1. From corpus linguistics to applied corpus linguistics?

Over the last 25 years there have been developments in corpus linguistics both on the part of European as well as North American researchers. Although not initially with a pedagogical goal in mind but with a research end, most corpus linguistics projects undertaken lately have recognized the necessity of bringing in a pedagogical aim towards the teaching and learning of a language, resulting in what I will call here applied corpus linguistics. European researchers in the study of corpus linguistics have been particularly active specially in the United Kingdom and Scandinavia as the prolific publications and creation of several corpora show. Take for instance the COBUILD project (earlier Bank of English (BoE project)) which started in the year 1991 and where the collaboration of John Sinclair towards a pedagogical lexical computing orientation was seminal (cf. Sinclair 1987 and 1991), it covers over 450 million words of British English (spoken and written), it is basically a specialized corpus devoted to single genres. Another important European corpus is the BNC (British National Corpus), a 100 million word collection of samples of mainly written but also spoken English language from a wide range of sources which was also started in the year 1991. Also relevant among corpora is the ICE (International Corpus of English) project, which began in 1990 with the main aim of collecting material for comparative studies of English worldwide. The British component of the corpus was published in 2001 with 1 million words. At the moment 13 research teams are participating in this project following Greenbaum’s (1991) initial idea. Also important is the CIC (Cambridge International Corpus); started in 1992, it holds 600 million words but this time including British and American English discourse. Within the ICI we can find corpora such as the British CANCODE (Cambridge and Nottingham Corpus of Oral Discourse in English) or the CAMSNAE (Cambridge Corpus of Spoken North-American English). The ICI tries to help writers in their task of writing books for learners of English and it is only available to authors and writers of books working for Cambridge University Press publishing house. More recently compiled is the BASE corpus project (The British Academic Spoken English Corpus), developed at the University of Warwick under the leadership of Hilary Nesi and Paul Thompson; it consists of 160 lectures and 40 seminars recorded in a variety of university departments, this corpus aims to be a counterpart of the North-American MICASE (Michigan Corpus of Spoken Academic English).

As to North-American scholars, there have also been remarkable developments in corpus linguistics, we could name Douglas Biber et al. at Northern Arizona University with the Longman Grammar of Spoken and Written English (1999) (the majority of co-authors being European, though), Michael Barrow at Rice University and other corporist groups at The University of California or the University of Pennsylvania. Standing out we find the ANC (American National Corpus), started in 1990, it includes a massive electronic collection of American English including texts of all genres and transcripts from spoken data that is still in progress. When completed, the ANC will contain a core corpus of at least 100 million words, comparable across genres to the BNC described above. Undoubtedly significant in North-
America is MICASE (the Michigan Corpus of Spoken Academic English) (Simpson et al. 2002) within the field of EAP. It started in 1997 at the University of Michigan and was designed to digitally audio-record and transcribe about 1.8 million words of contemporary academic speech at a major American research university, covering speech events ranging from freshman advising to doctoral defences along with traditional university lectures and class discussion. Among other specialized well-known spoken corpora we find the Cambridge and Nottingham Business English Corpus (CANBEC) and Corpus of Spoken Professional American English (CSPAE), both compiling professional spoken genres.

The majority of the corpora mentioned above have exerted in compiling spoken discourse samples (whether general or more specialized- eg. professional or academic). In this line, the British BASE and the North-American MICASE, both focusing on different spoken academic genres have been seminal for corpus linguists in the analysis of academic discourse. However, these corpora are examples of what Römer (2009) calls ‘indirect pedagogical corpus applications’ opposed to ‘direct pedagogical corpus applications’ that “hand on for learners and teachers (data-driven learning or DDL (Johns 1986,1994))” (Römer 2009: 20). The aim of this paper is to find a bridge between the indirect use of spoken academic corpora and the pedagogical applications addressed not only to English language students but also to lecturers/teachers that want to use English as a vehicular language for their lessons among higher education institutions.

2. English for Academic Purposes in the European Higher Education Area (EHEA)

To date, there is some relevant research within corpus linguistics that has dealt with the analysis of spoken language in the EAP field (Swales & Malczewski 2001; Mauranen 2001, 2003; Biber et al. 2002; Swales & Burke 2003; Biber 2003; Flowerdew 2003). However, studies up to now have approached the EAP analysis from two different perspectives: a) the instructional-pedagogical academic discourse analysis; b) the research-oriented academic discourse analysis. The former usually contributes to the how of teaching shedding some light to the improvements of teaching and learning practices, whereas the latter contributes to the what of teaching and learning bringing new knowledge into the field. Even though the distinction between these two approaches is clearly seen, the border line between them can usually be blurry. Spoken academic researchers are frequently language lecturers/teachers at the same time and therefore, they can largely benefit from the results of their research-oriented academic discourse analysis. In fact, these two different analysis orientations should benefit from one another; corpus linguistics research-oriented analysis in the field of EAP should try to aim for pedagogical/instructional purposes, finding a place of confluence among these two analysis.(see Figure 1).

![Figure 1. Confluence between the instructional-pedagogical and the research-oriented academic discourse analyses.](image-url)
At this point we could address the questions: “why the study of EAP?”, “why has EAP gained so much importance among corpus linguistics researchers?”.

Crucial for the survival of any language is the power of its scientific production and the teaching of that language in classrooms. Scientific discovering and technology can nowadays largely be spread worldwide thanks to a single global language for communication, English. This fact has led many scientists and professionals from non-English speaking communities to choose English as the unique and supreme scientific language to be used in academic and professional settings.

Students at universities are from an ‘international’ background (Benson 1994: 182). Although most of the English higher education teaching has been done in English speaking countries, such as Britain and the U.S. and it is also in these countries where the academic discourse started to be analysed (Mauranen 2001, Swales & Malczewski 2001, among others), English has dramatically expanded to other countries as the result of internationalisation. Globalising processes are forcing higher education institutions to internationalise. Internationalisation is no longer confined solely to aspects of knowledge, it also involves extensive student and staff exchanges, the joint development and implementation of new educational programmes, where two or more institutions from different countries are involved and recently, a worldwide focus on student and staff recruitment (Wilkinson 2004).

Due to this internationalisation of university lecturing mainly in Europe (e.g. Erasmus-Socrates programmes, Leonardo da Vinci projects, etc.), and also in the United States (e.g. postgraduate education), academic discourse in English has become a main concern in higher education institutions, focusing especially on spoken academic discourse, which should be looked at as paramount for both students and faculty. Students attending an academic event need to listen and understand in order to be able to take notes and hold concepts. On the other hand, faculty’s academic research life involves not only reading English publications, but attending as well as presenting papers at conferences, or lecturing in their own or other universities where English is used as the primary language. Both students and lectures at university have to use different academic genres, these genres have been studied and classified by some authors (Giménez 2000, Bellés-Fortuño & Fortanet 2004, Fortanet 2005).

A thorough classification of oral academic genres according to criteria of purpose, rather than to interaction between speaker and listener, has been provided by Fortanet (2005), who distinguishes three main levels of academic genres: classroom, institutional and research genres. The first level of academic genres(classroom genres) has a clear pedagogical-instructional aim, whereas the other two, although mainly research-oriented, can easily be instructional when approached from an applied corpus linguistics analysis.
Among some relevant studies on oral genres I would like to point out the one carried out by Räisänen (1999) on conference genres. She uses a corpus of engineering conferences that she classified as Conference Abstract (CA), the written Conference Proceedings Paper (CCP) and the oral Conference Presentation (CP) to carry out a quantitative study. Räisänen highlights the increasing mobility “over disciplinary, national and professional borders” (1999: i). and she emphasizes the idea that a substantial part of the world’s scientific dissemination is given through academic conference genres. Other academic genres recently analysed are guest lectures (Crawford 2004) where the lecturers are frequently recruited from different academic, professional and cultural backgrounds “adding that ‘international flavour’ which is now an essential ingredient in the era of global education”. (Crawford 2004: 91).

Already in the late 20s Mauranen (1998, 2001) pointed out on the important of English academic discourse among international academic institutions. She detected that academic genres are not isolated units, but chainlike formations within a single genre or across different genres:

Academic speech events are commonly organised in chainlike formations either within one genre (e.g. a lecture course or a linked series of seminars) or across different genres (e.g. lectures, followed by examinations; supervision and consultations, followed by a thesis defense and its important prior text, the thesis itself) (2001: 166).

In general, academic oral genres have aroused the interests of researchers, specifically the genre of lecture, being one of the most important genres within spoken academic discourse. A great part of university discourse research focuses on the lecture (Johns 1981, Richards 1983, Benson 1989). Knowing the best way for students to internalise and comprehend lecture content seems to be paramount for university success; that is why there is some research on spoken academic language centred on different aspects of lectures (Flowerdew 1994,

[…] can indicate to teachers and course designers what linguistic and discoursal features learners need to be familiar with in order to understand a lecture and what, therefore, should be incorporated into ESL courses. In addition a knowledge of the linguistic/ discoursal structure of lectures will be of value to content lecturers in potentially enabling them to structure their own lectures in an optimally effective way.

2.1. Academic English in the EHEA

In June 1999 the Bologna Declaration was signed. It was some years after the Bologna Declaration when the guidelines for the creation of a European Higher Education Area (EHEA) were decided, emphasizing on the promotion of the mobility of students and staff, the harmonization of European degrees, the establishment of two main cycles (undergraduate and graduate) and a common credit system for all the European universities.

One of the most important factors of the EHEA creation for the field of EAP is the emphasis on increasing students and staff mobility, which is more feasible and realistic when there is one language or more in common. Thus, the necessity towards (academic) language learning promotion is real. The European Commission has encouraged higher education institutions to promote language learning and has given several examples such as e-learning, distance learning or the use of a foreign language as the instruction medium in the classrooms, what has been coined as CLIL (Content Language Integrated Learning).

It is precisely this last factor the most controversial one. CLIL has aroused the interests of not only European university commissioners but also lecturers and even secondary school teachers. The European commission never sated that English was the language that had to be used in the classrooms as the instruction medium; however, as we already aforesaid in this article, English has gained the status of lingua franca and has become the language of the scientific and research community.

Another change that EHEA premises have brought to the academic world is the more and more widely extended incorporation of the Information and Computer Technologies (ICTs) into the university settings. The result is the emergence of new modes and types of lecturing, as for example blended learning, incorporating many different learning styles combining virtual and physical resources, where technology, instructor and peer interaction blend (Thorne 2003, French et al. 2003). As well, this brings the emergence of new academic genres with the incorporation of electronic visuals or the access to virtual classrooms, where you can find materials for the course, seminar, etc. on the internet, the possibility to be involved in a classroom forum or to submit exercises for correction without the presence of a lecturer, (e.g. Virtual Learning Environment). Take for instance the e-learning education (strictly related to Open University institutions) where classroom attendance is not necessary and therefore new modes of interaction and participation are used (on-line office hours and/ or tutorials, etc.). And of course, one of the possibilities that technology has brought into the academic setting is the video-conferencing or teleconferencing, this latter is a live genre, where the presence of the lecturer changes its original conception. These forms of teaching are already used by some English language lecturers as resources for their teaching practices.

Other changes do not affect the how of teaching in higher education setting but they are strictly related to what to teach. Curricula and syllabi have to adapt to the Tuning project claimed competences, this project:
 (...) is one of the few projects in Europe that actually links the political objectives set in the Bologna Declaration 1999 to the higher education sector. Tuning is a project developed by and meant for higher education institutions. (European Tuning Project: http://unideusto.org/tuning/)

This project does not focus on educational systems, but on educational structures and content of studies. The project develops competences distinguishing between subject specific and generic competences. Generic competences or transferable skills have an outstanding position within the European higher education setting since they can prepare students for the future role in society both, in the labour market and as citizens. The generic competences according to the European Tuning Project can be of three types:

- Instrumental competences: cognitive abilities, methodological abilities, technological abilities and linguistic abilities;
- Interpersonal competences: individual abilities like social skills (social interaction and co-operation);
- Systemic competences: abilities and skills concerning whole systems (combination of understanding, sensibility and knowledge; prior acquisition of instrumental and interpersonal competences required).

Curricula and syllabi have to aim at promoting the acquisition of competences as well as the learning of contents. Therefore, and in the case of language lecturers efforts have to be addressed towards the right selection of language contents (Perez-Paredes 2003:1) and towards “student-centred higher education and away from teacher driven provision” (London Communiqué, 2007) in an attempt to acquire generic and subject specific competences.

The question to be addressed now would be: how does the EHEA policies affect EAP? And the answer is that language promotion has to be stronger than in older times. English as a globalised *lingua franca* used all over the world has become a main concern among European universities: although English has a prior position, European universities are also trying to promote other minor or more local European languages in an effort to promote multilingualism. English in the academic world and therefore EAP has an important role especially in the CLIL projects at universities, where the necessity for teacher training courses in languages is a must (Räisanen and Fortanet-Gómez, 2008: 20). For this the collaboration between content and language teachers at a university level seems to be necessary. It is not only content teachers but also language teachers who have to make the effort to approach the other part. A good way to make content teachers more conscious about academic genres and to help them improve the specific English language content in their discipline is through specific training courses, and it is language teachers who are most suitable to teach these courses (Fortanet-Gómez and Bellés-Fortuño, 2008). However, as far as I know, very scarce or few pedagogical materials available in the market have specialised on EAP and only a few might be based on real language realisation extracted from corpora. Most of the current instructional materials available are based on written academic discourse rather than spoken, and they are presented in the form of textbook language. Having the availability to work on spoken academic corpora, the purpose of this paper is to show how a spoken academic corpus created *ad hoc* can be used with pedagogical objectives for the creation of instructional materials that can be of great help in the field of EAP teaching and learning.
3. A pedagogically-oriented spoken academic corpus analysis: The MASC example

In an attempt to deepen the analysis of spoken academic English and aiming at different spoken academic genres, the group of research GRAPE (Group of Research on Academic and Professional English, http://www.grape.uji.es) at Universitat Jaume I (Castellón, Spain) has compiled and digitalised the MASC corpus (Multimodal Academic and Spoken Language Corpus). The compilation of MASC started in 2004 and it is still in progress. It consists of spoken academic events in Spanish and mostly in English that have taken place at a university context such as lectures, conference presentations, seminars, guest lectures, and plenary lectures. MASC is a multidisciplinary corpus that includes events from the fields of Linguistics, History, Law, Business, Marketing, Biology, and Chemistry. These recordings are accompanied by other resources and materials that have also been included as part of the corpus such as PPT. slides, handouts, and any other resources delivered while the academic event recording.

The purpose of the MASC corpus is two-fold. On the one hand, the project of compiling a multimodal spoken academic corpus aimed at offering a research corpus in order to carry out linguistic and pragmatic analyses on the use of the English language in academic settings in Europe, and the analysis of another European language, that is, Spanish, opening possibilities to contrastive linguistics analysis. On the other hand, the GRAPE research group also aimed at using the advantages that a real language corpus on academic events has, to exploit them in order to create pedagogical materials that could be of interest, not only for undergraduate but also for postgraduate students and university lecturers. The members of the GRAPE were conscious of the changes taking place within higher education institutions in Europe and the need to promote English language learning to be able to go ahead with the EHEA premises that encouraged students and staff mobility as well as the adaptation of the language learning curricula according to the competences of the Tuning Project aforementioned.

With this in mind, the GRAPE group started working together in two different projects: one of the results of that co-work and cooperation is the publication of the book Hablar Inglés en la Universidad: Docencia e Investigación (Bellés-Fortuño et al 2008). The book was addressed to university lecturers with general knowledge of English that would like to improve their English language oral skills in order to use English as the instructional language in their subjects as well as in research settings such as conference presentations. At the same time, the group began to work on a way to exploit the multimodal academic corpus (MASC) for pedagogical purposes. This second project took longer than the first one and it is still work in progress but the result can partially be seen at GRAPE Online Activities for Academic English (Ruiz-Madrid, M.N. and Querol-Julián, M. 2008, http://www.grape.uji.es/activities/pagina%201/index.html) (see Figure 3 below).
Figure 3. GRAPE Online Activities for Academic English (Ruiz-Madrid, M.N. and Querol-Julián, M. 2008)

The activities embrace research and teaching purposes aiming at the improvement of spoken and written English academic language, they are divided into three main sections: 1) Writing a research article, 2) Teaching in English, 3) Participating in conferences (see Figure 3). These activities can be used by undergraduate, graduate or postgraduate students that want to improve their English academic language when writing subject projects in English, giving classroom presentations or even writing their MA thesis. Lecturers and researchers within the EHEA can also benefit from the content of these EAP on-line activities, when using English as the instructional medium for their subjects or when trying to participate in an academic event with a communication in English.

Section 3 of the activities is a clear example of how multimodal spoken corpora can be used as the source for the elaboration of activities that can be applied for English Language Teaching (ELT).
In the activity shown in Figure 4 above, behavioural aspects have been categorised in order to distinguish between two main conferencing styles in English (e.g. interacting with the audience, being serious, using jokes etc.). Other non-linguistics features such as body movement, gestures (kinesics) have been studied in order to show behavioural trends when giving an academic conference presentation in English. This activity example is the results of thorough video and audio examination of different academic conference presentations included in the MASC corpus. Activities such as the one presented here could not have been possible without the analysis of a multimodal corpus; transcripts are not enough to develop instructional materials that take into consideration the three aspects of human communication: verbal, kinesics and paralinguistic.

Going ahead with the exploitation of the MASC multimodal corpus, the activity in Figure 5 below, is an example of how a video recording can be a potential instructional material which, in this particular case, is used for the teaching and learning of EAP. Activity 6 of the GRAPE On-line activities contains two different clips that show the two different conferencing styles categorised in activity 5 before (see Figure 4). The learner watches these two clips recorded from real conference presentations, with real participants in authentic settings. The learner is asked to write down his/her own notes on his/her impressions and beliefs about the two clips shown. This type of instructional activity perfectly fits in the Tuning Project competences enhancing abilities such as cognitive, methodological, or individual among others, and promoting language learning from a more student-centred approach.
The pedagogical materials developed from the MASC corpus would like to pinpoint the importance of multimodal corpora in corpus linguistics applied to LT. Thanks to a more global multimodal perspective, applied corpus linguists could easier bridge the gap between discourse analysis using corpus linguistics and language teaching. In the next section, I will discuss on the characteristics, advantages, disadvantages and the compilation difficulties of multimodal corpora.

4. Spoken academic corpora: Towards a multimodal approach

Corpora such as the MASC presented here in this paper is just an example of an ad hoc EAP corpus that tries to incorporate other means of analysis different from the transcripts. Most very well-known and used spoken academic corpora or other non-academic corpora mentioned in section one of this paper consist of a compilation of transcribed language events, sometimes with tagging possibilities. However, as shown in the previous section, transcripts are usually not enough when exploiting corpora for researching or pedagogical aims. Spoken discourse, and in this case academic spoken discourse, is very complex; there are other non-linguistic features that take place in the realization of the event such as non-verbal communication. In fact, human communication involves more than only verbal elements, there are also kinesics (body movement, gestures, face expression) and paralinguistic (intonation, rhythm, pitch of the voice, etc.) features (Poyatos 2004). Corpus linguists facing research have to think before hand about the analysis to follow and start to consider aspects beyond the linguistic features such as intonation, speech rate, speaker’s idiolectal variation, gestures, etc., for the analysis of spoken discourse. To address the analysis of paralinguistic features, kinesics and prosodies, it is necessary to adopt a multimodal perspective of analysis combined together with a corpus linguistics
methodological approach (Querol-Julián, M. & Bellés-Fortuño, B (in press)). When approaching applied corpus linguistics towards language teaching these non-linguistic features mentioned above should also be taken into consideration as well as other cross-cultural features among universities worldwide.

In this line, spoken corpora are samples of real language as opposed to textbook language, unfortunately being this last the most extended one among pedagogical-instructional English materials up-to-date. The MASC corpus presented here is a tri-modal combination that amalgams audio, video and text (transcript). These are traits that undoubtedly count as advantages of using multimodal spoken corpora whether for research or pedagogical purposes. However, one of the most commonly encountered difficulties of compiling a multimodal corpus is the possibility of recording an event. It is not easy to get access to seminars, conference presentations, lectures, etc. when you are not part of them. The corpus linguist recorder is usually seen as an intruder. Permissions have to be asked but the answer is not always positive. On a second instance, it is not enough with having recorded the event; if this is going to be used as part of a corpus and is going to be exploited for research or pedagogical purposes, it is compulsory to ask participants for consent, unless we want to fail into a criminal act. These legal aspects are even emphasized in the case of the United States where individual privacy is exalted.

Even counting on the disadvantages and difficulties of compiling a multimodal spoken corpus, the possibilities of having a multimodal corpus, whether ad hoc or larger samples of corpora, are infinite when using corpora for research and teaching purposes, and the result is worthwhile. Moreover, I consider that a multimodal spoken corpus offers more possibilities to bridge between corpus linguistics and language teaching. The corpus linguist has to be seen as an applied corpus linguist able to guide language learners and lecturers/teachers, as well as novice corpus linguists in the selection of language contents and researchable linguistic and non-linguistic parameters.

5. Conclusion

This article has looked at the importance of spoken academic corpora and the relationships with language teaching and research. I have revised here the most well-known spoken corpora in general, academic and professional settings. I aimed to picture the current situation in the European Higher Education Area (EHEA) after the Bologna Process and the subsequent Tuning Project with the development of competences, a more student-centred approach and language learning promotion. The globalisation process has stressed the consideration of English as a lingua franca and therefore English has become a key language among academic settings. In this line, English for Academic Purposes (EAP) has gained the interests of language researchers who have seen the necessity of directly analyse academic language use. This paper has also tried to show how corpus linguistics, or more concretely applied corpus linguistics, and the exploitation of spoken corpora can easily adapt to the current university situation in Europe, becoming potential pedagogical corpus applications that provide samples of real and authentic language. By means of a multimodal ad hoc compiled corpus such as MASC, I have shown the research and pedagogical applications in the field of EAP. The activities presented are also an example of linguistic and non-linguistic features analysis transformed into language instructional materials. The advantages of a multimodal corpus that amalgams video, audio and text (transcript) have been shown. However, the compilation of a spoken multimodal corpus can arouse some difficulties such as the access to record the events or the participants’ consent to be recorded.
This article invites applied corpus linguists to consider EAP as their field of study in order to enlarge the possibilities of having larger spoken academic corpora that can become not only discourse researchable corpora but also potential pedagogical corpus applications. It goes without saying that the boundaries between research and pedagogical purposes is very thin in academic settings, it is up to applied corpus linguists to find the joint between research and language teaching in the future.

References


Appendix: A selection spoken on-line corpora

- **British Academic Spoken English (BASE) Corpus**
  [http://www.rdg.ac.uk/AcaDepts/ll/base_corpus/](http://www.rdg.ac.uk/AcaDepts/ll/base_corpus/)
- **The Bergen Corpus of London Teenage Language – (COLT)**
  [http://torvald.aksis.uib.no/colt/](http://torvald.aksis.uib.no/colt/)
- **The Brown Corpus**
  [http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM](http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM)
- **Lancaster/IBM Spoken English Corpus (SEC)**
  [http://nora.hd.uib.no/icame/lanspeks.html](http://nora.hd.uib.no/icame/lanspeks.html)
- **Michigan Corpus of Academic Spoken English (MICASE)**
  [http://www.lsa.umich.edu/eli/micase/index.htm](http://www.lsa.umich.edu/eli/micase/index.htm)
- **Cambridge and Nottingham Business English Corpus (CANBEC)**
  [http://www.cambridge.org/elt/corpus/cic.htm](http://www.cambridge.org/elt/corpus/cic.htm)
- **Corpus of Spoken Professional American English (CSPAE)**