http://dx.doi.org/10.18842/klaces.2016.12.3.1

Speech Corpora for Foreign Language Education: Methods and Tools[*]

Julien Eychenne(한국외국어대학교)

<Abstract>

줄리앙 에센. 2016. Speech Corpora for Foreign Language Education: Methods and Tools. *The Language and Culture* 12-3: 1-26. 본 연구는 현대의 기준에 맞추어 잘 디자인된 발화 코퍼스의 구축 및 활용이 외국어 교육에 있어 교사, 학습자 및 연구자들에게 구체적으로 어떤 도움을 줄 수 있는지 보이고자 하였다. 이를 위해 400명에 이르는 프랑스어권 모국어 화자의 발화를 체계적인 프로토콜 하에서 녹음한 자료로 이루어진 대규모 코퍼스 프로젝트인 Phonologie du Français Contemporain(PFC), 그리고 이를 모체로 하여 출발한 프로젝트인 Interphonologie du Français Contemporain(IPFC)에 대해 소개한다. 특히 IPFC는 다양한 나라에서 프랑스어를 배우는 학습자들의 발화를 PFC와 동일한 프로토콜 아래에서 녹음하여 구축한 학습자 말뭉치로서 말뭉치 구축을 위하여 사용하는 방법론, 프로젝트의 중요한 성과 및 구축 과정에서 대두된 이슈와 한계점 등을 제시하고, 이 같은 프로젝트가 제2외국어 학습 및 교육 측면에 어떤 시사점을 던져줄 수 있는지 논의한다. 마지막으로 두 프로젝트에서 구축된 발화 코퍼스를 바탕으로 이루어진 연구들에서 산출된 여러 교수학습자료들과 발화 말뭉치를 분석하기 위해 개발된 프로그램인 Dolmen을 소개한다. (Hankuk University of Foreign Studies)

Keywords: speech corpus(발화 말뭉치), annotation(주석), French(프랑스어), Dolmen(돌멘), data-driven learning(데이터 기반 학습)

[*] This research was supported by the National Research Foundation of Korea Grant funded by the Korean Government (Project Number: NRF-2015R1A5A7036384). It was presented at the 22nd Spring Conference of the Korean Language and Culture Education Society.

1. Introduction

Over the last 15 years or so, there has been a growing interest in the use of speech corpora, as evidenced by the emergence of the field known as corpus phonology (see Durand, Gut & Kristoffersen (eds), 2014, and contributions therein). Corpus-based phonology has allowed researchers to gain new qualitative and quantitative insights that go beyond what can be obtained solely on the basis of casual observation or introspective judgments, both for native and non-native speech. In parallel, the field of non-native corpus-based research has experienced a tremendous growth, with the development an international scholarly society (the Learner Corpus Association, which contains a large searchable database), two recent handbooks (Granger, Gilquin & Meunier (eds), 2015; Colantoni, Steele & Escudero, 2015) and a new dedicated journal (the International Journal of Learner Corpus Research, launched in 2015 and published by John Benjamins). There are now several large written corpora available, such as the International Corpus of Learner English, which contains about 2.5 million words and covers no less than 11 mother tongues (Granger, Dagneaux & Meunier, 2002). However, there are still few spoken corpora available today, which can undoubtedly be attributed to the fact that building a speech corpus involves a significant investment in time and resources, and requires in addition some expertise in data acquisition and processing. Given this relatively high cost, instructors and practitioners may understandably be reluctant to investing significant resources in the creation of a new corpus. Nevertheless, I will try to demonstrate in this contribution that the initial investment is well worth the effort and that well-designed corpora can benefit instructors, students and researchers altogether.

2. Spoken corpora in first and second language research

In order to adequately frame the discussion, it is necessary to first define what we mean by spoken corpus. I will take as a starting point Gut & Voormann's (2014) definition of a phonological corpus, which according to these authors consists of:

> - *primary data* in the form of audio or video data;
> - *phonological annotations* that refer to the raw data by time information (time-aligned); and
> - *metadata* about the recordings, speakers and corpus as a whole.
>
> (Gut & Voormann's 2014: 16)

Primary data (or raw data) consist of a set of audio of video files which have been assembled into a relatively coherent whole. Although the data may have been collected for a specific purpose, Gut & Voormann argue that this should not be a required condition, and indeed many corpora are used for purposes others than those they were created for. A phonological annotation represents time-aligned textual information associated with the primary data. Annotations can take many forms, such as orthographic transcriptions, phonetic transcriptions using the International Phonetic Alphabet or, more generally, any kind of symbolic coding that might provide useful information about the content of the primary data. Needless to say, annotations need not be strictly phonological and may denote information about any level of linguistic structure (parts of speech, syntactic constituents, etc.). Finally, metadata are information that link primary data and phonological annotations together or indicate structural relations about the data which enable users to manage, organize and make use of the corpus.

Given such a broad definition, there are many ways the concept of ‘corpus’ can be operationalized in practice; in the remainder of this section, I discuss a number of motivations and design principles that can guide scholars in the creation (or selection) of a specific type of corpus.

2.1 Why build (and use) a speech corpus?

The use of speech corpora can bring several important benefits to the field of foreign language education, and we can identify three types of audience that can take advantage of them: researchers, instructors and students. First, the use of learner corpora can allow researchers to systematically and objectively assess language learners’ pronunciation, which in turn makes it possible to develop targeted teaching methods that can be tailored to specific learner populations. For example, Gut (2009), in a landmark corpus analysis of the speech of 101 non-native learners of English and German, showed that there exist important interactions between the different components of language acquisition (phonology, morpho-syntax, lexicon and fluency). This study revealed that there are strong connections between lexical richness, morphosyntactic complexity and general fluency. However, the relation between phonological domain on the one hand, and the morphosyntactic and lexical domains on the other, appeared to be less systematic and weaker. Such phenomena can only be discovered through the careful analysis of first-hand data.

Second, instructors can also take advantage of speech corpora in order to obtain authentic linguistic material to complement standard recorded material. Although this would appear to put an unnecessary burden on the learner, because such material may be more challenging than traditional material, there are some clear advantages. Eisenstein (1986) insists on the limitations of teaching a single 'standard' form of language, especially for populations of learners living in large (multi-ethnic) urban centers. For example, in New York, non-native speakers are exposed to no less than three distinct varieties: Black English, New Yorkese (a non-standard variety specific to New York), and the regional standard. Being able to understand these varieties is essential for learners to successfully negotiate social interactions in their environment. More generally, as pointed out by Detey (2009), who draws from results in phonetics and psycholinguistics, it seems that exposing learners to a rich and varied linguistic input can potentially help them build more robust phonological categories in the target language. For example, Iverson, Kuhl, Akahane-Yamada, Diesch, Tohkura, Kettermann & Siebert (2003) suggest that teaching the English contrast between /r/ and /l/ (cf. *rice* vs *lice*) to Japanese adults is more effective when learners are exposed to multi-talker, highly variable stimuli rather than an impoverished input. Similar effects have been found with native speakers as well. Clopper & Pisoni (2004) tested two groups of native speakers of English in a dialect perception experiment. Both groups were exposed to six different dialects of American English. The first group listened to only one talker per dialect, whereas the second group was exposed to three talkers per dialect. When tested on material to which they had been exposed during the training session, the group which had been exposed to one talker performed better; however, when tested on new stimuli, the group which had been exposed to more talkers performed better than the other group, suggesting that they generalized better to new inputs. Such results suggest, perhaps unsurprisingly, that both native and non-native speakers use similar cognitive mechanisms to build phonological categories, and that the perceptual categories that they build are more resistant to noise when they are inferred from a variable input. In other words, when the input is too homogeneous, the resulting categories tend to be too finely tuned to the input, a phenomenon known as *overfitting* in the statistical learning literature.

Finally, learners themselves can also benefit from using corpora, provided that these are made accessible in a suitable form, such as a concordancer to examine collocations. Granger (2008) discusses the benefits of contrastive interlanguage analysis, which consists in extracting collocations and letting learners compare L1 and L2, or L2 and L2 productions. For instance, Granger (2008) showed that the English verb *argue* appeared in a much more limited set of contexts in learners' productions than in native speakers'; letting students proactively analyze linguistic constructions in this way can help draw their attention to patterns that they might otherwise be unaware of. Indeed, this and other forms of *data-driven learning* (e.g. Boulton & Tyne 2013) has been argued to increase students' autonomy and develop their linguistic awareness. Tyne (2009) reports on an experiment which consisted in developing a corpus linguistics course tailored to second-language learners of French, where learners had to collect, transcribe and analyze authentic French data. Although the sample size was small (N=10), a post hoc evaluation of the course by students revealed that the experience had been overall very positive, many students observing that it had increased their awareness of a number of linguistic patterns (e.g. use of discourse markers, variable realization of the 'schwa' vowel).

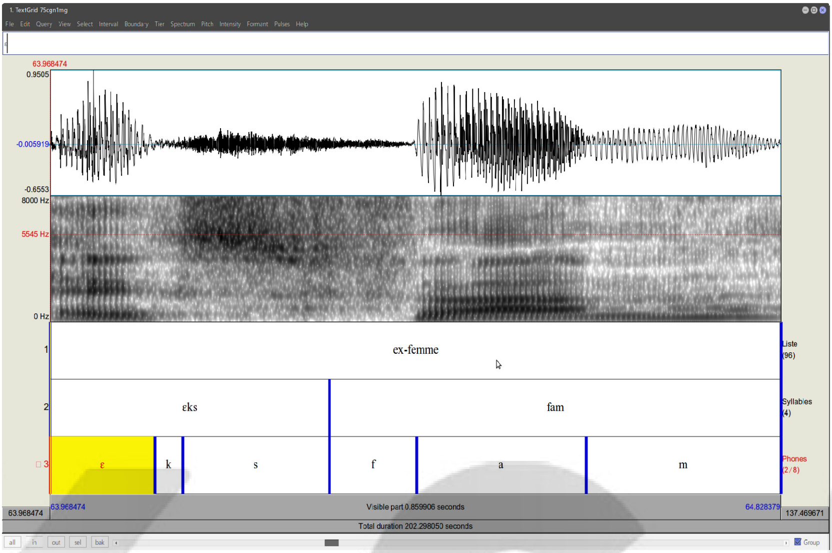
All things considered, it is clear that there are multiple benefits to the use of (spoken) corpora in foreign language education and teaching. It is therefore worth reviewing some important guidelines that can help in the construction of a speech corpus.

2.2 How to build a speech corpus?

Given the definition of a speech corpus that we have adopted at the beginning of this section, the first step that needs to be taken in the construction of a corpus is the acquisition of raw data. We can identify two broad types of data, *naturalistic* and *experimental*, which are probably best understood as the two ends of a continuum (Chaudron 2003: 763-765). Naturalistic data, such as guided interviews and conversation with a peer, can yield (relatively) spontaneous forms of speech, at the expense of comparability since there may be discrepancies among subjects in terms of fluency, lexical and grammatical richness, or talkativeness among other factors. On the other hand, experimental data allows for a very precise comparison of a number of subjects by rigorously controlling and selectively manipulating the variables of interest. This comes at the expense of authenticity and representativity, since the performance of a learner in a specific experimental setting may not accurately reflect their linguistic competence in a real communicational environment. (For example, learners who are highly reliant on pragmatic cues to process language may perform comparatively poorer than others in a decontextualized task.) A good compromise is to adopt a hybrid approach and collect both naturalistic and experimental data. This is the strategy that was adopted in the two projects that will be presented in section 3.

The next step, which can be the most time-consuming one depending on the nature of the corpus, is to annotate the raw data. Spoken corpora differ from written corpora in that they need to be time-aligned: elements of annotation are bound to a specific time point or stretch in the sound signal. For many purposes, it is useful to have at least one level of annotation in (quasi-)standard orthography, so as to be able to efficiently find and retrieve relevant subparts of the corpus. There exist a large number of tools for annotation (see Durand, Gut & Kristoffersen (eds), 2014: Part II), but the most popular one is Praat (see www.praat.org). Praat allows to annotate a sound file on a number of *tiers*, which are independent layers of annotation, where each tier represents either a series of time points or a sequence of intervals. Figure 1 shows an example annotation for the French word *ex-femme* ‘ex-wife’. From top to bottom, we can see the speech waveform, a wideband spectrogram, and three hierarchical levels of annotation (word, syllable and phoneme). Because Praat’s TextGrid annotation format is very simple and is stored in plain text files, it can be easily imported into other programs. However, it has a number of limitations. First, it offers no way to express structural relations between annotation elements. Thus, there is no direct way to represent hierarchical structure, as is commonly done in phonology or syntax (segment < syllable < word, or word < phrase < utterance): one needs to use a cumbersome and error-prone approach to represent such structure in a TextGrid file[1]. Non-local dependency relations, such as a disfluency and its repair (e.g. *I saw his fa/, uh I mean, his brother*) which are extremely common in spontaneous speech, are even more difficult to represent in this format. Finally, Praat currently offers no way to associate arbitrary metadata with a corpus

[1] While the annotation in Figure 1 seems to be hierarchical, this is only apparent. Interval boundaries happen to be aligned in such a way that they represent hierarchical structure, but nothing in Praat’s annotation format ensures that the hierarchical structure is correct.



<Figure 1> Annotation of a speech segment in Praat

Several efforts have been carried out to remedy such limitations (see Romary & Witt, 2014 for an overview of existing standards). Bird & Liberman (2001) developed a common framework for linguistic annotation. They observed that all existing data formats (including Praat's TextGrid) could be described using *annotation graphs*. For speech annotation, nodes in the graph represent time points, whereas edges, which are labeled, represent annotation elements. A more recent effort, which builds in part upon Bird & Liberman's (2001) approach, is the *Linguistic Annotation Framework* (LAF) (see Ide & Suderman, 2014). This is a recent standard specifically developed to address the needs of linguistic annotation, and which was published as an ISO standard (ISO 24612:2012). This specification describes how a linguistic annotation can be represented using a graph-based format. It also mandates that the data be encoded using Unicode for character encoding and XML (eXtensible Markup Language) for structural information, two widely used standards which facilitate the exchange and long-term conservation of data. This new standard offers a promising way of facilitating the exchange of data between applications, but given that it is so recent, it has not yet been widely adopted.

Finally, most corpora are accompanied by metadata, which are normally stored alongside the data and encode information about subjects, tasks, annotators, revision history, etc. The two most common methods of encoding metadata are structured text files in XML format (Broeder & van Uytvanck, 2014) and databases built on top of the Structured Query Language (SQL) standard. There are a number of free, open source SQL engines available, and they are usually used via a web-based and/or desktop user interface (see for example Eychenne, Navarro, Tchobanov & van Leussen, 2016).

A speech corpus built upon these guidelines will be inter-operable and easier to use, distribute and maintain on the long term. The next section briefly introduces two related corpora that have been designed according to these principles.

3. Illustration from French: the PFC and IPFC projects

3.1 Phonological variation in contemporary French: the PFC project

The project entitled “Phonologie du français contemporain : usages, variétés, structure” [Phonology of Contemporary French] (PFC) (see

Durand, Laks & Lyche, 2009), coordinated by M.-H. Côté (Laval, Québec), J. Durand (Toulouse, France), B. Laks (Paris, France) and C. Lyche (Oslo, Norway) is an international research programme that federates over 50 researchers and graduate students around the world. The project was initiated in the late 90's to offer a systematic and comprehensive description of the varieties of French spoken in French-speaking areas of the world (mainly Europe, Africa, North America and a number of overseas territories). The project now has over 40 survey points, representing speech gathered from more than 400 speakers. Most of the data have been transcribed and annotated, and are available through a dedicated website (http://www.projet-pfc.net).

Each survey point represents a balanced sample of about 10-12 subjects, with an equal number of men and women spread across two or ideally three age groups. All speakers are recorded following the same survey protocol, which comprises four tasks: a word list, a text that looks like a brief newspaper article, a guided interview and free conversation[2]. The word list contains 94 items which have been carefully selected to establish the subject's phonological inventory and to test the presence or absence of well known phonological contrasts, such as the opposition between a front low vowel in *patte* 'leg' vs a back low vowel in *pâte* 'dough', as well as a number of other phenomena (e.g. voicing assimilations). The text contains many of the minimal pairs present in the word list, but allows to study additional phenomena typically found in connected speech. The guided interview is usually

[2] The fact that this protocol includes reading tasks has been criticized because it cannot be applied to speakers who only know French as a spoken language. This unfortunately excludes a number of socio-demographic groups from the pool of communities that can be surveyed. However, it was felt that given the importance of writing in most French-speaking societies, it was necessary to include some reading tasks.

conducted by one of the researchers and typically contains a number of questions about work, family, neighborhood, etc. Finally, the free conversation is usually a discussion between two to three subjects together, on any topic. Each interview lasts from 15 to 30 minutes, and collectively they are supposed to provide two types of speech (formal and informal, respectively), although this has not always been achieved in practice. This standard protocol can be augmented by additional tasks. For instance, surveys conducted in Canada include an additional word list, which contains words designed to specifically test aspects of the phonology of Canadian French, such as the devoicing of high vowels.

All recordings are transcribed in standard orthography using Praat, and are then coded for two phonological phenomena important in French phonology, known as 'schwa' and 'liaison'. Although the project was originally conceived to study segmental phenomena in phonology, a number of studies have extended the core corpus to take into account additional phenomena such as morphosyntax (Christodoulides, Avanzi & Goldman, 2014) and prosody (e.g. Avanzi, Schwab, Dubosson & Goldman, 2012). Lack of space precludes a full overview of this project, but interested readers are referred to Gess, Lyche & Meisenburg (eds)(2012) and Detey, Durand, Laks & Lyche (eds)(2016), and the contributions to these volumes, for more thorough presentations of the project as well as linguistic analyses based on the PFC corpus.

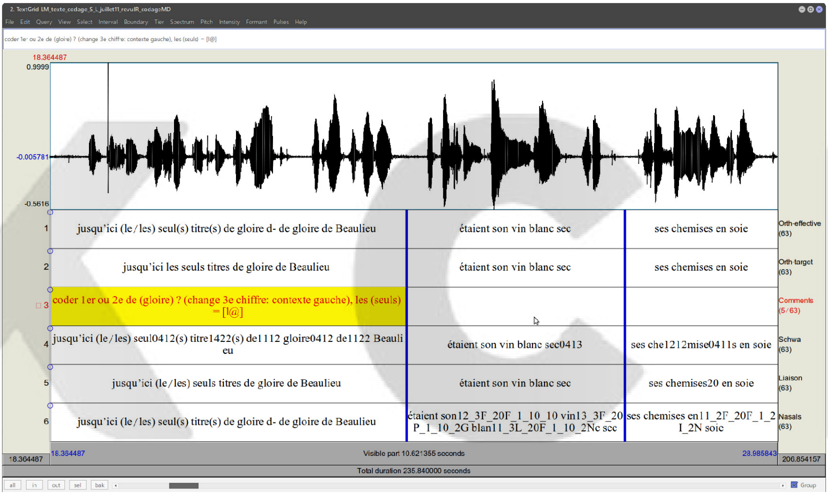
3.2 French as a foreign language: the IPFC project

A few years after PFC was started, a satellite project was launched to study the phonology of non-native speakers (and learners) of French. This international endeavour, entitled "Interphonologie du français contemporain" [Interphonology of Contemporary French] (IPFC) (see Detey, Racine, Kawaguchi & Eychenne (eds), 2016 for an overview), is coordinated by S. Detey (Tokyo, Japan), I. Racine (Geneva, Switzerland) and Y. Kawaguchi (Tokyo, Japan). It involves more than 50 researchers and contains data from no less than 16 learner populations (including native speakers of German, Spanish, English, Brazilian Portuguese and Turkish among others). See the project's website: http://cblle.tufs.ac.jp/ipfc/.

Following the same methodological principles as PFC, all subjects are recorded following a similar protocol, which is slightly adapted to take into account aspects specific to each population of learners. The protocol is progressive, and contains the following tasks: repetition of a word list, repetition of the same word list, reading of the PFC word list and text, guided interview with a (near-)native interviewer, and a 'semi-guided' interaction with other non-native speakers on a predefined topic. Not all populations of learners are able to perform all tasks, but advanced learners are expected to be able to perform all of them.

As in the PFC project, the data are orthographically transcribed using Praat and are coded to analyze a number of phonological phenomena. Sub-projects can decide to selectively apply some of the codings to their corpus, depending on the learning difficulties faced by the learners. As can be expected, not all phenomena pose the same difficulties to all learners. Syllabic structure, for example, is particularly challenging for Korean and Japanese learners but is virtually unproblematic for native speakers of Dutch. Nasal vowels, on the other hand, raise similar challenges for many learners since most languages do not have such vowels (see Detey, Racine, Eychenne & Kawaguchi, 2014 for an analysis of the acquisition of nasal vowels by Japanese learners based on IPFC data).

Data coding is currently performed by one, or sometimes two native speakers of French. Each coding contains a number of fields, several of which correspond to a subjective evaluation of the learner's production by the native coder. For instance, the coding scheme for nasal vowels includes, among other things, a subjective evaluation of the acoustic quality of the vowel, the presence or absence of a nasal appendix, the consonants surrounding the vowel and the position in the word.


2. TextGrid LM_texte_codage_S_L_juillet11_revuIR_codageMD
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
18.364487
0.9999
-0.005781
-0.5616
18.364487
Visible part 10.621355 seconds
28.985843
206.854157
Total duration 235.840000 seconds
all in out sel bak
Group

coder 1er ou 2e de (gloire) ? (change 3e chiffre: contexte gauche), les (seuls) = [l@]

1	jusqu'ici (le/les) seul(s) titre(s) de gloire d- de gloire de Beaulieu	étaient son vin blanc sec	ses chemises en soie	Orth-effective (63)
2	jusqu'ici les seuls titres de gloire de Beaulieu	étaient son vin blanc sec	ses chemises en soie	Orth-target (63)
□ 3	coder 1er ou 2e de (gloire) ? (change 3e chiffre: contexte gauche), les (seuls) = [l@]			Comments (5/63)
4	jusqu'ici (le/les) seul0412(s) titre1422(s) de1112 gloire0412 de1122 Beaulieu	étaient son vin blanc sec0413	ses che1212mise0411s en soie	Schwa (63)
5	jusqu'ici (le/les) seuls titres de gloire de Beaulieu	étaient son vin blanc sec	ses chemises20 en soie	Liaison (63)
6	jusqu'ici (le/les) seul(s) titre(s) de gloire d- de gloire de Beaulieu	étaient son12_3F_20F_1_10_10 vin13_3F_20P_1_10_2G blan11_3L_20F_1_10_2Nc sec	ses chemises en11_2F_20F_1_2I_2N soie	Nasals (63)

Figure 2. Annotation of a fragment of the PFC text read by a Japanese learner

<Figure 2> illustrates a fragment of the PFC text read by a Japanese learner, and coded for 3 phonological phenomena, namely the realization of schwa, liaison, and nasal vowels.

Both the PFC and IPFC projects offer useful frameworks to systematically analyze phonological variation in first and second language respectively. In addition, the fact that their experimental protocols partially overlap also allows one to compare non-native and native productions in read text (using the PFC word list and text). In the next sub-section, I provide an overview of the ways in which these corpora have been used in second language education to design innovative teaching resources for French.

3.3 Creating pedagogical resources from corpora

The PFC and IPFC corpora have been used in a significant amount of research, including several collective publications, journal articles, as well as a number of Masters’ and PhD theses. I will briefly discuss three collective publications that have been designed mostly for learners of French and teachers of French as a foreign language, and to which I contributed.

The first volume is entitled *Les variétés du français parlé dans l’espace francophone : ressources pour l’enseignement* [Varieties of French spoken in the French speaking world: resources for teaching] (Detey, Durand, Laks & Lyche (eds), 2010), is a book accompanied by a DVD containing 2:30 hours of recording. After a first part introducing general notions of linguistics, the book contains several parts covering the major French Speaking areas in the world (Northern France, Southern France, Belgium, Switzerland, Africa and North America). Each part contains several chapters, each of which is an analysis of an excerpt of an interview involving a native speaker representative of the variety. Each interview is transcribed orthographically, and commented at the phonological, lexical and syntactic levels. Excerpts are designed to be used in classroom, to introduce (advanced) learners to linguistic variation, syntactic patterns found in spoken French, as well as cultural aspects specific to each French-speaking area.

The second book (Detey, Racine, Kawaguchi & Eychenne (eds)(2016) to appear) is entitled *La prononciation du français dans le monde : du natif* à *l'apprenant* [The pronunciation of French: from native speakers to learners] contains a phonological description of the main varieties of French, along with descriptions of the phonology of many populations of learners (including English, Vietnamese, Korean, Arabic, Serbo-Croatian among many others). For each language, an overview of the phonological system is first provided, after which authors outline the main difficulties faced by speakers of this language in their acquisition of French. Each chapter is accompanied by sound samples illustrating the phonology of each language and the pronunciation of French by native and non-native speakers. Many sound samples are drawn from the PFC and IPFC corpora. This book was designed with two main use cases in mind: first, it allows French teachers to get a systematic and concise overview of the most important phonological characteristics of the main varieties of spoken French, going beyond what is found in typical textbooks on pronunciation. Second, it allows teachers, who are increasingly dealing with students from diverse backgrounds, to easily find information about the main pronunciation difficulties that their students might be facing, depending on their mother tongue.

The last book, entitled *Varieties of Spoken French* (Detey, Racine, Kawaguchi & Zay (eds), 2016), is written in English and is geared towards an international audience. The book contains three parts. The first part covers general concepts and approaches related to the study of variation in spoken French. The second part contains chapters describing excerpts of interviews from speakers representative of all the main varieties of French. The transcribed excerpt is available via an interactive web-based interface, so that users can click on the text and play the corresponding sound. This material has been formatted to be usable in the classroom. The last part is a more advanced introduction to the analysis of phonological intra- and inter-individual variation, focusing on several representative varieties. The book is accompanied by a companion website, accessible from the publisher's website. In addition to the multimedia versions of the book's chapters, it includes a large subset of the PFC database, as well as two computer programs that can be used to explore and analyze the data: Praat, which has already been mentioned, and Dolmen, a computer program that I have developed for the analysis of spoken corpora (see next section). These data and programs have been included in order to encourage students to explore phonological variation on their own or under the guidance of an instructor.

These three books show how results and data drawn from existing corpora, originally designed for research purposes, can be integrated into innovative teaching resources. In the next section, I offer a brief overview of the Dolmen computer program and show how it can be useful for Foreign Language Education researchers and teachers.

4. Analyzing speech corpora with Dolmen

Dolmen is a computer program for the analysis of aligned speech corpora. It runs on all major platforms (Windows, Mac OS X and Linux) and is freely available under an open-source license. The program can be downloaded on the author's website[3]. Although Dolmen is used within the PFC and IPFC projects, it is not tied to these particular projects and can be used with any speech corpus, provided that the annotations are stored in a supported format (for example, files created

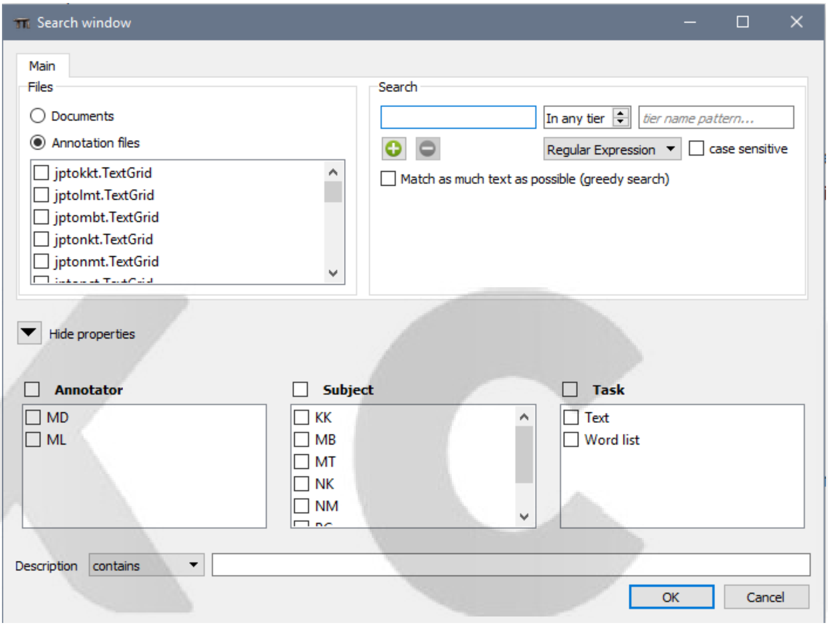
[3] http://www.julieneychenne.info/dolmen

by Praat or WaveSurfer). This section describes Dolmen 2.0, the new version which is currently under development and is due to be released in September 2016. See Eychenne, Navarro, Tchobanov & van Leussen (2016) for a general overview of the program, and Eychenne & Paternostro (2016) for a practical introduction.

Dolmen's core feature is a search engine which allows to retrieve (time-aligned) search results. The program does not impose any restriction on the organization of a corpus. Instead, it provides a mechanism to add metadata that can help the user keep their data organized. File names represent the most basic type of metadata and for small projects (containing a dozen of files or so) this may be all that is needed. When one needs to sort and organize a larger collection of files, Dolmen offers a flexible mechanism called *properties*, which represent key/value pairs. Each file can be tagged with an arbitrary number of such properties: the key represents a category, which is always a text string, and the value may be either a text string, a number or a logical value (true or false). For example, one could define a category entitled Subject, where each unique identifier represents a distinct value, Gender (with the values 'Male' and 'Female'), Task (e.g. 'Guided conversation', 'Free conversation', 'Word list', 'Text'), and so on. The properties that are created by the user are made available in the search window, which is displayed in Figure 3. The data used in this example are drawn from the IPFC project, and represent a subset of a corpus of Japanese learners of French coded for nasal vowels (see Detey, Racine, Kawaguchi & Eychenne, 2014 for an analysis of these data). The search window lets the user filter results by selecting individual files metadata. Text can be searched using plain text or string patterns known as *regular expressions* (see Eychenne & Paternostro 2016 for an overview).

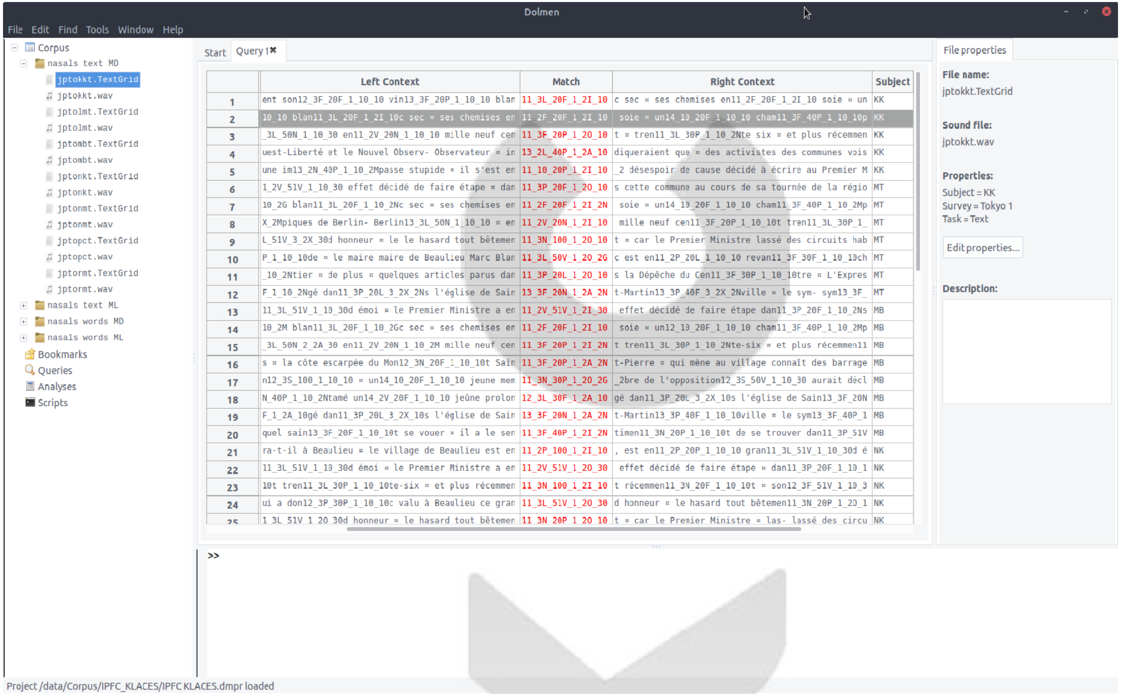
After validating the query, search results are displayed in the main interface, as shown in Figure 4. On the left-hand side, the *file browser* displays the structure of the current project, including all the files in the corpus. The *metadata panel,* located on the right-hand side, displays metadata about the currently selected file(s). The middle part of the interface is the *viewer*, which occupies the remainder of the window: it stores a number of 'views', which are similar to tabs in modern web browsers. Search results are displayed in a *query view*, as in this example. Each match is presented with its left and right context, as is usual in a concordancer, along with other metadata. Since annotations are time-aligned, it is possible to listen to individual matches, to directly modify the text of the match or to open the annotation and the sound file at the location where the match was found. Individual matches can also be saved, in which case they are stored in the "bookmarks" directory in the file browser. This feature can be particularly useful for instructors who want to save specific examples to illustrate a linguistic phenomenon. Results from a query can also be exported to a standard tabular format (CSV), which can then be imported into statistical software packages such as SPSS and R. This feature is mostly useful for researchers who would like to carry out a quantitative analysis of the data.

A very common use case in corpus phonology and sociophonetics is to devise coding schemes to analyze specific linguistic variables. As a matter of example, the IPFC project, as we mentioned earlier, has designed such a scheme for the analysis of nasal vowels. It is of course entirely possible to use Dolmen's search function to search for all the possible combinations allowed by the coding scheme. However, for all but the simplest schemes, looking for all possible combinations can be cumbersome and error-prone. For example, the IPFC coding for nasal vowels uses 6 fields which are represented using 12 digits; these fields encode large array of phenomena such as the quality and nasality of the vowel, the left and right segmental context, etc. In addition, users who will use or analyze the coded corpus are not necessarily those who have coded it, and they should not be required to know the details of the coding scheme.



<Figure 3> Dolmen's search window

To facilitate the use of coding schemes, Dolmen offers an extension mechanism known as "plugins". A plugin can contain information about one or several coding scheme(s), each of which is described in a simple text file using a widely used structured format (JSON). Such plugins can be created by users and can be easily shared and installed, which allows to customize Dolmen for the needs of a specific corpus or research project. Once the plugin is installed, it adds a customized search window for each coding scheme; users can then simply click on



	Left Context	Match	Right Context	Subject
1	ent son12_3F_20F_1_10_10 vin13_3F_20P_1_10_10 blan	11_3L_20F_1_2I_10	c sec ¤ ses chemises en11_2F_20F_1_2I_10 soie ¤ un	KK
2	10_10 blan11_3L_20F_1_2I_10c sec ¤ ses chemises en	11_2F_20F_1_2I_10	soie ¤ un14_10_20F_1_10_10 cham11_3F_40P_1_10_10p	KK
3	_3L_50N_1_10_30 en11_2V_20N_1_10_10 mille neuf cen	11_3F_20P_1_20_10	t ¤ tren11_3L_30P_1_10_2Nte six ¤ et plus récemmen	KK
4	uest-Liberté et le Nouvel Observ- Observateur ¤ in	13_2L_40P_1_2A_10	diqueraient que ¤ des activistes des communes vois	KK
5	une im13_2N_40P_1_10_2Mpasse stupide ¤ il s'est en	11_10_20P_1_2I_10	_2 désespoir de cause décidé à écrire au Premier M	KK
6	1_2V_51V_1_10_30 effet décidé de faire étape ¤ dan	11_3P_20F_1_20_10	s cette commune au cours de sa tournée de la régio	MT
7	10_2G blan11_3L_20F_1_10_2Nc sec ¤ ses chemises en	11_2F_20F_1_2I_2N	soie ¤ un14_10_20F_1_10_10 cham11_3F_40P_1_10_2Mp	MT
8	X_2Mpiques de Berlin- Berlin13_3L_50N_1_10_10 ¤ en	11_2V_20N_1_2I_10	mille neuf cen11_3F_20P_1_10_10t tren11_3L_30P_1_	MT
9	L_51V_3_2X_30d honneur ¤ le le hasard tout bêtemen	11_3N_100_1_20_10	t ¤ car le Premier Ministre lassé des circuits hab	MT
10	P_1_10_10de ¤ le maire maire de Beaulieu Marc Blan	11_3L_50V_1_20_2G	c est en11_2P_20L_1_10_10 revan11_3F_30F_1_10_10ch	MT
11	_10_2Ntier ¤ de plus ¤ quelques articles parus dan	11_3P_20L_1_20_10	s la Dépêche du Cen11_3F_30P_1_10_10tre ¤ L'Expres	MT
12	F_1_10_2Ngé dan11_3P_20L_3_2X_2Ns l'église de Sain	13_3F_20N_1_2A_2N	t-Martin13_3P_40F_3_2X_2Nville ¤ le sym- sym13_3F_	MT
13	11_3L_51V_1_10_30d émoi ¤ le Premier Ministre a en	11_2V_51V_1_2I_30	effet décidé de faire étape dan11_3P_20F_1_10_2Ns	MB
14	10_2M blan11_3L_20F_1_10_2Gc sec ¤ ses chemises en	11_2F_20F_1_2I_10	soie ¤ un12_10_20F_1_10_10 cham11_3F_40P_1_10_2Mp	MB
15	_3L_50N_2_2A_30 en11_2V_20N_1_10_2M mille neuf cen	11_3F_20P_1_2I_2N	t tren11_3L_30P_1_10_2Nte-six ¤ et plus récemmen11	MB
16	s ¤ la côte escarpée du Mon12_3N_20F_1_10_10t Sain	11_3F_20P_1_2A_2N	t-Pierre ¤ qui mène au village connaît des barrage	MB
17	n12_3S_100_1_10_10 ¤ un14_10_20F_1_10_10 jeune mem	11_3N_30P_1_20_2G	_2bre de l'opposition12_3S_50V_1_10_30 aurait décl	MB
18	N_40P_1_10_2Ntamé un14_2V_20F_1_10_10 jeûne prolon	12_3L_30F_1_2A_10	gé dan11_3P_20L_3_2X_10s l'église de Sain13_3F_20N	MB
19	F_1_2A_10gé dan11_3P_20L_3_2X_10s l'église de Sain	13_3F_20N_1_2A_2N	t-Martin13_3P_40F_1_10_10ville ¤ le sym13_3F_40P_1	MB
20	quel sain13_3F_20F_1_10_10t se vouer ¤ il a le sen	11_3F_40P_1_2I_2N	timen11_3N_20P_1_10_10t de se trouver dan11_3P_51V	MB
21	ra-t-il à Beaulieu ¤ le village de Beaulieu est en	11_2P_100_1_2I_10	, est en11_2P_20P_1_10_10 gran11_3L_51V_1_10_30d é	NK
22	11_3L_51V_1_10_30d émoi ¤ le Premier Ministre a en	11_2V_51V_1_20_30	effet décidé de faire étape ¤ dan11_3P_20F_1_10_1	NK
23	10t tren11_3L_30P_1_10_10te-six ¤ et plus récemmen	11_3N_100_1_2I_10	t récemmen11_3N_20F_1_10_10t ¤ son12_3F_51V_1_10_3	NK
24	ui a don12_3P_30P_1_10_10c valu à Beaulieu ce gran	11_3L_51V_1_20_30	d honneur ¤ le hasard tout bêtemen11_3N_20P_1_20_1	NK
25	1_3L_51V_1_20_30d honneur ¤ le hasard tout bêtemen	11_3N_20P_1_20_10	t ¤ car le Premier Ministre ¤ las- lassé des circu	NK

File name:
jptokkt.TextGrid

Sound file:
jptokkt.wav

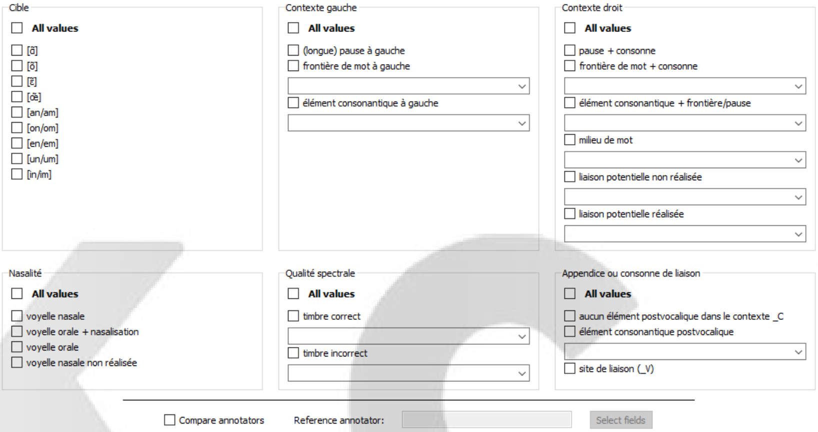
Properties:
Subject = KK
Survey = Tokyo 1
Task = Text

Description:

Project /data/Corpus/IPFC_KLACES/IPFC KLACES.dmpr loaded

<Figure 4> Search results in Dolmen

buttons to create their query, as in Figure 5, instead of using a string of alphanumeric characters (Figure 3). This also makes it particularly easy to extract all codings into a tabular file for further analysis, since leaving all buttons unchecked, as in Figure 5, will automatically extract all the matches that correspond to the coding scheme.



Cible
□ **All values**
□ [ɑ̃]
□ [ɔ̃]
□ [ɛ̃]
□ [œ̃]
□ [an/am]
□ [on/om]
□ [en/em]
□ [un/um]
□ [in/im]

Contexte gauche
□ **All values**
□ (longue) pause à gauche
□ frontière de mot à gauche
□ élément consonantique à gauche

Contexte droit
□ **All values**
□ pause + consonne
□ frontière de mot + consonne
□ élément consonantique + frontière/pause
□ milieu de mot
□ liaison potentielle non réalisée
□ liaison potentielle réalisée

Nasalité
□ **All values**
□ voyelle nasale
□ voyelle orale + nasalisation
□ voyelle orale
□ voyelle nasale non réalisée

Qualité spectrale
□ **All values**
□ timbre correct
□ timbre incorrect

Appendice ou consonne de liaison
□ **All values**
□ aucun élément postvocalique dans le contexte _C
□ élément consonantique postvocalique
□ site de liaison (_V)

□ Compare annotators
Reference annotator:

<Figure 5> Custom search window for the IPFC coding for nasal vowels

This brief presentation provides an overview of Dolmen's most important features. Interested readers are referred to Eychenne & Paternostro (2016) and to the program's manual for further information. Work is currently under way to further automate the annotation, analysis and visualization of speech corpus data.

5. Concluding remarks

In this contribution, I have tried to show how the use of spoken corpora can be beneficial in the field of foreign language education.

Corpora give us new tools to teach, learn and understand language, and they can be extremely valuable tools for research and teaching. Nevertheless, for them to be maximally useful, several factors must be kept in mind. First, corpora (and corpus-based resources) are not meant to replace more traditional ways of teaching. They can, and probably should, be incorporated with existing teaching methods, as an additional way to approach the target language. Second, for students to fully benefit from their use, corpora should not be provided in raw form but should be presented as usable resources and/or via appropriate tools. Failing that, it is clear that the introduction of corpora in the classroom may quickly be overwhelming for student, making the whole enterprise counterproductive. Finally, for a corpus to deliver its full potential, it is important that it be built following a sound methodology and that it be as inter-operable and open as possible. What has been achieved within PFC and IPFC has been possible because a large number of students and researchers have agreed to collaborate and share tools and resources, for the benefit of the community as a whole.

References

Avanzi, M., S. Schwab, P. Dubosson & J.-P. Goldman(2012) La prosodie de quelques variétés de français parlées en Suisse romande, in A.-C. Simon (ed). *La variation prosodique régionale en français*. Brussels: De Boeck/Duculot, 89-119.

Bird, S. & M. Liberman(2001) A formal framework for linguistic annotation. *Speech Communication,* 33(1-2), 23-60.

Boulton, A. & H. Tyne(2013) Corpus linguistics and data-driven learning: a critical overview, *Bulletin suisse de linguistique appliquée*, 97, 97-118.

Broeder D. & D. van Uytvanck(2014) Metadata formats. In Durand, Gut & Kristoffersen (eds)(2014), 150-165.

Chaudron, C.(2003) Data collection in SLA research. In C. J. Doughty & M. H. Long (eds)(2003). *The Handbook of Second Language Acquisition*. Malden: Blackwell, 762-828.

Christodoulides, G., M. Avanzi, M. & J.-P. Goldman(2014) DisMo: A morphosyntactic, disfluency and multi-word unit annotator. An evaluation on a corpus of french spontaneous and read speech. Proceedings of 9th International Conference on Language Resources and Evaluation, LREC2014, Island, 3902-3907.

Clopper, C. G. & D. B. Pisoni(2004) Effects of talker variability on perceptual learning of dialects. *Language and Speech*, 47(3), 207-238.

Colantoni, L., J. Steele & P. Escudero(2015) *Second Language Speech: Theory and Practice*. Cambridge University Press.

Detey, S.(2009) Phonetic input, phonological categories and orthographic representations: A psycholinguistic perspective on why language education needs oral corpora. The case of French-Japanese interphonology development. In Y. Kawaguchi, M. Minegishi & J. Durand (eds)(2009). *Corpus Analysis and Variation in Linguistics*, 179-200.

Detey, S., J. Durand, B. Laks & C. Lyche (eds)(2010) *Les variétés du français parlé dans l'espace francophone : ressources pour l'enseignement*. Paris: Ophrys.

Detey, S., J. Durand, B. Laks & C. Lyche (eds)(2016) *Varieties of Spoken French*, Oxford: Oxford University Press.

Detey, S., I. Racine, J. Eychenne & Y. Kawaguchi(2014) Corpus-based L2 phonological data and semi-automatic perceptual analysis: the case of nasal vowels produced by beginner Japanese learners of French, Proceedings of Interspeech 2014, 539-544.

Detey, S., I. Racine, Y. Kawaguchi & J. Eychenne (eds)(in press)(2016) *La prononciation du français dans le monde : du natif* à *l'apprenant*. Paris: CLE International.

Detey, S., I. Racine, Y. Kawaguchi & F. Zay(2016) Variation among non-native speakers: the InterPhonology of Contemporary French. In Detey, Durand, Laks & Lyche (eds)(2016), 489-500.

Durand, J, B. Laks & C. Lyche(2009) Le projet PFC: une source de données primaires structurées. In J. Durand, B. Laks & C. Lyche (eds)(2009) *Phonologie, variation et accents du français*. Paris: Hermès, 19-61.

Durand, J., U. Gut & G. Kristoffersen (eds)(2014) *The Oxford Handbook of*

Corpus Phonology. Oxford: Oxford University Press.

Eisenstein, M. R.(1989) Dialect variation and second-language intelligibility. In M. R. Eisenstein (ed). *The Dynamic Interlanguage*, New York & London: Plenum Press, 175-184.

Eychenne, J., S. Navarro, A. Tchobanov & J. W. van Leussen(2016) Approaching variation in PFC: the tools. In Detey, Durand, Laks & Lyche (eds)(2016), 387-398.

Eychenne, J. & R. Paternostro(2016) Analyzing transcribed speech with Dolmen. In Detey, Durand, Laks & Lyche (eds)(2016), D35-D52.

Gess, R., C. Lyche & T. Meisenburg (eds)(2012) *Phonological Variation in French: Illustrations from three continents*. Amsterdam: John Benjamins Publishing Company.

Granger, S.(2008) Learner corpora. In A. Lüdeling & M. Kytö (eds), *Corpus Linguistics: An international handbook*, Berlin & New York: Mouton de Gruyter, 259-274.

Granger, S., E. Dagneaux & F. Meunier(2002) The International Corpus of Learner English. Handbook and CD-ROM. Presses universitaires de Louvain: Louvain-la-Neuve.

Granger, S., G. Gilquin & F. Meunier (eds)(2015) *The Cambridge Handbook of Learner Corpus Research.* Cambridge: Cambridge University Press.

Gut, U.(2009) *Non-native Speech: A Corpus-based Analysis of Phonological and Phonetic Properties of L2 English and German.* Frankfurt am Main: Peter Lang.

Gut, U. & H. Voorman(2014) Corpus design. In Durand, Gut & Kristoffersen (eds)(2014), 13-26.

Ide, N. & K. Suderman(2014) The Linguistic Annotation Framework: a standard for annotation interchange and merging. *Language Resources and Evaluation*, 48(3), 395-418.

Iverson, P., P. K. Kuhl, R. Akahane-Yamada, E. Diesch, Y. Tohkura, A. Kettermann & C. Siebert(2003) A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, 87(1), B47-B57.

Romary, L & A. Witt(2014) Data formats for phonological corpora. In Durand, J., U. Gut & G. Kristoffersen (eds)(2014), 166-190.

Tyne, H.(2009) Corpus oraux par et pour l'apprenant. *Mélanges CRAPEL*, 31, 91-111.

이름: Julien Eychenne
소속: 한국외국어대학교 언어인지과학과
주소: 경기도 용인시 처인구 모현면 외대로 81
전자우편: jeychenne@hufs.ac.kr

투고일자: 2016년 7월 1일
심사일자: 2016년 7월 20일
게재확정일자: 2016년 8월 5일