Corpora in language teaching and learning

Kate Beeching

Electronic version

URL: <https://journals.openedition.org/rdlc/1672>

DOI: 10.4000/rdlc.1672

ISSN: 1958-5772

Publisher

ACEDLE

Electronic reference

Kate Beeching, "Corpora in language teaching and learning", *Recherches en didactique des langues et des cultures* [Online], 11-1 | 2014, Online since 07 January 2014, connection on 16 February 2023. URL: <http://journals.openedition.org/rdlc/1672> ; DOI: <https://doi.org/10.4000/rdlc.1672>

This text was automatically generated on 16 February 2023.

Creative Commons - Attribution-NonCommercial-NoDerivatives 4.0 International - CC BY-NC-ND 4.0
<https://creativecommons.org/licenses/by-nc-nd/4.0/>

Corpora in language teaching and learning

Kate Beeching

Introduction

- 1 As Aijmer (2009: 2) points out:

Discussions of the pedagogical implications of corpora can take two forms. They focus on the use of corpora in the classroom. Moreover they deal with the use of corpora for applied linguistics research in particular the use of learner corpora to get a better picture of how advanced learners write and speak.

- 2 The ways that (pre-existing) corpora can be exploited in the classroom is addressed elsewhere (Aijmer, 2009; Frankenberg-Garcia, this volume). The current article reverses the assumption that the corpus comes first – and that it can be exploited in the classroom – in order to put primacy on language teaching and learning, and the corpora which emerge if the learner takes centre-stage. The article suggests that the relationship between language teaching and corpora can take at least three forms, and these are illustrated in its three main sections: from syllabus to corpus (for general language courses), from corpus to syllabus (for specialist language learners) and learner corpora (in this case a role-play corpus). First of all, however, in order to answer the question « What is a corpus in language teaching and learning? », we have to define what we mean by a corpus and what we mean by language teaching. It is to these definitions that I turn my attention in the following sub-sections. The first question is perhaps more straightforward than the second.

What is a corpus?

- 3 McEnery *et al.* (2006: 4) provide the following definition of what is normally meant by the word corpus:

The term corpus as used in modern linguistics can best be defined as a collection of sampled texts, written or spoken, in machine-readable form which may be annotated with various forms of linguistic information.

- 4 Corpora provide a body of data which is representative of the language at a particular point in time. The *British National Corpus* is perhaps the best-known example of a corpus with representative texts gathered from the 1980s-1993. Arising out of Quirk's *Survey of English Usage*, the BNC contains a 100-million-word text corpus of samples of written and spoken English from a wide range of sources. The project to create the BNC involved the collaboration of three publishers (with Oxford University Press as the lead collaborator, along with Longman and Chambers), two universities (Oxford and Lancaster) and the British Library. The creation of the BNC started in 1991 under the management of the BNC consortium and the project was finished by 1994. There have been no additions of new samples after 1994, but the BNC underwent slight revisions before the release of the second edition BNC World (2001) and the third edition BNC XML (2007). A new project has been launched, funded by the ESRC, in a collaboration between Lancaster University and Cambridge University Press, to create a further spoken corpus (<http://cass.lancs.ac.uk/>), the BNC Spoken Corpus 2014.
- 5 90% of the BNC is made up of samples of written language use. These samples were extracted from regional and national newspapers, published research journals or periodicals from various academic fields, both fiction and non-fiction books, leaflets, brochures, letters, essays written by students of differing academic levels, speeches, scripts and many other types of texts. The remaining 10% of the BNC is composed of samples of spoken language. The spoken corpus consists of two parts: one part is demographic, containing the transcriptions of spontaneous natural conversations produced by volunteers of various age groups, social classes and originating from different regions. The second part comprises « context-governed » samples such as transcriptions of recordings made at specific types of meetings and events. These are subdivided into Business, Leisure, Education and Institutional, and the latter contain extracts from courts of law, amongst other institutional contexts. The original recordings transcribed for inclusion in the BNC have been deposited at the British Library Sound Archive and the sound-files are now being linked to the electronic transcriptions by researchers at the University of Lancaster and may be accessed via <http://bncweb.lancs.ac.uk>.
- 6 As the BNC is a large mixed corpus which set out to be representative of British English as a whole, it is unsuitable for the study of highly specific text-types or genres, as any one of them is likely to be inadequately represented. Those wishing to explore their specific conventions of particular genres would do better to compile a small corpus including only texts of those types.
- 7 To sum up this section, we can say that a corpus is generally understood to be a collection of:
 - authentic texts (including transcriptions of spoken data) which have been
 - sampled so that they are
 - representative of a particular language or variety of a language, and which are
 - machine-readable.

What do we mean by language teaching?

- 8 Before turning to the more general question of what might be covered by the broad term *didactique des langues* (language teaching and learning), let us consider the two general ways in which corpus material can be used in language teaching. Firstly, publishers and researchers can use corpus samples to create language-learning syllabuses and materials. The learners themselves do not have access to the corpus but the corpus informs the way that language is presented to students in learning materials. Secondly, the analysis of corpus data can be incorporated directly into the language teaching and learning environment. With this method, language learners are given the opportunity to categorize language data from the corpus and subsequently form conclusions about the patterns and features of the target language from their categorizations. This method involves a greater amount of work on the part of the language learner and is referred to as « data-driven learning » or as « hands-on » corpus use (see Frankenberg-Garcia, this volume). Thirdly, in a « hands-off » approach, a tutor can use corpus examples to illustrate particular language points. This demands considerable insight and work on the part of the tutor. A large representative corpus like the BNC is particularly useful as a reference source when studying the use of individual words in different contexts, so that learners become familiar with the different ways to use particular words in context. As Hunston (2009) points out, however, explanations of this sort only accentuate our perceptions of the complexity of language rather than providing the type of straightforward « rule » that learners crave. Arguably, a representative corpus can show what company a word keeps (its collocations) and also its frequency, so that translators, for example, could select a word which is equally frequent in the target language as in the source language.
- 9 The theory and practice of language teaching and language learning is a vast field which it would be inappropriate to attempt to encapsulate here. In general, we can say that, since the 1960s, the field has moved from a focus on grammar-translation (the aim of learning a language was to read its literature) towards an interest in communicative competence (the aim is to be able to function practically in daily interactions of different types). This communicative revolution took hold in the 1970s and 1980s and there was a strong focus on « authentic » language which I will come back to in Section 1.3. Most language teaching materials take an eclectic approach which covers the acquisition of grammar and vocabulary in everyday and thematic situations which are relevant to the student body in question. Most syllabuses also highlight the four skills (listening, speaking, reading and writing) and are adapted to the level of the student (beginner, intermediate or advanced, to give but the broadest categories). Over the last 40 years, the focus in language syllabuses has shifted from grammar to situations, themes, functions and notions, to task-based learning, the lexical syllabus and more learner-centred approaches built around needs analyses. It was, however, back in the 1980s that these developments were beginning to shape language teaching and learning. As Nunan (2007: 10) suggests:

The 1980s was the decade in which the principles of communicative language teaching, which had evolved in the preceding decade, began to gain traction in the classroom. We began to see curricula and materials that took as their point of departure an analysis of learners' communicative needs, rather than inventories of language systems. Needs analysis procedures and needs based programming emerged to support the development of differentiated curricula to meet different learner needs.

- 10 Theories of language learning and corpora came together particularly forcefully in the early 1980s when I began my French text-book writing career: the focus was on « authenticity », the provision of samples of language which were produced in the target culture in real communicative situations. This was due in part to a reaction against the unnatural model sentences favoured by the grammar-translation approach which were fabricated to illustrate particular aspects of structure. As far as I was aware, there was no material of an authentic sort of this type which was easily available for adaptation for the teaching of French in schools, though researchers/teachers from the University of Reading had been trail-blazers in this area in creating the *Enquête Sociolinguistique sur Orléans Corpus* in 1968.

What is authenticity?

- 11 Through the 1970s and on into the 1990s and the first decade of the current century, the team of researchers at the Crapel (*Centre de Recherches et d'Applications Pédagogiques en Langues*) in Nancy were at the forefront of discussions of the interaction between « authenticity » and what might be « authentic » for a language learner (Duda *et al.*, 1972; Holec, 1974, 1990; Boulton, 2009; Duda & Tyne, 2010). Authentic texts are generally described as spontaneous, spoken, non-scripted texts produced in a real communicative situation. For a learner, however, in order to be « authentic », a text must be both relevant and accessible. As Widdowson (1998: 714-715) pointed out, language learning tasks "must take account of the interests, attitudes, and dispositions of the learners... the appropriate language for learning is language that can be appropriated for learning". Rühlemann (2008: 685) remarks that:

authenticity in Widdowson's sense does not depend on the text being invented by a materials designer or captured in a spoken corpus, but on the successful mediation through careful selection and motivating teaching.

- 12 The notion of authenticity, then, is not something which is inherent to the text but is a negotiation between the teacher and the learners.
- 13 The remaining sections of this article illustrate three very different types of spoken corpora which I have created and used for language teaching and learning: firstly, in Section 2, a corpus which resulted from the development of a series of text-books published by Oxford University Press and Cambridge University Press between 1981 and 1993 and aimed at the different needs of British students studying for their GCSE and A-level examinations in French; secondly, in Section 3, a small corpus of French for Specific Purposes (to meet the needs of student gardeners and golf greenkeepers) created in 1995; and, thirdly, in Section 4, a corpus composed of a collection of role-plays in English, designed to help researchers and learners better understand the role of pragmatic markers in spoken discourse.

From syllabus to corpus

Figure 1 – Covers of text-books of French (1980-1993).

- 14 From 1980 to 1993, on the back of my MA dissertation on the exploitation of authentic resource materials (ARMs) and an article on this topic (Beeching, 1982), I was commissioned by Oxford University Press and Cambridge University Press to write a series of French text-books which addressed the need for samples of spontaneous spoken (and written) French for students preparing for GCSE and A-level¹ examinations in the UK. The first, *Vrai de Vrai! Authentic French for Listening and Reading Comprehension* (Beeching, 1985), focused on receptive skills and was tailored to the GCSE syllabus with chapters devoted to transactions in shops and at the station, hotel and post-office, but also discussions about clothes and fashion, music, feminism and nuclear power. Authentic materials were expected to be not only relevant but also motivating for students. As I say in the introduction:

The exploitation of authentic material is being increasingly recommended both because of the obvious relevance of such things as menus and tourist information brochures and for the effect they have on students » motivation. It has been shown that motivation is one of the factors – if not the factor – which is crucial in learning a foreign language. The national criteria for GCSE French stress that examination tasks should be of value outside the classroom and that the material used should be carefully selected authentic materials. (Beeching, 1985: 3).

- 15 Not only were the recordings collected for this particular purpose but the exploitation of them is carefully graded, and teachers are enjoined to adapt the way that the texts are presented to suit the level of proficiency of the students. The notion of « gist » listening was at that time quite a new one and the caveat is issued that the student is not expected to understand every word. Various tactics are deployed to ensure that students are supported in their understanding, however: the provision of vocabulary lists to be presented before listening, followed by « sign-post » questions, a range of multiple-choice, gap-fill, matching or other types of activity in English or in French. « Authenticity » is thus constructed as being both naturally-occurring French speech (and writing) and a negotiation which involves a careful consideration of learner proficiency and learner needs.
- 16 A similar approach was taken in the ensuing publications. All of these were based on recordings of spontaneous speech. *A Vrai Dire... Authentic French for Role-Play* (Beeching, 1986), focuses on productive skills at GCSE. *Contrastes* (Beeching & Page, 1988) was aimed at the development of a range of skills in A-level students. *Ça se Dit et ça s'écrit* (Beeching & Le Guilloux, 1990) was designed to develop advanced spoken and written skills, and to differentiate between the two modes, while *La Passerelle* (Beeching & Le Guilloux, 1993) provided a much-needed bridge between GCSE and A-level, focusing on developing basic grammar in students whose syllabus up till then had not done so.
- 17 After more than 10 years of recording and transcribing spoken French, I had collected a fairly large corpus of spoken language, 17½ hours and 155,000 words of which are

available as an online corpus: www2.uwe.ac.uk/faculties/CAHE/ELC/Documents/iclu/corpus.pdf. The corpus features speakers from Paris, Brittany, the Lot, the Minervois and from Belgium and includes school-children, bakers, butchers, nurses, doctors, salespersons and even a *toiletteur de chiens/dog stylist*.

- 18 In 2014, there is still no fully-fledged reference corpus of French to equal the *British National Corpus* though a number of projects are afoot to create one. In addition to my small *Beeching Corpus* (1980-1993) mentioned above, the following are accessible online:
 - ELILAP (Étude Linguistique de la Langue Parlée) and LANCOM (Langue et Communication) (<http://bach.arts.kuleuven.be/elicop/ProjetLANCOM.htm>) ;
 - ESLO and ESLO 2 Enquête Sociolinguistique sur Orléans (<http://eslo.tge-adonis.fr/>), which will contain the ESLO Corpus from 1968, replicated in 2014, as ESLO2 ;
 - CRFP (Corpus de Référence du Français Parlé) ;
 - CFPP (Corpus du Français Parlé Parisien) (<http://cfpp2000.univ-paris3.fr/>) ;
 - CLAPI (Corpus de Langue Parlée en Interaction) (<http://clapi.univ-lyon2.fr/>).
- 19 There is a fuller list of French spoken corpora here:
 - Inventaire des corpus oraux de la DGLFLF (Délégation Générale à la Langue Française et aux Langues de France). Disponible en ligne. http://www.culture.gouv.fr/culture/dglf/recherche/corpus_parole/Inventaire.pdf
 - CNRTL (Centre National de Ressources Textuelles et Lexicales). Disponible en ligne. <http://www.cnrtl.fr/>
- 20 These corpora provide data from different genres and regions of the French-speaking world, some with aligned audio, video and transcriptions. All of them can be exploited by the language teacher (or language researcher). However, until they are grouped together, and selections are made, none can be said to have been sampled so as to be representative of the French language as a whole (see Pierrel, this volume, for an initiative to make French corpora more accessible).

From corpus to syllabus: French for gardeners and golf greenkeepers

- 21 In the early 1990s, I was asked to teach French to a group of trainee gardeners and golf greenkeepers from the nearby Agriculture College in Bridgewater. These students had GCSE French and a very remote interest in French grammar. They undertook a field-trip each year to a partner-college in Coutances in Normandy and some of them went on to do internships at gardening establishments in France such as a *bambouseraie* (bamboo plantation) in Lyon. In order to meet their needs and to harness their motivation, I developed the course to include references to their specific subject-area: Horticulture. Drawing on horticultural magazines in French, the students prepared talks – in French – on subjects as diverse as maintaining the lawn-courts at Wimbledon, dealing with moles and badgers, using vegetation shredders, and so on. The level of language achieved was astonishing despite the students' lack of knowledge of basic grammar rules. The students invited me to join them on their annual field trip to Coutances. This was my opportunity to collect the data I needed to tailor next year's course to their specific needs. The trip was organised around a series of visits to experimental stations where a range of plants were grown. One of the students video-recorded the presentations in French while I did some improvised interpreting where required. There were frequently

occasions when I filled in the gaps in the students » French while the students filled in the gaps in my knowledge of horticulture. So, when *periodicité* was mentioned, I could translate it as « periodicity » without having a clear knowledge of what was meant. The students were able to explain to me the process that was being referred to, and were proud to do so. This type of collaboration between language and subject specialists is particularly fruitful in specific purposes contexts leading to a greater sense of equality between teacher and learner and to collaborative and negotiated learning (for more on Specific Purposes, see Kübler, this volume).

- 22 On our return to the UK, I transcribed the video-recordings and analysed the data; the results are published in Beeching (1997). The main findings which informed the syllabus for the following year in terms of grammatical items, important structures and lexis were as follows:

23 Grammatical items:

- the synthetic future was not used at all, so priority was placed on the analytic future (aller + infinitive) ;
- the *qui/que* distinction which poses a problem for English learners of French was not crucial – most uses were of *qui* in presentatives (see below).

24 Important structures:

- 25 Structures which were very frequent in the data and which are not traditionally covered in intermediate courses included characteristically spoken structures such as presentatives and also structures used to express cause and effect. These are illustrated below:

- Presentatives

Il y a / On a un/e qui

il y a une usine qui s »est construite

on a deux fois 4500 plants qui arrivent en même temps.

C »est un... qui/que/qu »...

c »est des variétés qui sont assez porteuses

c »est un chauffage qu »on met au coeur de la végétation

- Condition/Cause > Effect

on met le chauffage 60 degrés > effect > ça permet de chauffer le cœur

si on plante au mois d »avril un cutting dans le sol > effect > il ne va pas fleurir ici où vous avez une forte luminosité > effect > vous avez une chute importante de l »hydrométrie

26 Lexical items:

- Names of plants and parts of plants ;
- the processes applied to plants, equipment, including computer-assisted administration ;
- growth conditions ;
- commercial terminology.

- 27 Most of the plants have cognate forms in English, e.g. *aster*, *astroméria*, *carotte*, *chrysanthèmes*, *escarole*, *frisée*, *frésia*, *géranium*, *gerbera*, *orchidée*, *pelargonium*, *rununcule*. What is more, a high proportion of semi-technical terms can be inferred from English, e.g. *bulbe*, *botritis*, *cellule*, *climat*, *compartement*, *CO₂*, *concentration*, *cutting*, *évaporation*, *humidité*, *injecter*, *irrigation*, *monopole*, *modem*, *périodisme*, *PPM*, *programmer*, *qualité*, *quantité*, *température (ambiante)*, *tester*, *variété*, *végétation*. In terms of pedagogy, grouping terms together can help to focus the learner »s attention on families of forms (*fleur*, *fleur coupée*, *double fleur*, *simple fleur*, *mise à fleur*, *floraison*, *date de floraison*, *fleuron*, *fleurir*) or on forms

which look similar but which have different translations in English: *plant* (« seedling »), *plante* (« plant »), *plante à massifs* (« bedding plant »), *plantation*, *planter*.

- 28 By corpus standards, this corpus was extremely small. It was, however, precisely targeted at this group of learners » specific needs and thus forms an authentic representative sample of « French for Amenity Horticulture ». Collecting a spoken corpus on the ground allows the teacher of Language for Specific Purposes to discover the communicative acts, the lexis and the grammar which are vital in the contexts in which the learners can envisage operating. The transcription and detailed linguistic analyses bring important insights which can inform the construction of the syllabus. In this case, the data collection and transcription also brought to light certain characteristics of spontaneous spoken French and of multimodality which are interesting from a theoretical point of view and form part of a lively and ongoing field of research (Bruxelles *et al.*, 2006; Mondada, 2009).

A corpus of role-plays

- 29 The final example illustrates the benefits of creating a learner corpus (for more on this topic, see De Cock, this volume; Granger, 2009). A comparison of native speaker and learner productions in specific contexts can highlight the differences between L1 and L2 usages, and can help to inform language teaching. In this case, I wanted to investigate the use of discourse markers (DMs) (see Beeching, forthcoming, for fuller details). Previous studies of DMs conducted by applied linguists and language teachers have suggested that, far from being irritating « tics » or « parasites », markers have an important role to play in spoken interaction. Back in the 1980s, Svartvik noted that:

If a foreign language learner says five sheeps or he goed, he can be corrected by practically every native speaker. If, on the other hand, he omits a « well », the likely reaction will be that he is dogmatic, impolite, boring, awkward to talk to, etc., but a native speaker cannot pinpoint an « error ». (Svartvik, 1980: 171)

- 30 Fung and Carter (2007), too, recommend the inclusion of markers in the curriculum:

Incorporation of DMs into the language curriculum is necessary to enhance fluent and naturalistic conversational skills, to help avoid misunderstanding in communication, and, essentially, to provide learners with a sense of security in L2. (Fung & Carter, 2007: 433)

- 31 Further research is, however, required into the functions, polysemies and sociolinguistic salience of individual markers before language teachers can feel confident about presenting them as part of the language curriculum. As Kasper (1997: n.p.) has pointed out, pragmatic competence is not easily amenable to intuition:

Because native speaker intuition is a notoriously unreliable source of information about the communicative practices of their own community, it is vital that teaching materials on L2 pragmatics are research-based (Myers-Scotton & Bernstein, 1988; Wolfson, 1989; Olshtain & Cohen, 1991; Bardovi-Harlig *et al.*, 1991).

- 32 Research of this sort can most reliably be conducted using corpus data.
- 33 Typically, and understandably, learners are expected to simply « pick up » DM-usage as they become more integrated into a particular speech community (see Hellerman & Vergun, 2007). Given that the meanings and functions of DMs can be complex and opaque and that the sociolinguistic salience of particular markers may vary across different parts of the English-speaking world, it seems premature to rush into the teaching of them; we

may not know, for example whether a learner will end up in the UK, the US, Australia, New Zealand or Canada, not to mention countries where English is a widely spoken colonial language, like India, or others where it is a Lingua Franca. Similarly, teachers typically do not know whether their learners are destined for either further study or employment in one of the countries in which English is spoken or will for example require a purely reading knowledge of the language to facilitate study in the host country. A grasp of more colloquial spoken forms like DMs will more likely prove to be an asset for the former group of students than for the latter.

- 34 Table 1 (from Beeching, forthcoming) summarises the range of studies which have previously been conducted on the acquisition of different DMs across a range of L1s, levels of competence and contexts, and employing different research methods.

Table 1: Previous studies of native and non-native usages of discourse markers.

Study	L1	Level of competence in English	Context	Research method and markers studied
Müller, 2005	German	9-12 + years of formal instruction university students (95,555 words)	Silent movie stimulus > oral narrative in pairs, followed by exchange of opinions – conducted in experimental conditions	- Information gap oral narrative - Exchange of opinions - « Use of English » questionnaire <i>So, well, you know, like</i>
Fung & Carter, 2007	Cantonese	49 Intermediate-advanced school pupils (17-19 year-olds) (14,157 words)	Tape-recorded group role-play recorded in Hong Kong in a pedagogic context	« Naturally occurring » interaction (a very wide range of markers are included)
Denke, 2009	Swedish	Advanced (37,152 words)	Video-recorded natural science seminars – mainly monological	Naturalistic spoken data <i>Y »know, well, I mean</i>
Liao, 2009	Chinese	6 Advanced graduate students in a study-abroad context 6 x 40-50 min. lectures + 6 x 40 min. interviews > 1,422 tokens of 9 markers	Teaching-Assistant led discussions in the US; sociolinguistic interviews	Lecture with quantitative analysis + qualitative ethnographic interviews <i>Yeah, oh, you know, like, well, I mean, ok, right, actually</i>

Baumgarten & House, 2010	US English and mixed non-natives: Indonesian, German, Chinese, Korean, French, Nepali, Gujarati	University students in Germany; 3 x 30-min. conversations on a discursive topic: English L1 ELF1 ELF2	Audio-taped elicited conversations on a topic of general interest « the role of English in the world » or « men and women in the contemporary arts scene »	Quantitative and qualitative analysis of the uses of <i>I think</i> and <i>I don't know</i>
Polat, 2011	Turkish	5 years formal EFL instruction in Turkey; emigrated to US aged 25	One individual who has emigrated (married to a US citizen). Audio-taped conversations every 2 weeks for 1 year	Developmental learner corpus – one speaker »s development over a year <i>You know, like and well</i>
Beeching, forthcoming	Chinese L1 Mixed L1 recorded at UWE Native speakers recorded by Qun Zheng at UWE	Intermediate / advanced university students, in 2 groups: 1. 24 Beijing Foreign Studies University: SWELC Corpus (6,326 words) 2. 28 L1 Albanian, Czech, French, German, Spanish and Vietnamese students at UWE (7,419 words) 3. 44 native speakers of English, students at UWE (14,308 words)	3-min. argumentative role-play in pairs; exam situation; « practice » exam situation; experimental situation	Quantitative (role-play data) + qualitative: some brief ethnographic interviews of mixed L1 subjects <i>Well, you know, I don't know, like, sort of, I think, I mean</i>

- 35 The scope of the current article does not permit detailed elaboration on each of these studies. Some studies (Denke, 2009; Liao, 2009) look purely at DM usage in non-natives while others compare native and non-native usages. Fung and Carter (2007) compare native and non-native usage in corpora which were collected independently of each other (they compared their own data with occurrences in the *Cancode* corpus). Polat (2011) compares the data from her one individual subject that she studies longitudinally with native-speaker data drawn from the *Louvain Corpus of Native English Conversation* (Locnec: De Cock, 2004). Müller (2005), Baumgarten and House (2010) and my own forthcoming

study draw on comparable corpora: native and non-native data were collected following the same protocol with both natives and non-natives in order to ensure that like was being compared with like.

- 36 Of the studies which did not include detailed comparison with a native-speaker corpus, Denke »s (2009) investigation looks specifically at repair and repetition in the context of monological seminar presentations. The more interpersonal functions of DMs are less frequent in such a context. The interest of Denke »s study lies in the way that genre influences both the frequency of occurrence and the functions which DMs can have. Liao »s (2009) study of Chinese Teaching Assistants in the US highlights the extent to which the speaker »s desire to integrate in the target culture influences their rate of usage of DMs. One of her participants lived with her Chinese husband, had little contact with natives of English and intended to return to China – her attitudes to DMs as informality markers were largely negative and she avoided using them. Another participant was, however, intending to settle permanently in the US, had a number of American friends and used DMs plentifully as a means of creating friendly rapport. Ethnographic studies of this sort, which couple attitudinal interviews with transcribed recordings of spoken data which can be drawn on for detailed frequency counts, allow the researcher to correlate language usage with extralinguistic information in a particularly interesting way. Polat (2011) extends this type of approach to her longitudinal study of a single individual, a native speaker of Turkish who emigrated to the US aged 25 and married an American native speaker of English. Polat used the developmental learner corpus she created to trace patterns of use of three markers: *you know*, *like* and *well*. She found that *you know* was heavily used by the participant though rates of usage dropped by half over the year of recordings. *Like* increased from almost no usages at the beginning to 230 per 10,000 words by the mid-point, then reduced by half by the end of the study. Rates of both *you know* and *like* far exceeded those of the native speakers in the *Locnec* corpus. Finally, *well* was not used at all as a discourse marker.
- 37 Turning now to the studies which explicitly focus on eliciting native and non-native performance, Müller (2005) compares the use of markers by American and German native speakers performing the same tasks in experimental conditions. She has a large corpus to draw on (the *Giessen – Long Beach Corpus*) and her study compares two communicative functions, narrating and opinion-giving. She provides an exhaustive analysis of the uses of four markers: *so*, *well*, *you know* and *like*. She found that the Americans used *so*, *you know* and *like* to a statistically significantly greater extent than the Germans. *Well*, on the other hand, was used more by the Germans than the Americans, though not to a statistically significant extent. The native speakers of German used *well* for different functions – when searching for the right phrase, giving an indirect answer, continuing an opinion and concluding (to a statistically significant degree).
- 38 Fung and Carter's (2007) study compares rates of DM usage in the interactive classroom discourse of secondary school pupils in Hong Kong with usage in the *Cancode* corpus; native speakers were found to use DMs for a wider range of pragmatic functions. Baumgarten and House (2009), on the other hand, found that *I think* was used more frequently and in a wider range of contexts by ELF than by L1 English speakers: it was the most common way of « expressing subjective meanings and taking stances ». This finding was replicated in my own forthcoming study with respect to the Chinese learners of English. In this study, native speakers, a mixed non-native group on a stay in the UK and Chinese non-natives who had not been to an English-speaking country were recorded

doing a 3-minute role-play in pairs. Their conversations were transcribed and raw distributional frequencies were charted for a range of lexemes which have DM functions. Figure 1 charts rates of occurrence per 10,000 words of the lexemes *well*, *you know*, *I don't know*, *like*, *sort of*, *I think* and *I mean*. It is important to point out that these are raw rates, in other words, counts for *like* include all occurrences of *like* (as in *I like bananas* and *He looks just like his father*, not just discourse-marking *like*).

Figure 2 : Rates of occurrence per 10,000 words of a range of discourse markers in role-plays conducted by Chinese and mixed non-native and native speakers of English.

- 39 Figure 2 shows that *well* is most often used by the natives at a rate of just over 40 occurrences per 10,000 words, with the mixed non-natives coming in a rate of 20 and the Chinese non-natives using it about 5 times per 10,000 words. In general, the Chinese beginners have low rates of DM usage apart from, and quite markedly so, *you know* and *I think*. It seems that these learners rely quite heavily on these two markers to elicit agreement or approval from their interlocutors and to express hesitation – or, indeed, their opinion. The mixed non-natives who have spent some time in the UK have rather higher rates than the Chinese non-natives and equal, or exceed, the native speakers in their rates of usage of *you know*, *I think* and *I mean*. That is to say that these learners are very good at using interactional markers which are addressee-oriented (*you know*) and speaker-oriented (*I think* and *I mean*). They also use *I don't know* to a greater extent than native speakers do. These learners use DMs as compensation strategies which help them to fill their pauses with appropriate fillers. Polat (2011: 3752) suggests that her participant uses *you know* as a « catch-all gap-filler » and finds that native speakers have far higher rates of other hesitation markers such as *uh/eh*, *um* and *er* which would support such a hypothesis. In my own data, the mixed non-natives have lower rates of both *like* and *sort of* than the native speakers but they have clearly picked up the usage of *like* from their native-speaking peers and are beginning to use it appropriately from both a functional and sociolinguistic point of view –for their age-group, this is the mitigating marker of choice. Corpora of this sort allow not only raw frequency counts, but also access to the contexts in which the markers were used. In the vast majority of cases, *like* was used as an approximative and not in its focaliser or quotative functions (see Dailey O'Cain, 2000; Macaulay, 2001; Buchstaller & D'Arcy, 2009). The example below shows the way in which *like* was used alongside other hesitation gap-filling markers, such as *I mean*,

you know and I don »t know to “act as a stock-phrase mainly used to help speakers process and plan their output, and link spans of discourse” (House, 2009: 187):

L: yeah sure but I mean like you can we can probably work for money during the year and you know it »s much more rewarding when you know you get to work for free and you know like for me for example I I I »m going to do volunteering this summer you know helping people er like I don »t know feel more yeah rewarding in the end you should do that.

- 40 Brief conversations were held with the participants after their role-plays were recorded, and subject « L » in particular had not been in the UK for long (under 3 months). She declared herself to be very interested in staying on in the UK – indeed, she claimed that she wanted to marry an Englishman! This desire in principle to be accepted and integrated into UK society appears to play a role in the adoption of markers such as *like*.
- 41 In conclusion, there is considerable variability in the use of DMs across native and non-native speakers, in terms of both their distributional frequency and the functions for which they are pressed in to service. Natives tend to use more, but this is not invariably the case: non-natives can also depend on certain markers like « lexical teddy bears » (Hasselgren, 2002) to bridge the gaps between sequences of talk, to hesitate and fill pauses. These appear to vary according to their native language: there is some evidence that natives of German use *well*, Dutch-speaking Belgians use *so* (Buysse, 2012) and Chinese learners use *I think*. Most learners depend on *you know* to gain confirmation from their interlocutor that they have said the right thing. As Polat »s participant Alex puts it:
I wanted to make sure you understood me. I don »t trust my English. I am saying right or wrong, and I want to get approval from you. (Polat, 2011: 3754).
- 42 As Polat remarks, this indicates that DMs may play a particular role in non-native discourse, one which is similar but more crucial than that played in native discourse, as a means of negotiating meaning.
- 43 There are, however, other factors, too, which come into play in the acquisition of DMs. Some learners acquire markers in a naturalistic way in conversation with native speakers once they are in the host country, but identity factors play a large part in their desire to do so. DMs are complex, multifunctional and may be stylistically stigmatised. Teaching them may rely on developing students » observation using appropriate naturalistic data and allowing time for conversation of an interactional sort in the classroom.
- 44 In this section, we have surveyed studies of the non-native acquisition of DMs which draw on different types of corpora, recorded in different communicative situations. The advantage of the experimental design of the role-play corpus is that it generates data which allow the researcher to compare like with like. It is relatively easy to collect and transcribe such a corpus: each role-play is only 3 minutes in length and a representative corpus of different speakers can therefore be collected fairly quickly. The role-play scenario was sufficiently stimulating to produce lively discussion – and, most importantly, a good number of DMs is generated. The data are thus robust and it is possible to make both statistical and qualitative analyses.
- 45 On the downside, the role-play situation could be considered to be unnatural, set in a particular communicative situation and thus ungeneralisable. Also, as Müller (2005) points out, there may be significant differences depending on whether speakers took Role A or Role B – eliciting or giving information, or in the case of the UWE Role-play, defending the volunteering or the working-for-money position.

Conclusions: Criteria for corpora for language teaching and learning

- 46 Table 2 attempts to summarise the criteria that need to be borne in mind when considering a corpus which might be suitable for language teaching. In this article, we have considered three different scenarios: firstly, in « From syllabus to corpus », a corpus is collected and transcribed by a materials developer specifically with a large general population of students in mind, who are studying for exams based on a national curriculum. These data are authentic, non-scripted, representative, machine-readable and pedagogically relevant, with support in the form of vocabulary help, sign-post questions, and a range of gap-fill, matching and multiple choice comprehension and production exercises. Secondly, in « From corpus to syllabus », authentic, non-scripted and relevant recordings are made *in situ* in a Language for Specific Purposes situation. These recordings and their transcriptions are analysed in order to inform the syllabus in terms of the most common lexical and structural features which are pertinent in this specific communicative situation. This corpus may be small, and is not generalizable to students following a more global language course (though aspects of it, such as features of the spoken language which have not traditionally been taken on board in language syllabuses, may be of more general interest). Finally, in the section on « The role-play corpus », the place of corpora for research on the acquisition of particular features was highlighted, in this case the growing study of the functions of DMs in native and non-native discourse.

Table 2 : Criteria for corpora for teaching and learning

	CORPORA			FOR TEACHING AND LEARNING	
	authentic? (non-scripted)	Representative?	machine-readable?	pedagogic?	
				relevant	support
Syllabus > corpus	✓	✓	✓	✓	✓
Corpus > syllabus	✓	✓ (of a particular genre)	□	✓	no
Role-play	✓ (guided, improvised)	✓ (of a particular population)	✓	✓	no yet

- 47 Looking to the future, developments will no doubt include video as well as audio as a necessary part of what is deemed authentic for the 21st century, with students » increasing access to *YouTube* clips and internet media. Multimedia corpora coupled with the necessary pedagogic support online in the form of sub-titles, transcriptions and language-learning activities can be tailored to students » specific needs. Multimedia

applications are not only motivating and appropriate to the internet age, but provide greater contextualisation, including paralinguistic features and sociolinguistic background information, to better situate extracts in their authentic communicative situation.

- 48 At more advanced levels and in university language learning situations, students themselves can create their own « do-it-yourself » corpora, which can be either « throwaway » or archived to be used by their peers (see Tyne, 2009; Boulton & Tyne, 2013: 108). The active engagement of students in the creation of corpora raises a more general and crucial point with respect to the relationship between corpora and pedagogy. Hands-on corpus activities involving concordance lines do not suit all learners: some prefer to see language embedded in social situations in which real speakers can be seen engaging with each other in specific sociolinguistic contexts. If students select extracts for transcription themselves, we can be more certain that not only the material itself but also the engagement that the student has with it is « authentic ».

BIBLIOGRAPHY

- Aijmer, K. (ed.) (2009). *Corpora and language teaching*. Amsterdam: John Benjamins.
- Baumgarten, N. & House, J. (2010). "I think and I don't know in English as lingua franca and native English discourse". *Journal of Pragmatics*, vol. 42, n° 5. pp. 1184-1200.
- Beeching, K. (1982). "Authentic materials". *British Journal of Language Teaching*, vol. 20, n° 1. pp. 17-20.
- Beeching, K. (1985). *Vrai de vrai ! Authentic French for listening and reading comprehension*. Oxford: Oxford University Press.
- Beeching, K. (1986). *A vrai dire... Authentic French for role-play*. Oxford: Oxford University Press.
- Beeching, K. (1997). "French for specific purposes: the case for spoken corpora". *Applied Linguistics*, vol. 18, n° 3. pp. 374-394.
- Beeching, K. (Forthcoming). "Variability in native and non-native use of pragmatic markers: the example of *well* in role-play data". In Beeching, K. & Woodfield, H. (dir.), *Researching sociopragmatic variability: perspectives from variational and interlanguage pragmatics*. Basingstoke : Palgrave Macmillan.
- Beeching, K. & Le Guilloux, I. (1990). *Ça se dit et ça s'écrit*. Oxford: Oxford University Press.
- Beeching, K. & Le Guilloux, I. (1993). *La passerelle: French grammar in use*. Cambridge: Cambridge University Press.
- Beeching, K. & Page, B. (1988). *Contrastes*. Cambridge: Cambridge University Press.
- Boulton, A. (dir.) (2009). "Des documents authentiques oraux aux corpus : questions d'apprentissage en didactique des langues". *Mélanges Crapel*, vol. 31. Available on line. <http://www.atilf.fr/spip.php?rubrique584>

- Boulton, A. & Tyne, H. (2013). "Corpus linguistics and data-driven learning: a critical overview". *Bulletin Suisse de Linguistique Appliquée*, vol. 97. pp. 97-117.
- Bruxelles S., Mondada, L., Simon, A. C. & Traverso, V. (dir.) (2009 [2007]). *Grands corpus de français parlé : bilan historique et perspectives*. *Cahiers de Linguistique de Louvain*, vol. 33, n° 2.
- Bruxelles S. & Traverso, V. (2006). "Usages de la particule voilà dans une réunion de travail : analyse multimodale". In Drescher, M. & Frank-Job, B. (dir.), *Les marqueurs discursifs dans les langues romanes : approches théoriques et méthodologiques*. Francfort: Peter Lang. pp. 71-93.
- Buchstaller, I. & D'Arcy, A. (2009). "Localized globalization: a multi-local, multivariate investigation of quotative *be like*". *Journal of Sociolinguistics*, vol. 13, n° 1. pp. 291-331.
- Buysse, L. (2012). "So as a multifunctional discourse marker in native and learner speech". *Journal of Pragmatics*, vol. 44. pp. 1764-1782.
- Dailey-O'Cain, J. (2000). "The sociolinguistic distribution and attitudes toward focuser *like* and quotative *like*". *Journal of Sociolinguistics*, vol. 4. pp. 60-80.
- De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures*, vol. 2. pp. 225-246.
- Denke, A. (2009). *Native-like performance: pragmatic markers, repair and repetition in native and non-native English speech*. Sarrebrück: VDM Verlag Dr. Müller.
- Duda, R., Esch, E. & Laurens, J.-P. (1972). "Documents non didactiques et formation en langues". *Mélanges Pédagogiques*, vol. 1972. pp. 1-47. Available on line. <http://www.atilf.fr/IMG/pdf/melanges/4duda-esch-laurens.pdf>
- Duda, R. & Tyne, H. (2010). "Authenticity and autonomy in language learning". *Bulletin suisse de linguistique appliquée*, vol. 92. pp. 87-106.
- Fung, L. & Carter, R. (2007). "Discourse markers and spoken English: native and learner use in pedagogic settings". *Applied Linguistics*, vol. 28, n° 3. pp. 410-439.
- Granger, S. (2009). "The contribution of learner corpora to second language acquisition and foreign language teaching: a critical evaluation". In Aijmer, K. (dir.), *Corpora and language teaching*. Amsterdam: John Benjamins. pp. 13-32.
- Hasselgren, A. (2002). "Learner corpora and language testing: smallwords as markers of language fluency". In Granger, S., Hung, J. & Petch-Tyson, S. (dir.), *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins, pp. 143-173.
- Hellerman, J. & Vergun, A. (2007). "Language which is not taught: the discourse marker use of beginning adult learners of English". *Journal of Pragmatics*, vol. 39. pp. 157-179.
- Holec, H. (1974). "Cours initial d'anglais oral : une approche fonctionnelle". *Mélanges Pédagogiques*, vol. 1974. pp. 13-28. Available on line. <http://www.atilf.fr/IMG/pdf/melanges/2.holec.pdf>
- Holec, H. (1990). "Des documents authentiques, pour quoi faire ?" *Mélanges Pédagogiques*, vol. 1990. pp. 65-74. Available on line. <http://www.atilf.fr/IMG/pdf/melanges/5holec-2.pdf>
- House, J. (2009). "Subjectivity in English as Lingua Franca discourse: the case of *you know*". *Intercultural Pragmatics*, vol. 6, n° 2. pp. 171-193.
- Hunston, S. (2009). "The usefulness of corpus-based descriptions of English for learners: the case of relative frequency". In Aijmer, K. (dir.), *Corpora and language teaching*. Amsterdam: John Benjamins. pp. 141-154.

- Kasper, G. (1997). "Can pragmatic competence be taught?" *NFLRC NetWork 6*. Honolulu: University of Hawai'i, Second Language Teaching & Curriculum Center. Available on line. <http://www.nflrc.hawaii.edu/networks/NW06/>
- Liao, S. (2009). "Variation in the use of discourse markers by Chinese teaching assistants in the US". *Journal of Pragmatics*, vol. 41. pp. 1313-1328.
- Macaulay, R. (2001). "You're like « why not? » The quotative expressions of Glasgow adolescents". *Journal of Sociolinguistics*, vol. 5, n° 1. pp. 3-21.
- McEnery, P., Xiao, R. & Tono, Y. (2006). *Corpus-based language studies*. Abingdon: Routledge.
- Mondada, L. (2009). "Emergent focused interactions in public places: a systematic analysis of the multimodal achievement of a common interactional space. » *Journal of Pragmatics*, vol. 41. pp. 1977-1997.
- Müller, S. (2005). *Discourse markers in native and non-native English discourse*. Amsterdam: John Benjamins.
- Nunan, D. (2007). "The nineteen eighties". In Brown, H. Tarone, E. Swan, M., Ellis, R., Promdomou, L., Jung, U. Bruton, A., Johnson, K., Nunan, D., Oxford, R., Goh, C., Waters, A. & Sauvignon, S. Forty Years of Language Teaching. *Language Teaching*, vol. 40. pp. 1-15.
- Polat, B. (2011). "Investigating acquisition of discourse markers through a developmental learner corpus". *Journal of Pragmatics*, vol. 43. pp. 3745-3756.
- Rühlemann, C. (2008). "A register approach to teaching conversation: farewell to standard English?" *Applied Linguistics*, vol. 29, n° 4. pp. 672-693.
- Svartvik, J. (1980). "Well in conversation". In Greenbaum, S., Leech, G. & Svartvik, J. (dir.), *Studies in English linguistics for Randolph Quirk*. Londres : Longman. pp. 167-177.
- Tyne, H. (2009). "Corpus oraux par et pour l'apprenant". *Mélanges Crapel*, vol. 31. pp. 91-111. Available on line. http://www.atilf.fr/IMG/pdf/melanges/05_Tyne.pdf
- Widdowson, H. (1998). "Context, community and authentic language". *TESOL Quarterly*, vol. 32, n° 4. pp. 705-716.

Corpora

- Beeching Corpus* (BC). (1988). 17½ hours of spoken French, 155,000 words, 95 speakers. <http://www.uwe.ac.uk/hlss/llas/iclru/corpus.pdf>
- British National Corpus* (BNC). (1980s-1994). 100 million words of British English, available on line via <http://corpus.byu.edu/>. Further information on all versions available at <http://www.natcorp.ox.ac.uk/corpus/index.xml>.
- British National Corpus 2014 Spoken*. Project announcement at <http://cass.lancs.ac.uk/?cat=630>.
- Cambridge and Nottingham Corpus of Discourse in English (Cancode)*. Cambridge University Press. 5 million words of spoken English from a variety of settings in the United Kingdom and Ireland, with metadata on participants, settings, and conversational goals. It forms part of the 1 billion word CIC (Cambridge International Corpus), but is not freely available.
- Corpus de Référence du Français Parlé* (CRFP). (2002). From 40 towns in France, 400,000 words. Available via a concordancer at <http://sites.univ-provence.fr/delic/corpus/index.html>.
- Enquête Sociolinguistique d'Orléans* (ESLO). (1966-1970). 109 hours of spoken French (902,755 words transcribed), available at <http://bacharts.kuleuven.ac.be/elicop>.

Giessen – Long Beach Chaplin Corpus. (1999-2005). Interactions between native and non-native English speakers with a given task, available at <http://ota.ahds.ac.uk/desc/2506>.

NOTES

1. School exams taken at the age of 16 and 18 respectively, in England, Wales and Northern Ireland.

ABSTRACTS

This article argues the case for regarding the relationship between corpora and the French term "didactique des langues" (i.e. theory and practice of language teaching and learning) as precisely that: a relationship. Relationships are negotiated and it is this negotiation which is managed by teachers with support from curriculum and materials developers. The article begins by attempting to define what is meant by the terms "corpus" and "teaching" and addressing the thorny question of "authenticity". It proceeds to give illustrations, drawn from my own experience, of three ways in which corpora and language teaching/learning are connected. The first, "From syllabus to corpus", describes how a small corpus of spoken French was created incrementally over more than 10 years of collecting authentic spoken recordings for exploitation in textbooks for the teaching of French to GCSE and A-level students in the UK from 1980 to 1993. The second, "From corpus to syllabus", describes how a specialist spoken corpus with a focus on amenity horticulture was collected, transcribed and analysed to inform a language learning syllabus. The third section, "A corpus of role-plays", investigates the role of learner corpora and how they can inform language teaching, with specific reference to the acquisition of discourse markers. In its consideration of the links between corpora and language teaching, the conclusion highlights the role of motivation, the importance of engaging students' interest and thus rendering activities « authentic ».

Répondant à la question "Qu'est-ce qu'un corpus en didactique des langues?", cet article se propose de mettre en questions la prééminence accordée au corpus lui-même pour privilégier le point de vue de l'apprenant. La relation entre didactique et corpus est en effet prioritairement de l'ordre d'une relation. Cette relation s'établit normalement par la médiation de l'enseignant, aidé souvent en cela par les concepteurs de programmes et de ressources pédagogiques. L'article commence par une définition des termes "corpus", "didactique" et de celui, plus épineux, d'"authenticité". Pour suivre, trois aspects du lien entre didactique et corpus sont alors illustrés, articulés avec mon expérience propre. La première partie, "Du syllabus au corpus", présente un petit corpus de français parlé créé à partir d'enregistrements authentiques réunis et transcrits durant plus de 10 ans entre 1980 et 1993. Des manuels de français langue étrangère inspirés de ces enregistrements authentiques visaient la préparation d'apprenants britanniques aux examens de niveau GCSE et A-Level. La deuxième partie, "Du corpus au syllabus", montre comment le recueil, la transcription et l'analyse des corpus constitués dans un contexte spécialisé, celui de stations expérimentales d'horticulture, illustrent la constitution d'un syllabus langagier. La troisième partie, "Un corpus de jeux de rôle", étudie le rôle des corpus d'apprenants et leur contribution pour l'enseignement-apprentissage, tout particulièrement pour l'acquisition

des marqueurs discursifs. Au regard du lien entre didactique et corpus, la conclusion souligne le rôle de la motivation et l'importance de l'engagement de l'intérêt de l'apprenant qui, seul, garantirait l'"authenticité" des activités proposées.

INDEX

Mots-clés: syllabus, authenticité, besoins spécifiques, marqueurs discursifs, corpus pédagogiques

Keywords: syllabus, authenticity, specific purposes, discourse markers, do-it-yourself corpora

AUTHOR

KATE BEECHING

Kate Beeching is Associate Professor in Applied Linguistics at the University of the West of England, Bristol, United Kingdom, where she is head of the International Corpus Linguistics Research Unit (<http://www1.uwe.ac.uk/cahe/research/bristolcentreforlinguistics/iclu.aspx>) in the Bristol Centre for Linguistics (<http://www1.uwe.ac.uk/cahe/research/bristolcentreforlinguistics.aspx>). Her research interests include corpus approaches to the analysis of the functions, sociolinguistics and historical evolution of pragmatic markers in French and English, the teaching and learning of features of spoken interaction, and the analysis of interactions between carers and those living with dementia.

Email : Kate.Beeching@uwe.ac.uk

Website : <http://people.uwe.ac.uk/Pages/person.aspx?accountname=campus%5ck-beeching>

Address : University of the West of England, Dept. of Arts, Faculty of Arts, Creative Industries and Education, Frenchay Campus, Coldharbour Lane, Bristol BS16 1QY, United Kingdom.