Journal of Celtic Language Learning

Volume 12 2007

Research Articles

Teaching Forum

Book Reviews

KYG: A Corpus of Spoken Breton for Both Researchers and Advanced Learners

Christophe Ropers
Université de Bretagne Sud, Lorient (France)

For various reasons, the situation of advanced learners of Breton who reside in Brittany is not unlike that of the learners of any foreign language who do not live in the country where the language is spoken. Although linguistic immersion remains possible in theory, its practical conditions are rarely met, and some alternatives must be found for those advanced students who are asked not only to produce grammatically correct Breton utterances but also to copy, as much as possible, the type of utterance that native or near-native speakers would spontaneously produce. This paper suggests that, in such a case, partitioned electronic corpora used by researchers can also be useful to advanced learners. It describes the tools used to build and search the corpus, and suggests ways in which corpora may be used by learners in order to find the kind of contextualized information that dictionaries and grammars do not necessarily provide.

The Korpus ar Yezh dre Gomz[1] (KYG) project and its aims

As is often the case with most of the languages that are now commonly referred to as *less commonly taught*, students who wish to become proficient in Breton may face grave difficulties in the final stage of their language learning experience. Researchers in the teaching of English for Speakers of Other Languages (ESOL) would refer to students who have reached such a stage as *advanced*

learners. Although this paper does not deal with ESOL, the term *advanced learner(s)* will also be used here. The sort of difficulty encountered by advanced learners of Breton is closely related to the demographic and geographic gaps they must bridge when looking for the kind of non-academic linguistic immersion that is supposed to help them step up to near-native proficiency. Indeed, the speakers who could provide such a linguistic environment as that required for immersion live, for the most part, in rural areas of Brittany whereas most learners of Breton as a second language (who also happen to be rather young) live mostly in urban areas. Moreover, those learners who would still be willing to experience linguistic immersion in spite of the efforts involved face one more deterrent, as described in (Timm, 2005, pp. 41-42):

> " [...] the teaching of neo-Breton has produced new cohorts of speakers who do not share many of the same expressive and creative resources with the native speakers; a sort of linguistic dissonance is the result when neo- and paleo-speakers[2] attempt to engage in conversation. I have suggested that more attention might have been paid, and might still be paid, in the construction of grammars and dictionaries to some traditional genres in vernacular forms of the language – e.g., conversational styles that incorporate more idioms, proverbs, and sayings to help provide a bridge between the native and neo-Breton speakers."

Such is the goal that KYG has set out to reach.

The KYG project is currently hosted by the LiCoRN research team at Université de Bretagne Sud, which specializes in corpus-based linguistic studies. It brings together researchers from both the LiCoRN team and the Rennes branch of the CRBC research laboratory. It has been underway since September 2006 and is

expected to continue until at least 2012. As will be explained below, its ultimate goal is to provide the kind of spoken language data which is needed by researchers who study spoken language phenomena in context, and by those advanced learners who do not have as much access to naturally occurring language as they would like to. What is presented below are the preliminary results of ongoing work on newscast language samples. Ultimately, KYG aims to comprise all types of spoken material while keeping in mind that some types of speech will certainly be harder to come by than newscast speech.

In its first section, the paper examines, from a theoretical point of view, the link that can be drawn between teaching advanced learners and needing access to naturally occurring language. The second section shows how such access can be granted through the use of electronic corpora. In so doing, it discusses general and specific issues of corpus design, and explains how design choices were made in the specific case of KYG. The last section gives examples of the kind of information that advanced learners can get from corpus search results, which they could not readily get from a paper dictionary or grammar.

Theoretical grounding

Timm suggests that the dissonance between neo- and paleo-speakers of Breton is a result of the "language ideologies in Brittany in the past century", which "have promoted an emphasis on a form of Breton, the one currently being taught as a second language, which is in many ways quite different from the native, spoken forms of the language" (op.cit.). It is not certain, however, that language ideologies alone are to be blamed for the dissonance. All things considered, the situation depicted here is not at all unlike that of any learner of a second language who decides to go beyond the classroom environment and face the reality of listening to, or speaking, the language in its natural environment. The only difference with Breton is that the simple fact of living in, or going to, Brittany is not enough to guarantee

immersion in the Breton language. What is therefore needed for advanced learners of Breton as a second language is similar to what is needed to train advanced learners of any other second language who do not have regular access to immersion: a tool that would help them bridge the gap that lies between the artificial world of normative language and the world of naturally occurring speech.

It may seem obvious that language teaching should ultimately be concerned with teaching the kind of language that occurs naturally. Implementing such a principle, however, is not as easy as it may seem. Teaching naturally occurring language implies having extensive prior knowledge of how it is organized. Particular attention must therefore be paid to how such knowledge is acquired. Indeed, studies by Hill (1961), Quirk and Svartvik (1966), Greenbaum and Quirk (1970), and Labov (1975) have shown that native informants are highly unreliable in that, when asked about their own natural use of language, they give inconsistent and misleading information. This is probably what Hjelmslev (1961) was already alluding to by stating that "language wants to be ignored." Although the above-mentioned experiments were conducted among native speakers of English, there is no reason to believe that native speakers of Breton would be any less inconsistent. Studies by Spencer (1973) have gone even further by showing that what was true of so-called naïve informants was also true of trained language professionals. In other words, any attempt to gather data about naturally occurring speech productions through elicitation or introspection is bound to be biased by the speakers' or linguists' own representations or misconceptions of what naturally occurring language is or ought to be.

The only way to overcome the problems caused by such inconsistency is to resort solely to naturally occurring linguistic data. Linguists do not necessarily agree on what constitutes such data. Some consider that recorded observation, provided it is properly obtained—i.e. by trying to keep the observer's paradox (the potential influence that the observer's presence may have on

the speakers' productions) to a minimum—is an acceptable form of data. This is probably what Per Denez referred to in the preface to the first edition of Yann Desbordes' grammar handbook (Desbordes, 1990), when he stated that, in spite of its being normatively organized, the book's material relied on "competent observation"[3]. Others will argue that, although it may be possible for a trained observer to keep track of data that have not been corrupted by the observer's paradox, the practice of observation and note-taking will always lead to collecting data out of context. In that case, one alternative solution could be to build an electronic corpus (i.e. a collection of texts in an electronic database) of transcribed speech. For instance, the material for Favereau's grammar (1997) and dictionary (2000) was extracted from corpora. KYG, although it differs significantly in its format from Favereau's Poher corpus, follows the same line of thought.

As the idea behind the use of corpora is to represent language phenomena the way they are actually produced—and not the way some native speakers think they are or the way some language theoreticians would like them to be—the first issue corpus designers may still be faced with is that of representativeness (Biber, 2004). For instance, studies as early as those reported in Zipf (1932), showed that any language corpus will be significantly skewed. In any given language, a small number of elements will occur extremely frequently whereas a large number of elements will occur much more rarely. This was also Chomsky's main argument against the use of corpora in linguistic research. However, Zipf also alludes to the fact that language data are not the only ones to form a highly skewed population sample: most samples of naturally occurring phenomena follow the same pattern (also referred to as Zipf's law). While Zipf's law forces us to abandon the hope of ever working with normal distributions and easily predictable patterns, it also tells us that statistical methods commonly used to curve the skew of distributions can be applied to naturally occurring language data (Oakes, 1998). There are at least two ways of reducing the impact of

Zipf's law on the statistical analysis of such data:

1. significantly increasing the amount of population samples (i.e. the number of texts in the database);
2. partitioning the population into subgroups and conducting comparative analyses on the general population and on each of its subgroups.

Both these statistical devices have direct consequences on corpus design and use.

First of all, since the corpus must contain a large amount of data, it must be built in an automatically searchable (i.e. electronic) format. Secondly, the tools used to design and analyze the corpus must allow for partitioning so as to run analyses either on the whole corpus or on any of its subcorpora. Given the limited workforce available to build a corpus of spoken Breton, many years may elapse before its completion and, as a consequence, before the extraction of the data it contains is possible. For reference, the British National Corpus, on which the Oxford Advanced Learners' Dictionary is based, took three years in development and only ten percent of it consists of spoken material, most of which is non-spontaneous speech. According to Chafe et al. (1991, p. 70), the builders of the New Corpus of American Spoken English reported an estimated building cost of roughly 6 hours of work per minute of recorded speech. However, although a corpus may not be representative enough to serve as reference for the publication of a learners' dictionary, it might still be perfectly usable for specific learning activities. It might therefore be necessary, while thinking about corpus design and use, to add one more item to the tool selection criteria, i.e. that it should give teachers and students the possibility of searching the data contained in intermediate versions of the corpus. This can be done through the use of a concordancer—a computer application similar to a search engine which shows the results of the search in the form of lines referred to as *concordance* lines or *KWIC* (Key Word In Context) lines, as shown in Figure 1 and discussed more fully later in the paper. The application can then be made available from an institutional server, as is the case for MiCASE (Michigan Corpus of Academic Spoken English).

evit bezañ otopsiet da c'houzout ha	marv	eo diwar an taolioù pe diwar ar veuzadenn
penaos oa graet lidañ tout ar re zo	marv	er mor dibaoe pell zo
evit enoriñ ar vartoloded a zo	marv	evit ar vro
Jean-Jacques Lesieur hag a zo	marv	evit ar vro
diskouezh traoù personel ar vartoloded	marv	

Figure 1. Concordance lines for the word *marv*

Since the corpus will have to be divided into subcorpora, a definition must be given to the notion of subcorpus before corpus building work actually begins. Many dictionaries give information about the context in which a given word takes on such or such a meaning. The contexts are usually based on genres or registers. For spoken language corpora to be of any use to lexicographical work, they must at least allow for partitioning along those lines. However, results of a comparative study reported in Biber (1995) show that, in many languages, productions are not homogeneous even within a single genre. Any statistically relevant subcorpus must therefore be situated below the level of the genre. Biber (2004, pp. 176-178) gives further elements as to the situational and functional parameters which can be used to define subcorpora. For example, the first genre represented in KYG is the newscast genre. Within the genre, at least three types of speaker can be found:

1. newsreaders and voice-over presenters, who have had ample time to prepare their speech, and are often reading it from a prompt as they are talking or being recorded;

2. live reporters, who have had time to prepare notes, and must often improvise in the amount of time given;
3. witnesses/interviewees, who have not been able to prepare their answers, and have no control over the time they are given to speak.

Each set of speakers will produce utterances that may differ both in structure and lexis. For the newscast genre, each set represents a separate subcorpus, and the corpus must be designed in such a way as to allow partitioning between each of those subcorpora.

Methodology of corpus design: Basic aspects

Some aspects of corpus design remain unchanged no matter what the search tool happens to be. When dealing with spoken language corpora, one of the most important of these aspects is transcription. Indeed, since automatic language data retrieval is not yet possible directly from sound files, it can only be implemented on textual data. If, moreover, the corpus is to be made available to users other than the designers, much thought must be given to the transcription protocol, and particularly to the issues of phonetic versus orthographic transcription, spelling conventions, use of contractions, and annotation.

As mentioned in Baude (2006, p. 73 onwards), the issue which is bound to come up first is that of orthographic versus phonetic transcription, which is still much debated within the linguistics community. For many, orthographic transcriptions of speech can be equated with a form of corruption, and only phonetic transcriptions can accurately describe spoken language data. However, what is referred to as *phonetic transcription* can vary greatly among researchers: some will use broad phonetic transcriptions that could best be termed *phonemic*, others will use various shades of narrower transcriptions based on their understanding of underlying articulatory processes. Those who reject orthographic transcriptions should also reject broad phonetic transcriptions as they do not represent speech as it is pronounced but

rather as it is conceptualized. Only very narrow phonetic transcriptions which also include information about intonation and rhythm could serve a purpose here, but they have at least two drawbacks:

1. They use extremely complex symbols that cannot be processed easily by computers (Du Bois, 1991, p. 88).
2. No amount of detail in a phonetic transcription can take the place of actually listening to recordings or carrying out a computerized acoustic analysis of utterances (for more on phonetics and acoustics, see Ladefoged, 1996).

Keeping in mind those two drawbacks and the fact that transcribed speech corpus building is time-consuming (and therefore very costly), it seems much wiser to try to find a technical solution that would enable the corpus users to retrieve (legally, of course) relevant pieces of sound from the source data rather than invest considerable amounts of time in parallel phonetic transcriptions that would add technical difficulties while yielding potentially useless data. For all those reasons, it was agreed that KYG would be made up of orthographically transcribed documents in a way that would make it possible, at a later stage, to align segments of texts with their corresponding sound segments.

Having ruled in favor of orthographic transcriptions as being the lesser of two evils, two other elements of the transcription protocol must be clarified. The first one, which may not come up for all languages, deals with the type of the standard which is going to be used. Modern Breton writing has typically followed one of three spelling standards (Favereau, 2000, p. VII):

- *Skolveurieg* (academic)
- *Etrerannyezhel* (interdialectal)
- *Peurunvan* (unified)

To these the *Gwenedeg* (the southernmost dialect of Breton) version of *etrerannyezhel* may be added. Keeping in mind that the main goal of the project is to provide helpful information for advanced learners, and that most textbooks, grammars, and dictio-

naries are now making use of *peurunvan*, it seemed counter-productive to choose any other spelling standard, though the processing of *etrerannyezhel*, which makes use of fewer special characters (such as *n̄*) could have been technologically less difficult to implement.

The last element of the transcription protocol that requires clarifying is the use of contractions and other artifacts, the purpose of which is to make transcribed speech look similar to the way it was actually uttered. Studies as early as Ochs (1979) have shown that the use of artifacts, in addition to the higher degree of inconsistency it may generate, may surreptitiously add pragmatic information that the transcriber did not intend to add. The use of contractions also causes problems related to computer-assisted searching techniques. Contractions are usually shown by substituting an apostrophe for the contracted segment, for instance *ba'n ti* instead of *barzh an ti* (*in the house*). Supposing the corpus users are aware of such a fact, they may think about searching the corpus for instances of *ba*—instead of instances of *barzh*—if they want to learn about all the various contexts in which *barzh* is used in the spoken language. It would be easy to reset the parameters of the search engine in such a way as to consider blank spaces and apostrophes as word separators. Two major problems can however be foreseen. The first problem lies in the fact that contractions are not compulsory in the spoken language. Speakers may or may not choose to contract words. Moreover, there may be more than only one way to contract some of the longer words. If contractions are transcribed, a simple search will return parts of the data at a time but not all of the data at once. The second problem is that at least two Breton spelling standards make use of apostrophes in cases which are not related to contractions, i.e. to make a distinction between the <ch> cluster (pronounced as <sh> in English *ship*) and the <c'h> cluster (pronounced as <ch> in Scottish *loch*). If apostrophes were to be considered as word separators, a form such as *gourc'hemmenn* (*recommendation*) would be presented as two different words: *gourc* and *hemmen* (none of which is part of the Breton lexicon). For these reasons, it was agreed that KYG should use conventional *peurunvan* spelling without contractions or other related artifacts.

After having dealt with transcription issues, the next step concerns annotation. The choice of conventional spelling for transcripts represents an advantage for the common user but a substantial drawback for the researcher, who loses the pragmatic information contained in the speakers' pronunciation or intonation. To avoid such loss, corpus designers can insert the extra information in a separate yet complementary set of annotations. Some designers consider annotation as part of the transcription process and append the extra information to the words using an underscore symbol as separator, as shown in Figure 2.

labourioù_NoN2 bras_ADJoN2 zo_VPo bremañ_AVo

(lit. works_ big_ be_ now_)

Figure 2. An example of search result when using an underscore annotation scheme

However, the fact that the annotation is inserted as supplementary text means that, no matter who the end user of the corpus may be, any search will return results which will include full annotation. The potential complexity of such results will certainly be welcomed by researchers but will constitute a deterrent for other users, such as language learners for instance. As one of the main features of the KYG project is to give teachers and learners access to the corpus as it is being developed, it seemed preferable to place annotation in separate tags which may or may not be viewed by the user depending on their level of expertise. It was therefore decided that transcriptions should be encoded in the XML format, which is slightly more difficult to

process than the raw text (.txt) format but is supported by most corpus-searching tools and ensures user-friendliness, as shown in Figure 3.

Full view	<w ana="#NoN2">labourioù</w> <w ana="#ADJoN2">bras</w><w ana="#VPo">zo</w> ...
Partial view	labourioù bras zo bremañ

Figure 3. Viewing possibilities when using an XML annotation scheme

Search tools and corpus design implications

Other aspects of corpus design are highly dependent on the choice of the tool that will help the user search the corpus. In this case, the fact of choosing XML as the format for all incorporated documents implies having access to a transcription interface which outputs transcription files in the XML format. As the first genre to be transcribed was that of the newscast, the corpus designers opted for Transcriber (Barras et al., 1998), which was originally conceived as a newscast transcription application.

The choice of the search tool (concordancer) also had consequences for that of the annotation scheme. There are many open-source concordancers freely available for download but they do not all have the features required for the project. Among all the possibilities, the designers selected XAIRA for the following reasons:

1. It is designed to index XML files (XAIRA stands for XML Aware Indexing and Retrieval Architecture);
2. Its client/server configuration makes it possible to use it as a stand-alone application or as a network application;
3. It allows for partitioning provided the texts in the corpus are encoded according to the guidelines given by the Text Encoding Initiative (Burnard & Bauman, 2007);
4. It has already proved reliable when used with very large corpora such as the BNC or the ANC (American National Corpus).

Teaching and learning implications

Working with corpora in the classroom may imply approaching the issue of linguistic acceptability in a different way. Indeed, by trying to discover invariance through the building of decontextualized models of language, linguists may have sometimes forgotten that the constraints that shape utterances are not necessarily inherent to the model alone but are also due to contextual circumstances. For Coseriu (1969), contextual constraints make up what he calls the "norm space". Another way of stating the concept would be to say that an utterance is not necessarily acceptable or unacceptable (following a boolean type of logic, i.e. a choice between only two possible values), but rather more or less contextually expected (following a bayesian type of logic, i.e. a location within a range). The degree of expectedness depends very much on the speaker's own linguistic proficiency and habits. Very few speakers of any given language have access to the full extent of the lexicon or grammar of their language, and different speakers of the same language have access to different stocks of lexical or grammatical items. In the case of lexical items, although the various stocks may have a lot of items in common, the semantic charge they carry vary from speaker to speaker simply because native speakers of a language do not learn their usual vocabulary from a prescriptive source but from the normative source of their own speech communities. Likewise, in the case of grammatical items, similar structures may differ in their pragmatic values from one discourse community to the next. As a consequence, no representative corpus will ever give its user a clear-cut view of the language. Attested data will show patterns of lexical or grammatical usage which could start to look like

rules. At the same time, the same attested data will feature a small number of occurrences that do not follow those patterns, thereby breaking what were hoped to become rules. Before teachers make use of corpora in the classroom, learners should therefore have reached a level of language awareness at which the absence of formal rules is no longer a problem. This is precisely what is asked of advanced learners, and what should constitute the main difference between upper-intermediate learners and advanced learners.

Learners who will be exposed to corpora may also benefit from being introduced to the notions of collocational framework and lexico-grammatical patterns. In other words, they should be made aware of the fact that, as suggested in Halliday (1991, p. 32), grammar and lexis can be seen as two ends of the same linguistic continuum rather than two distinct systems. Both Halliday (1991) and Sinclair (1991), for instance, show that lexical items are not selected randomly from an open list on the sole basis of their grammatical type. A speaker's choice of one particular item constrains that of others in its vicinity, thereby further limiting the alleged infinitely creative potential of language. This phenomenon, termed collocation by Sinclair, is easily perceived in the case of fixed expressions (De Cock, 1998) or extended metaphors (Lakoff & Johnson, 1980), and can be made apparent in all speech data through the viewing of concordance lines (cf. Figure 3). Although it is important to show how concordancing can be used at this point, the reader should bear in mind that the corpus from which the concordancer draws its data is still in the making, and that any example that could be shown here is yet far from reflecting the final product that a full spoken language corpus can give.

While running comparative frequency counts on Breton and French newscasts, since the concordancing tool also automatically provides word frequency counts, Breton newscasting was noticed to contain comparatively very few occurrences of words which refer to death, i.e. Breton *marv*, French *mort(s)*. The results

of a query on the word *marv* ran as shown in Figure 3. Concordancing shows that, out of five occurrences of *marv*, two are part of the expression *marv evit ar vro* (lit. *dead for the country* > *dead in the line of duty*). Although there are not yet enough data to weigh the relative importance of such an expression within all the genres of spoken Breton, it is interesting to note that an advanced learner of Breton—who is assumed here to be a native speaker of French—does not necessarily have direct access to it through a bilingual dictionary entry. Favereau (2000), for instance, gives many expressions related to *mort*—as shown in Figure 4—but the expression *mort pour la patrie* (lit. *dead for the homeland*) is not one of them. Learners may still try to resort to a more literal style of translation by finding separate equivalents for each word, in which case they will certainly produce the less idiomatic *marv evit ar vammvro* because *mammvro* (*homeland*) is the clearest equivalent they will find for the French *patrie*. Furthermore, it may be remarked that the same concordance lines give instances of the preposition *diwar* that learners will probably not expect, since they usually take *diwar* as an equivalent for *à propos de* (*regarding*) and not as a possible equivalent for *des suites de* (*as a result of* | *fol-*

French	***Breton***
à la mort	e par ar marw
entre la vie et la mort	en e enkoù
mort de fatigue	skuizh-marw
ivre-mort	mezw-dall
mort de peur	dindan e aon
faire le mort	ober ar marw bihan
se donner la mort	en em zistrujañ (cf. klask e varw)
frapper à mort	skoet marw

Figure 4. French-Breton equivalents for expressions containing *mort*

lowing)—although Favereau (2000, p. 1302) suggests such use.

Perspectives

Before moving on to other genres, KYG's designers would like the newscast genre to comprise at least four hours' worth of transcribed material. As soon as this first goal is reached, the next genre to be dealt with will be that of talk shows, both from radio and webTV. Two other genres are also considered: college lectures and casual conversations. The former should be rather easy to obtain and transcribe. The latter, which may also turn out to be the most interesting of all, will bring many technical difficulties, and possibly some legal ones as well. Special care will be given to the documentation of the corpus for it is hoped that, once its use has gathered momentum within the teaching community, other researchers and teachers will like to contribute to its development.

The ultimate goal is, of course, to use the corpus material to produce reference books (dictionaries and grammars) and learning resources for advanced learners, which take into account the specificities of the spoken language. Since such resources require giving access to a large number of examples taken from the recorded data, paper editions may not be the most adequate format. The editors would indeed be hard-pressed to find publishers willing to produce a two-thousand-page grammar of Breton or a six-thousand-page bilingual dictionary. It should therefore be worth considering the electronic alternative through the use of software such as Tshwanelex, which is also TEI-XML compatible.

Finally, collaborating with researchers in natural language processing (NLP) could be mutually beneficial to both parties. Most language models designed by NLP researchers are based on stochastic methods. NLP researchers are aware of the fact that linguistic data are highly skewed. To avoid the impact of the skew on their stochastic models, they have had the habit of restraining their experiments to very specific communication tasks within a particular genre. When they reach satisfactory

results in one or more of those tasks, they usually think about testing the robustness of their models in task-independent environments. However, they very often lack information about the kind of linguistic variation to which their models will be exposed. Corpus analysis can help define variation in a way that is suitable for NLP methods. In return, corpus analysis can benefit from NLP research. Sufficiently robust stochastic language models can help enrich the annotation of a corpus automatically. For example, results obtained from a first version of KYG will provide information that will help set parameters for an automatic part-of-speech tagger originally developed for English (Mason, 2000, pp. 195-210). In return, the tagger will automatically add part-of-speech information in the annotation, thereby making it possible to fine-tune the more specific study of linguistic variation not only across genres but also across the subcorpora of a given genre. Eventually, as finer linguistic analysis can help widen the scope of robustness of NLP models, the latter could be included in multimedia language applications such as real-time speech synthesis or automatic speech recognition.

References

Barras, C. et al. (1998). Transcriber, a free tool for segmenting, labeling and transcribing speech. In *Proceedings of the 1st Conference on Language Resources and Evaluation (LREC 98)*, 1373-1376.

Baude, O. (2006). *Corpus oraux: guide des bonnes pratiques 2006*. Paris: CNRS éditions.

Biber, D. (1993). Using register-diversified corpora for general language studies. *Computational Linguistics, 19(2)*, 243-258.

Biber, D. (1995). *Dimensions of register variation: a cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, D. (2004 [1992]). Representativeness in corpus design. In Sampson, G. & McCarthy, D. *Corpus Linguistics: readings in a widening discipline* (pp. 174-197). London: Continuum.

Burnard, L. & Bauman, S. (2007). *TEI P5: Guidelines to Electronic*

Text Encoding and Interchange. Oxford: www.tei-c.org/release/doc/tei-p5-doc/en/html/.

Chafe, W. et al. (1991). Towards a new corpus of spoken American English. In Aijmer, K. & Altenberg, B. *English Corpus Linguistics* (pp. 64-82). London: Longman.

Coseriu, E. (1969). *Sistema, norma et parola, studi linguistici in onore Vittorio Pisani*, Brescia: Paideia Editrice.

De Cock, S. (1998). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. In *International Journal of Corpus Linguistics, 3(1),* 59-80.

Desbordes, Y. (1990). *Petite grammaire du breton moderne*, 2e édition, Lesneven: Mouladurioù Hor Yezh

Du Bois, J. (1991). Transcription design principles for spoken discourse research. In *Pragmatics 1:1*, 71-106.

Favereau, F. (1997). *Yezhadur ar brezhoneg a-vremañ*, Montroules: Skol Vreizh

Favereau, F. (2000). *Geriadur ar brezhoneg a-vremañ*, Montroules: Skol Vreizh

Greenbaum, S. & Quirk, R. (1970). *Elicitation Experiments in English: Linguistic Studies in Use and Attitude*. London: Longman

Halliday, M.A.K. (1991). Corpus studies and probabilistic grammar. In Aijmer & Altenberg, *English Corpus Linguistics* (pp. 30-43). London: Longman.

Halliday, M.A.K.; Teubert, W.; Yallop, C.; Cermakova, A. (2004). *Lexicology and Corpus Linguistics: an introduction*. London : Continuum.

Hill, A. (1961). "Grammaticality" in *Word, 17*, 1-10.

Hjelmslev, L. (1961). *Prolegomena to a Theory of Language*. Madison: University of Wisconsin press.

Kervella, F. (1947). *Yezhadur bras ar brezhoneg*. Al Liamm: Ar Baol.

Labov, W. (1975). *What is a linguistic fact?* Lisse: Peter Ridders Press.

Ladefoged, P. (1996). *Elements of acoustic phonetics*, second edition. Chicago: University of Chicago Press.

Lakoff, G. & Johnson, M. (1980). *Metaphors we live by*. Cambridge: Cambridge University Press.

Mason, O. (2000). *Programming for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Oakes, M.P. (1998). *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Ochs, E. (1979). Transcription as theory. In Ochs, E. & Schieffelin, B. *Developmental pragmatics* (pp. 43-72). New York: Academic Press.

Quirk, R. & Svartvik, J. (1966). *Investigating linguistic acceptability*. The Hague: Mouton.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Spencer, N. (1973). Differences between linguists and non-linguists in intuitions of grammaticality-acceptability. In *Journal of Psycholinguistic Research, 2,* 83-98.

Svartvik, J. (1992). *Directions in Corpus Linguistics*. Berlin: Mouton de Gruyter.

Timm, L. (2005). Language ideologies in Brittany, with implications for Breton language maintenance and pedagogy. In *Journal of Celtic Language Learning, 10,* 34-42.

Zipf, G.K. (1932). *Selected studies of the principle of relative frequency in language*. Cambridge (Mass.): Harvard University Press.

Common European Framework: *http://www.coe.int/lang/*
Lextutor : *http://www.lextutor.ca*
MiCASE : *http://quod.lib.umich.edu/m/micase/*
Text Encoding Initiative: *http://www.tei-c.org*
Transcriber : *http://trans.sourceforge.net*
Tshwanelex: *http://tshwanedje.com/tshwanelex/*
XAIRA: *http://xaira.sourceforge.net*

Notes

(1) Corpus of Spoken Breton

(2) Timm makes a distinction here between, on the one hand,

speakers who, prior to the 1960s, were raised in a Breton-speaking environment, learned French at school at a later stage, and have gone on speaking Breton usually only in informal situations (paleo-speakers) and, on the other hand, speakers who belong to younger generations, have learned Breton at school, but do not necessarily live in a Breton-speaking environment (neo-speakers).

(3) Yann Desbordes added one chapter—just as short as the others—relative to the main rules of Breton pronunciation. In keeping with the rest of the book, it is normative. But the norm it depicts is based on competent observation of modern Breton. (translated by the author from the French: "Yann Desbordes a ajouté un chapitre—toujours succinct—sur les principales règles de la prononciation du Breton: comme le reste de l'ouvrage, il est normatif. Mais d'une normativité basée sur l'observation compétente du breton Moderne."

Ethnicity of Irish Language Learners in Canada

Aralt Mac Giolla Chainnigh
The Royal Military College of Canada

The ethnicity of students studying Irish as an optional credit-course in a Canadian high school between 1997 and 2004 is discussed. No positive correlation is found between Irish ethnicity and enrollment in the course. Instead, statistics tend to reflect the ethnic mix of the community at large. This finding may have implications for marketing Irish, and perhaps other Celtic languages, to students of high school age in North America.

Introduction

Irish was taught as a credit course for ordinary, day-time, high school students at Kingston Collegiate and Vocational Institute (KCVI) in Kingston, Ontario, Canada in the years 2000, 2002 and 2004. The number of students during these years was, respectively: 15, 18, and 22. The Irish course was part of a full-semester Celtic Studies Program, which involved four separate courses: Celtic Literature, Celtic History, Celtic Music, and Irish. Students from high schools across the school district, both Public (Limestone District School Board) and Catholic (Algonquin and Lakeshore Catholic District School Board), were eligible to enroll in the program, and were bussed free of charge to KCVI. The ethnic distribution of the students taking the course is of interest, since this is the first time to our knowledge that Irish has been offered as a credit course at an Ontario high school.