscourse strategies in Spanish and US business websites. *ATLANTIS Journal of the Spanish Association of Anglo-American Studies*, *36*(2), 73–95.

doi Jiang, F.K., & Hyland, K. (2017). Metadiscursive nouns: Interaction and cohesion in abstract moves. *English for Specific Purposes*, *46*, 1–14.

doi Jin, X., & Shang, Y. (2016). Analyzing metadiscourse in the English abstracts of BA theses. *Journal of Language Teaching & Research*, *7*(1), 210–215.

doi Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics*, *29*(3), 333–347.

doi Kuhi, D., & Mojood, M. (2014). Metadiscourse in newspaper genre: A cross-linguistic study of English and Persian editorials. *Procedia – Social and Behavioral Sciences*, *98*, 1046–1055.

doi Kuteeva, M., & Mauranen, A. (2018). Digital academic discourse: Texts and contexts. Introduction. *Discourse, Context & Media*, *24*, 1–7.

doi Le, E. (2004). Active participation within written argumentation: Metadiscourse and editorialist's authority. *Journal of Pragmatics*, *36*(4), 687–714.

doi Lee, J.J., & Casal, J.E. (2014). Metadiscourse in results and discussion chapters: A cross-linguistic analysis of English and Spanish thesis writers in engineering. *System*, *46*(3), 39–54.

doi Lee, J.J., & Deakin, L. (2016). Interactions in L1 and L2 undergraduate student writing: Interactional metadiscourse in successful and less-successful argumentative essays. *Journal of Second Language Writing*, *33*, 21–34.

doi Lin, K.L., & Evans, S. (2012). Structural patterns in empirical research articles: A cross-disciplinary study. *English for Specific Purposes*, *31*(3), 150–160.

Livingstone, S., & Lunt, P. (2013). Mediated frameworks for participation. In M. Böck & N. Pachler (Eds.), *Multimodality and Social Semiosis* (pp. 79–88). Routledge.

doi Mauranen, A. (2010). Discourse reflexivity – A discourse universal? *Nordic Journal of English Studies*, *9*(2), 13–40.

doi McGrath, L., & Kuteeva, M. (2012). Stance and engagement in pure mathematics research articles: Linking discourse features to disciplinary practices. *English for Specific Purposes*, *31*(3), 161–173.

doi Mur Dueñas, P. (2007). 'I/we Focus on…': A cross-cultural analysis of self-mentions in business management research articles. *Journal of English for Academic Purposes*, *6*(2), 143–162.

doi Mur Dueñas, P. (2011). An intercultural analysis of metadiscourse features in research articles written in English and in Spanish. *Journal of Pragmatics*, *43*(12), 3068–3079.

doi Orgeira-Crespo, P., Míguez-Álvarez, C., Cuevas-Alonso, M., Doval-Ruiz, M.A. (2020). Decision algorithm for the automatic determination of the use of non-inclusive terms in academic texts. *Publications*, *8*(3), 41.

doi Saidian, S., & Jalilifar, A. (2016). "Mayhem! Absolute Mayhem!" Exploring the promotional metadiscursive features in the sportscasts of the 2014 FIFA World Cup semifinal between Brazil and Germany. *Discourse, Context and Media*, *14*, 9–17.

Saki, M. (2019). Metadiscourse and stance-taking in prefaces: A diachronic analysis. In S. Carter-Thomas & C.E. Hamilton (Eds.), *Science, Systemic Functional Linguistics and Language Change: A Festschrift for David Banks* (pp. 121–139). Cambridge Scholars Publishing.

doi Salager-Meyer, F. (1994). Hedges and textual communicative function in medical English written discourse. *English for Specific Purposes*, *13*(2), 149–170.

Scott, M. (2012). *WordSmith Tools* (Version 5.0) [Computer Software]. Available from http://www.lexically.net/software/index.htm

doi Shokouhi, H., Norwood, C. & Soltani, S. (2015). Evidential in Persian editorials. *Discourse Studies*, *17*(4), 449–466.

doi Sorahi, M., & Shabani, M. (2016). Metadiscourse in Persian and English research article introductions. *Theory and Practice in Language Studies*, *6*(6), 1175–1182.

doi Suau Jiménez, F. (2015). Quality translation of hotel websites: Interpersonal discourse and customer's engagement. *Onomázein*, *32*, 152–170.

Tajeddin, Z., & Alemi, M. (2012). L2 learners' use of metadiscourse markers in online discussion forums. *Issues in Language Teaching (ILT)*, *1*(1), 93–121.

Thompson, S.K. (1992). *Sampling*. John Wiley & Sons.

doi Vande Kopple, W.J. (1985). Some exploratory discourse on metadiscourse. *College Composition and Communication*, *36*, 63–94.

Vande Kopple, W.J. (2002). Metadiscourse, discourse, and issues in composition and rhetoric. In F. Barton & C. Stygall (Eds.), *Discourse studies in composition* (pp. 91–113). Hampton Press.

Wei, J., Li, Y., Zhou, T., & Gong, Z. (2016). Studies on metadiscourse since the 3rd millennium. *Journal of Education and Practice*, *7*(9), 194–204.

doi Xiao, W., & Sun, S. (2020). Dynamic lexical features of PhD theses across disciplines: A text mining approach. *Journal of Quantitative Linguistics*, *27*(2), 114–133.

doi Yoon, H., & Römer, U. (2020). Quantifying disciplinary voices: An automated approach to interactional metadiscourse in successful student writing. *Written Communication*,

doi Zhang, M. (2019). Exploring personal metadiscourse markers across speech and writing using cluster analysis. *Journal of Quantitative Linguistics*, *26*(4), 267–286.

Appendix 1. Complete list of search terms

10
Interactive metadiscourse categories – Transitions/logical markers
36
"además", "por otro lado", "por otra parte", "por su parte", "igualmente", "asimismo", "y", "adicionalmente", "a su vez", "de igual forma", "de igual modo", "análogamente", "incluso", "sin embargo", "no obstante", "por el contrario", "pero", "en cambio", "ahora bien", "a pesar de ello", "con todo", "aun así", "por tanto", "así", "por ello", "por todo ello", "por casi todo ellos", "por causa de ello", "así pues", "por lo tanto", "pues", "por este motivo", "por consiguiente", "por esta razón", "entonces", "en consecuencia"
Interactive metadiscourse categories – Code glosses
26

"es decir", "en concreto", "esto es", "concretamente", "esto significa", "se trata de", "en otras palabras", "nos referimos a", "en todo caso", "en cualquier caso", "finalmente", "por último", "en definitiva", "para finalizar", "en conclusión", "resumiendo", "en resumen", "en síntesis", "como", "por ejemplo", "tales como", "e.g.", "ej.", "(", ":", "-"
Interactive metadiscourse categories – Frame markers
43
"primero", "primera", "en primer lugar", "segundo", "segunda", "en segundo lugar", "tercero", "tercera", "en tercer lugar", "cuarto", "cuarta", "en cuarto lugar", "quinto", "quinta", "en quinto lugar", "sexto", "sexta", "en sexto lugar", "séptimo", "séptima", "en séptimo lugar", "octavo", "octava", "en octavo lugar", "noveno", "novena", "en noveno lugar", "décimo", "décima", "en décimo lugar", "por último", "finalmente", "por un lado", "por otra parte", "a continuación", "en cuanto a", "con respecto a", "respecto a", "por lo que se refiere a", "en relación a", "en relación con", "relacionado con", "por lo que respecta a"
Interactive metadiscourse categories – Endophoric markers
22
"anterior", "anteriores", "hipótesis", "anteriormente", "previo", "previamente", "apartado", "anterioridad", "siguiente", "siguientes", "adelante", "como sigue", "seguidamente", "epígrafe", "a continuación", "tabla", "figura", "gráfico", "cuadro", "fig.", "sección", "capítulo"
Interactional metadiscourse categories – Evidentials
1
"ref."
Interactional metadiscourse categories – Hedges
72
"poder", "*ría", "parecer", "caber", "soler", "observar", "considerar", "plantear", "pensar", "apreciar", "indicar", "pretender", "entender", "tratar de", "sugerir", "intentar", "suponer", "tender a deducir", "esperar", "creer", "implicar", "interpretar", "querer", "estimar", "prever", "asumir", "predecir", "apuntar", "bastante", "casi", "aproximadamente", "posiblemente", "normalmente", "quizá", "quizás", "probablemente", "frecuentemente", "habitualmente", "tal vez", "previsiblemente", "ligeramente", "aparentemente", "apenas", "posible", "planteado", "propuesto", "probable", "pretendido", "aproximado", "indicado", "hipótesis", "probabilidad", "posibilidad", "tendencia", "idea", "planteamiento", "visión", "interpretación", "intento", "propuesta", "observaciones", "concepción", "percepción", "argumentos", "en general", "en parte", "en principio", "en cierta medida", "en cierta forma", "hasta cierto punto", "a priori", "en términos generales"
Interactional metadiscourse categories – Boosters
77
"mostrar", "determinar", "destacar", "comprobar", "confirmar", "corroborar", "demostrar", "afirmar", "poner de manifiesto", "constatar", "verificar", "resaltar", "revelar", "concluir", " evidenciar", "enfatizar", "subrayar", "saber", "remarcar", "probar", "arrojar", "fundamentalmente", "significativamente", "especialmente", "principalmente", "ampliamente", "generalmente", "efectivamente", "siempre", "sustancialmente", "claramente", "obviamente", "realmente", "indudablemente", "netamente", "predominantemente", "profusamente", "claro", "evidente", "cierto", "demostrada", "amplio", "considerable", "inequívoco", "hecho", "determinación", "evidencia", "mayoría", "argumentos", "muestra", "afirmación", "verdad", "de hecho", "en mayoría", "en la mayoría", "en esta mayoría", "en esa mayoría", "en aquella mayoría", "en abrumadora mayoría", "en amplia mayoría", "en evidente mayoría", "en gran mayoría", "en la gran mayoría", "en la inmensa mayoría", "en casi la mayoría", "en aproximadamente la mayoría", "en una importante mayoría", "en una relevante

mayoría", "en efecto", "sin duda", "sin lugar a dudas", "en gran parte", "en gran medida", "en buena medida", "de manera sustancial", "de manera significativa", "de un modo amplio"
Interactional metadiscourse categories – Attitude markers
89
"principal", "importante", "relevante", "gran", "necesario", "bueno", "válido", "escaso", "amplio", "lógico", "conveniente", "fundamental", "fiable", "clave", "especial", "fuerte", "interesante", "apropiado", "básico", "coherente", "difícil", "fácil", "grandes", "imprescindible", "mejor", "pobre", "significativo", "sólido", "útil", "aceptable", "congruente", "curioso", "decisivo", "esencial", "esperanzador", "deseable", "desfasado", "excelente", "indispensable", "irremediable", "nefastas", "el primer", "primer", "problemático", "prudente", "radical", "razonable", "replicado", "riguroso", "robusto", "sencillo", "singular", "sorprendente", "suficiente", "vital", "convenir", "aportar", "garantizar", "contribuir", "ir", "ir más allá", "limitar", "ignorar", "subsanar", "importancia", "limitación", "problema", "fiabilidad", "aportación", "interés", "dificultad", "validez", "carencia", "complejidad", "relevancia", "falta", "problemática", "utilidad", "conveniencia", "disparidad", "diversidad", "negativamente", "solamente", "suficientemente", "adecuadamente", "debidamente", "fielmente", "paradójicamente", "rotundamente"
Interactional metadiscourse categories – Engagement markers
39
"*mos", "nos", "nuestro", "nuestros", "nuestra", "nuestras", "uno", "vosotros", "vosotras", "ustedes", "usted", "vuestro", "vuestra", "vuestros", "vuestras", "su", "suyo", "suya", "suyos", "suyas", "?", "ver", "tener que", "deber", "haber que", "deber", "haber que", "tienen que", "deben", "tenéis que", "debéis", "tiene que", "debe", "tuvieron que", "debieron", "tuvisteis que", "debisteis", "tuvo que", "debió"
Interactional metadiscourse categories – Self-mentions
13
"yo", "nosotros", "nosotras", "el autor", "la autora", "los autores", "las autoras", "*mos", "nos", "nuestro", "nuestros", "nuestra", "nuestras"

Appendix 2. Mathematical calculations used

From a mathematical perspective, the extracted data can be understood as the repetition frequencies of specific items. The weight of a string of text is equivalent to its number of occurrences. The calculation of this weight provides information beyond just pointing out its existence in a binary way. This is because, as a rule, the terms that appear frequently are usually the most relevant to the text. Caution should be taken with this approach because the number of occurrences of a string of text can lead to errors when the repeating string is a non-valid one for the study, such as an empty string.

Table 6. Symbols and correspondence

	Nomenclature	
Symbol	**Parameter**	**Units**
n	Population size that meets a condition	(–)
q	Complementary set of n, population that do not meet the condition	(–)

	Nomenclature	
Symbol	**Parameter**	**Units**
k	Confidence coefficient deviation	(–)
Pk	Confidence coefficient	(–)
e	Maximum error allowed	(–)
w	Term or item	(string)
d	Document	(fie)
D	Corpus, set of documents	(file array)
Ψ	File term or item	(string)
Ψn	Total elements of a file	(–)

For a string of text or term *w* (item) of document *d*, which is a part of a corpus *D*, its total frequency of occurrence (*ttf*) is defined by the summation in *D* of the number of times (occurrences) that term *w* appears in each document *d* or *tf*(*w, d*):

$$ttf(w, D) = \sum_{i=0}^{D} tf\,(w, d_i) \qquad \text{Equation 4}$$

In turn, *tf*(*w, d*) is supported by the following equation:

$$tf(w, d) = \sum_{j=0}^{\Psi_n} cmp\,(w, \Psi_j) \qquad \text{Equation 5}$$

Where for term *w* in the specific document *d, d* is traversed item to item, from *j = o* to *j = n*, with *Ψn* being the total number of items of *d*. This verifies that term *w* is equal to each item *Ψ*. The number of occurrences of *w* in *d* is the result of the summation in *d* of comparisons *cmp*(*w, Ψ*) of term *w* with various items in file *d*.

$$cmp(w, \Psi_j) = \begin{cases} 1 & \Psi_j = w \\ 0 & \Psi_j <> w \end{cases} \qquad \text{Equation 6}$$

Identificación de marcadores de metadiscurso en trabajos de fin de grado en español: Introducción de una herramienta de minería de textos

Resumen

En este artículo se presenta una herramienta de minería de textos que permite extraer información de forma automática utilizando el modelo de análisis de Hyland como marco teórico para analizar el uso y las características del metadiscurso en grandes cantidades de textos académicos. Para comprobar su validez, presentamos los resultados obtenidos utilizando esta herramienta en varios trabajos fin de grado (TFG), con un enfoque particular en el ámbito de la ingeniería. Nuestros resultados sobre un corpus de 6,9 millones de palabras extraídas de 680

TFG mostraron que los marcadores interactivos del metadiscurso son frecuentes en la elaboración de TFG de ingeniería, así como en los patrones de metadiscurso de los autores. Por otra parte, también hemos comparado los resultados obtenidos en esta investigación con trabajos previos relacionados con los marcadores del metadiscurso. Este estudio puede servir para identificar el uso de varios tipos de marcadores del metadiscurso durante la producción de textos y para el desarrollo de aplicaciones software que empleen métodos lingüísticos cuantitativos para la producción de textos académicos.

Palabras clave: metadiscurso, marcadores del metadiscurso, análisis de corpus, marcadores interactivos, marcadores interaccionales, programa de software, minería de textos, trabajo fin de grado

Address for correspondence

Carla Míguez-Álvarez
Department of Specific Didactics, Area of Didactics of Language and Literature
Universidade de Vigo
Campus Lagoas Marcosende
Vigo, 36310
Spain
camiguez@uvigo.es
https://orcid.org/0000-0001-6355-579X

Co-author information

Luis G. Varela
Industrial Engineering School
Universidade de Vigo
luisgonzalez@uvigo.es
https://orcid.org/0000-0003-1284-1379

Miguel Cuevas-Alonso
Department of Spanish Linguistics
Universidade de Vigo
School of Philology and Translation
https://orcid.org/0000-0001-7656-2374

Publication history

Date received: 26 August 2020
Date accepted: 28 May 2021