Interdisciplinary and interlinguistic perspectives on Academic Discourse: the mode variable

Introduction to the special issue on the French EIIDA[1] project

Shirley Carter-Thomas°, Marie-Paule Jacques*
°Institut Mines-Télécom (TEM), LATTICE (UMR 8094) CNRS, ENS (PSL research university), Paris 3-Sorbonne Nouvelle, *ESPE & LIDILEM, Université Grenoble-Alpes

Keywords: contrastive corpora, discipline, langage, mode

1. Introduction and context of project

The objective of the French EIIDA[2] project was to compare written and oral academic discourse, and to examine the impact of mode on the way scientific discourse is formulated and structured across languages (English, French and Spanish) and disciplines (hard sciences vs. humanities).

Research revolving around the study of academic discourse and languages for specific purposes has long recognised the need for specialised (discipline-specific) corpora to supplement general purpose academic corpora. Although studies based on large multi-purpose academic corpora are undoubtedly useful for describing general tendencies and for providing reference baseline

[1] Etude Interdisciplinaire et Interlinguistique du Discours Académique
Estudio Interlingüístico y Interdisciplinario del Discurso Académico
Interlinguistic and Interdisciplinary Study of Academic Discourse

[2] The 3-year EIIDA project (2012-2014), directed by Shirley Carter-Thomas and Jeanne-Marie Debaisieux, was funded as part of the Labex TransferS programme of the Ecole Normale Supérieure (Paris).

frequencies of lexico-grammatical features (see for example Biber et al 1999), they do not enable disciplinary specificities to be taken into account. As Bazerman (1981) and Hyland (2000) for example have shown, hard and soft disciplines can vary considerably in the way they approach their object of study, evaluate knowledge claims and cite previous research. In recent years, we have therefore witnessed a wealth of studies, particularly in ESP, focusing on discipline-specific features of academic discourse.

Discipline-specific corpora are also needed however in languages other than English. The process of knowledge transfer in the academic context can be considerably influenced by language and by culture-specific norms conveyed both by traditions and the educational system (cf. for example, Bennett, 2010; Clyne, 1996; Molino, 2011). The KIAP project (Cultural Identity in Academic Prose) led by K. Fløttum at Bergen University in Norway (Fløttum et al., 2006; Fløttum ed., 2007) was designed specifically so that these issues could also be addressed. The KIAP contrastive corpus created covers three disciplines (medicine, economics and linguistics) and three langages (English, French and Norwegian) and thus enables a number of interesting comparisons between discipline and language variables. Only one text genre is included in the KIAP corpus however: the scientific article.

Although there have also been a number of studies contrasting French and English academic discourse in France, it is not easy to survey the field as research is scattered and has not been widely disseminated. Many researchers have compiled their own “private” ad hoc corpora for specific research or pedagogical projects, such as the contrastive Earth Science Corpus (ESC) at University Paris Diderot[3]. Other corpora have been collected in the context of a PhD thesis, examples being the comparable (French/English) collection of mainly popularized texts in volcanology, created by Josselin for her PhD in lexicography (Josselin 2005), and the corpus created by Bordet for her contrastive analysis of PhD abstracts (Bordet 2011). Efforts though are now underway to pool resources. Since 2012, for example, the LIDILEM laboratory at Grenoble Alpes University has made available a searchable online collection of scientific texts in French and in English that is accessible to the general public. The texts making up the Scientext collection are tagged, and searches can be made for specific features using the query tool (Tutin et al 2013). They cover several written genres, including research articles and student essays. As yet, however, the collection has no oral component.

[3] See Kübler, N. & Volanschi 2012 for a description of the comparable ESC corpus.

According to Robles Garrotte's survey (2016), the situation in Spain is fairly comparable, with a significant lack of academic oral corpora. In the early 2000's a corpus of lecture courses in different languages was created as part of the ADIEU project. For her own studies, Robles Garrotes (2013) drew upon a corpus of conference presentations in Spanish and Italian, recorded at various conferences devoted to teaching in several disciplines. Robles Garrotes also mentions other studies carried out in Latin America (Venezuela, Argentina), with the aim of characterising some specific features of scientific presentations. However, to the best of our knowledge, these corpora do not cover a range of disciplines and more importantly do not provide a written counterpart with which the talks can be compared.

The comparative lack of interest in oral academic corpora and in comparisons between spoken and written academic discourse is surprising in that research on academic registers has found that mode differences are extremely important in accounting for linguistic variation (e.g. Biber, 2006; Swales, 2004). Spoken academic genres, such as conference presentations, display a very different range of lexico-grammatical features from written academic genres (Carter-Thomas & Rowley-Jolivet, 2001). It was therefore with the aim of redressing this balance that the EIIDA project was created. The main rationale behind the EIIDA project[4] was to create a corpus enabling the comparison of academic discourse from a triply contrastive perspective: discipline, language and mode.

2. Design of the EIIDA corpora

We therefore needed to build a corpus which would give us access to samples of academic discourse in different languages, modes and disciplines.

Accordingly we collected data in:

- three languages: English, French, Spanish;
- two modes: written (research articles) *vs* oral (conference presentations);
- two sets of disciplines: humanities (linguistics) *vs* sciences (geo-chemistry and water sciences).

[4] The EIIDA corpus is currently being cleaned and formatted. It will soon be freely available (at least in part) to the research community via the site of Scientext (http://scientext.msh-alpes.fr).

We ensured that the oral presentations recorded were all given by a native speaker of the respective language. They were transcribed using Transcriber (Barras et al. 2001), in most cases also by native speakers. The written texts were in some cases the articles from the conference proceedings corresponding to the conference presentations. In other cases, research articles by the same researcher on a similar subject were selected from academic journals in the field. Whatever the origin of the written articles, they were all anonymised, and Acknowledgments and References were removed. Abstracts that were in a different language from the main text were also discarded from the articles sub-corpora.

Examples in the linguistics subsets were tagged so as to ensure they were distinguishable from the article or from the presentation itself, as shown in (1), an extract from an English article, and (2), a transcript from an English conference presentation.

(1) For instance, consider the following sentences:
<seg ana="exemple">(3) Sam climbed over the wall.</seg>
<seg ana="exemple">(4) The bird flew over the wall.</seg>
<seg ana="exemple">(5) Sam walked over the hill.</seg>

(2) okay and here we have some examples, so <seg ana="exemple">he grassed</seg>, <seg ana="exemple">he boarded</seg>, <seg ana="exemple">he grassed me up</seg> and <seg ana="exemple">he's a supergrass</seg>

We will now describe the two subcorpora for each language, a description which is represented visually in Figure 1 at the end of this section. The word counts were provided by Transcriber, which indicates the number of words in the oral transcripts, or by the text processors we used, or by Antconc (Anthony 2013).

2.1 English sub-corpora

2.1.1 *Science corpus*

The science subset in English comprises 15 articles and 15 transcripts of oral conference presentations. The talks and articles correspond very closely to each other. Eleven of the talks and corresponding proceedings articles were taken from the 7th Novatech 2010 conference, an international gathering focusing on water management and related geological and environmental issues. The remaining four presentations deal with plasma chemistry and oceanography and

were recorded at two other recent scientific conferences. The articles were extracted from peer-reviewed journals or conference proceedings in which the presenters published their findings.

The talks are relatively short (12-15 minutes), amounting to a total of 36,665 words, while the word count for the articles is considerably higher (58,122 words).

2.1.2 *Linguistics corpus*

The linguistic subset also comprises 15 articles and 15 transcripts of oral conference presentations. Thirteen of the transcripts are from the John Swales Conference Corpus, a tribute to John Swales on his retirement. The remaining two are from the Verbes et Complexités Verbales conference organised by the 'Sorbonne Nouvelle' University. Although the conference took place in Paris, the people making the presentations were native speakers of English.

The 15 articles are not the corresponding articles published in the proceedings of the conferences (which is generally the case in the English science subcorpus), but instead were taken from various journals such as *English for Specific Purposes* and *Discourse & Society*. To guarantee a common basis for comparison, we ensured that the articles were written by the same authors as the conference presentations. For example, if we had a recording of John Swales, we then included an article from him in our linguistics corpus. As far as possible, the articles cover the same issues as the conference presentations. This was relatively easy to achieve, as a researcher often tends to give talks and write articles on the same subjects.

The length of the talks vary from 17 to 40 minutes, amounting to a total of 65,687 words, while the articles contain approximately 97,600 words.

2.2 French corpora

2.2.1 *Science corpus*

The French science corpus is made of 15 transcripts and 15 research articles.

The French talks were transcribed from recordings made at a conference organised by FROG (French Researchers in Organic Geochemistry). The presenters of the conference were mainly young researchers completing their PhD or doing postdoctoral studies.

The French science articles were extracted from the journal Quaternaire, a peer-reviewed geology journal publishing articles in French and English[5]. Al-

[5] https://quaternaire.revues.org/

though not an exact match, the French articles were selected so as to be as close as possible thematically to the French conference talks.

The talks are short (15-20 minutes) and represent 37,881 words. The article sub-corpus contain 109,312 words. This ~~huge~~ difference is due to the fact that the articles subset is published in a peer-reviewed journal while conference proceedings articles are usually shorter in length.

2.2.2 *Linguistics corpus*

As in the case of the science corpus in French and English and the linguistics corpus in English, the linguistics corpus in French also comprises 15 talks and 15 articles.

The linguistics talks came from 5 different events. One was recorded at the "La Réanalyse" seminar held at Neufchâtel; four were recorded at the CMLF (Congrès Mondial de Linguistique Française); five came from the Colloque "Quand les genres de discours provoquent la grammaire... et réciproquement". For these ten talks, our corpus includes the corresponding articles, published in the proceedings of the conference.

Three talks were recorded at the workshop on "Cohérence discursive et Prosodie", and two at the workshop on "Verbes et complexité verbale". As we did not have proceedings articles corresponding to these five recordings, we selected articles by the same authors, covering the same topics taken from peer-reviewed journals such as *CORELA*[6].

The average length of the talks is 20 minutes, and the total word count is 65,237. As the articles are in most cases the corresponding proceedings texts, the word count is relatively close to that of the talks: 74,669 words.

2.3 Spanish corpora

2.3.1 *Science corpus*

In Spanish, the science corpus only contains 15 written articles, in the fields of geology. The articles come from various journals such as *Geogaceta*[7], or from conference proceedings such as *CONAMA 9*[8]. Their number of words vary from 1,806 to 15,716, for a total amount of 113,117 words.

It proved impossible to collect a spoken science corpus Spanish for several reasons. Firstly, in the hard sciences more than in the humanities, English is usually adopted as a *lingua franca*. As research results are mostly presented in

[6] https://corela.revues.org/
[7] http://www.geogaceta.com/GEOGACETA/Presentacion%20Geogaceta.htm
[8] http://www.conama9.conama.org/conama9/index.php

an international contexts, this means that there are only a few events at which scholars can present work in their own language. Secondly, some scientists are unwilling to be recorded, or to give access to recordings if they are made. Linguists tend to give a higher priority to making data available, and are more inclined to give their agreement for the constitution of a corpus.

2.3.2 *Linguistics corpus*

Due to the difficulty of obtaining access to recordings, the linguistics corpus in Spanish has a slightly smaller number of texts than the French and English subsets. It comprises 10 talks and 10 articles.

The talks were recorded at three scientific events: the 'Jornada 10 años de Filología Catalana' at the UOC (Universitat Oberta de Catalunya, May 2009); the 'Congreso Internacional de Pragmática del Español Hablado' (Universitat de València, November 2009); and the 'XLI Simposio Internacional de la Sociedad Española de Lingüística' (Universitat de València, January-February 2012).

Of the ten talks, six had corresponding articles that were included in the corpus. For the four talks with no immediately corresponding article, we selected articles addressing the same issue by the same author published before the conference.

Despite the smaller number of texts, the size of the linguistics corpus in Spanish is fairly comparable to the English science corpus: 38,476 words in the oral subcorpus, 64,427 words in the written subcorpus. The recording lengths range from 14 to 40 minutes.

Figure 1 below displays the structure of these corpora and indicates the number of texts (articles or transcripts) and the number of words for each subcorpus.

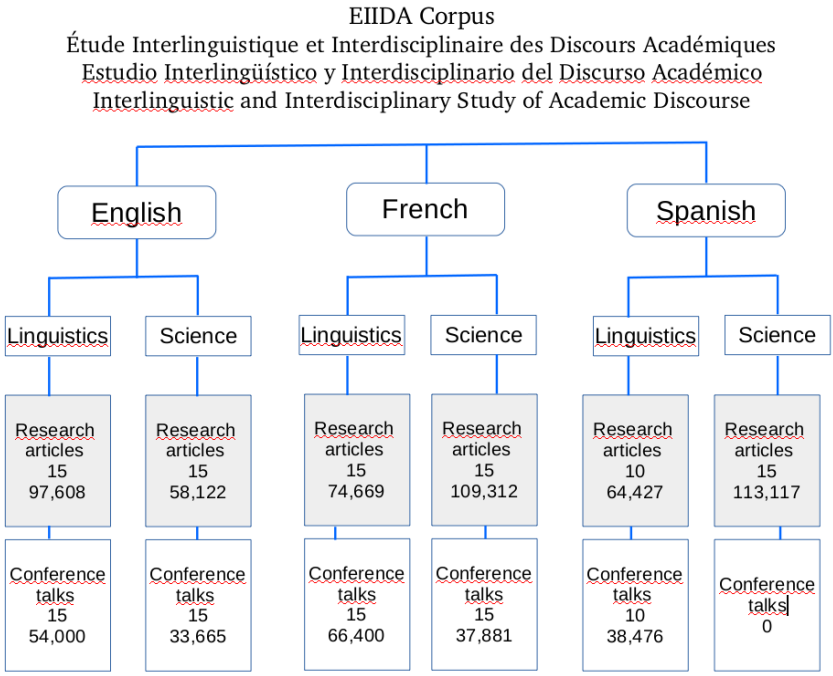

Figure 1. Description of the structure and quantities of texts and words of the EIIDA corpus

The articles in the volume refer to the various sub-corpora with a coding in order to identify the examples taken from them. Two letters indicate the language: FR for French, EN for English, ES for Spanish ("Español"). The next letter indicates the discipline: L for Linguistics, S for Science. The following letter refers to the mode: O for the spoken mode, E for the written mode. Finally, a number identifies the text inside the corpus. For example, 'ES-L-E-08' means that we are referring to the Spanish sub-corpus of Linguistics, and that the text was originally a written document and bears the number 8.

3. The five articles in the volume

The first two articles focus on morpho-syntactic features in the French and English subsets. In "Competing influences: the impact of mode and language on

verb type and density in French and English scientific discourse", Clive Hamilton and Shirley-Carter Thomas contrast lexical verb use in the conference presentations and articles in the science part of the EIIDA corpus. They compare the frequency and types of verbs used to evaluate whether certain verbs are more characteristic of the written or oral mode and whether the mode variable has the same impact in English and French. Initial results suggest that whilst English is more sensitive to the mode distinction, French is particularly sensitive to the issue of language reuse, with lexical variation being more pronounced.

The article by Laura Hartwell, Emmanuelle Esperança-Rodier and Agnès Tutin, entitled "*I think we need…*: Verbal expressions of opinion in conference presentations in English and in French", focuses on the use of opinion markers and their diverse range of functions in conference presentations. In their analysis of both the science and linguistics sections of the EIIDA corpus, the authors contrast the way French and English presenters express and buttress their opinions in the presence of their audience. Although *think* and *penser* were the most frequent opinion verbs, an array of opinion verbs, such as *say/dire* and *believe/croire*, are also employed to express the functions of a general or strong opinion, to negotiate, to suggest a hypothesis or a doubt or to classify data. The results also suggest that there are overall slightly more markers in linguistics, especially linked to a negotiation, than in the harder sciences, where markers of opinion are more often related to an observation.

The following two articles address phenomena at the textual level and compare spoken and written discourse within the same language.

In "Deixis textual y discursiva en el discurso científico-académico oral y escrito", Anna López Samaniego carries out a contrastive analysis of the use of deictic and anaphoric devices in Spanish. She takes into account the body of research on the differences between oral and written modes on the one hand, and between deixis and anaphora on the other hand, to contrast "textual deixis" and "discourse deixis". Broadly speaking, the former points to a referent previously referred to in the text while the latter provides a means to introduce a reference to a process or a propositional content. Textual deixis seems more prevalent in spoken discourse while discourse deixis occurs more frequently in written discourse. Since a conference presentation is a spoken mode which is nonetheless based on written material, questions arise concerning the frequency of the phenomenon, generally attributed to one or other of the modes, such as the type of deixis (textual or discourse), the category of deictic devices (pronouns or lexical items), and the type of anaphora (repetitive or unfaithful). Anna López Samaniego manually analyses and tags the lexical and pronominal items usually

associated with the deictic and anaphoric functions, in order to shed light on the specific choices made in academic discourse.

In her study entitled "La structuration textuelle en discours scientifique : comparaison oral / écrit", Marie-Paule Jacques focuses on text organisation and macrostructure. She investigates the means by which the reader/hearer is guided towards an understanding of this organisation. She compares the written and oral science subcorpora in 15 conference presentations and 15 research articles in the French section of the EIIDA corpus. Her point of departure is the system of headings which, in written articles, both indicates the structure of the article and provides landmarks to the readers. In oral presentations, such as conference presentations, there are no headings, unless some of the slides accompanying the presentation fulfil this role of segmenting and naming the segmented parts. She then examines both the slides and the spoken discourse to highlight the differences and similarities in text structure and in guidance between the oral and written modes.

The final article in the collection focuses on the enunciative dimension of scientific discourse and discusses the way authors and conference presenters refer to their own and other researchers' work.

In their article "Rôles d'auteur et références à d'autres sources" Fanny Rinck, Kjersti Fløttum and Céline Poudat focus on two main issues: authorial positioning and ways of referring to other sources. Basing the study on their extensive past research into these issues in the research article, the authors' aim is to determine to what extent the existing typologies can also be applied in the context of the conference presentation. Their analysis of the French linguistics and science conference presentations in the EIIDA corpus leads them identify a certain number of features specific to conference presentation genres, such as the much higher recourse to personal pronouns, and the first-person pronoun "*je*" in particular, although the impersonal "*on*" remains interestingly the most frequent overall. Another point to emerge is a pronounced tendency for presenters to refer to research projects, rather than to particular articles or authors.

References

Anthony, L. 2013. Developing AntConc for a new generation of corpus linguists. *Proceedings of the Corpus Linguistics Conference (CL 2013)*, July 22-26, 2013. Lancaster University, UK: 14-16.

Barras, C., Geoffrois, É., Wu, Z., Liberman, M. 2001. Transcriber: development and use of a tool for assisting speech corpora production, *Speech Communication*, 33, 1-2: 5-22.

Bazerman, C. 1981. "What written knowledge does: three examples of academic discourse." *Philosophy of the Social Sciences* 11, 3: 361-387.

Bennett, K. 2010. Academic discourse in Portugal: a whole different ballgame? J*ournal of English for Academic Purposes* 9/1: 21-32

Biber, D. 2006. *University Language: A corpus-based study of spoken and written registers.* Amsterdam and Philadelphia: John Benjamins.

Biber, D., Johansson, S., Leech, G., Conrad, S & Finegan, E. 1999. *Longman Grammar of Spoken and Written English.* Harlow: Pearson Education Limited.

Bordet, G. 2011. *Étude contrastive de résumés de thèse dans une perspective d'analyse de genre.* Thèse de doctorat non publiée. Université Paris Diderot-Paris VII. Accessible en ligne sur TEL (thèses-en-ligne) : http://tel.archives-ouvertes.fr/tel-00650637/fr/ (accessed December 9, 2016).

Carter-Thomas, S. & Rowley-Jolivet, E. 2001. Syntactic Differences in Oral and Written Scientific Discourse: The Role of Information Structure. *ASp* 31-33: 19-37. http://asp.revues.org/1752 (accessed December 9, 2016).

Clyne, M. 1996. *Inter-cultural communication at work. Cultural Values in Discourse.* Cambridge: Cambridge University Press.

Fløttum, K. (ed.) 2007. *Language and Discipline Perspectives on Academic Discourse.* Cambridge: Cambridge Scholars Publishing.

Fløttum, K., Dahl T. & Kinn, T. 2006. *Academic Voices. Across languages and disciplines.* Amsterdam and Philadelphia: John Benjamins.

Hyland, K. 2000. *Disciplinary Discourses: Social Interactions in Academic Writing*. London: Longman.

Josselin-Leray, A. 2005. *Place et rôle des terminologies dans les dictionnaires généraux unilingues et bilingues. Etude d'un domaine de spécialité : volcanologie.* Thèse de doctorat non-publiée, Université Lumière Lyon II.

Kübler, N. & Volanschi, A. 2012. Semantic prosody and specialised translation, or how a lexico-grammatical theory of language can help with specialised translation. In Boulton, A, Carter-Thomas, S. & Rowley-Jolivet, E. (eds). *Corpus-Informed Research and Learning in ESP: Issues and Applications.* Amsterdam and Philadelphia: John Benjamins.

Molino, A. 2011. A Contrastive Study of Knowledge Claims in Linguistics Research Articles in English and Italian. *ESP Across Cultures*, 8, 89-101

Robles Garrote, P. 2013. La conferencia como género monológico: análisis macroestructural en español e italiano. *Boletín de filología*, 48 (1), 127-146. http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-93032013000100006&lng=es&nrm=iso (accessed December 5, 2016).

Robles Garrote, P. 2016. Aportaciones de la Lingüística de Corpus al estudio de la conferencia como género académico de divulgación científica. *Chimera* 3 (1). https://revistas.uam.es/index.php/chimera/article/view/2282 (accessed November 22, 2016).

Swales, J. 2004. *Research Genres.* Cambridge: Cambridge University Press.

Tutin, A., Grossmann, F., Falaise, A. & Kraif, O. 2013. Autour du projet Scientext: étude des marques linguistiques du positionnement de l'auteur dans les écrits scientifiques. *Texte et corpus* 4: 333-349.