KULA
Knowledge Creation, Dissemination, and Preservation Studies

Modes of Annotation in the Video-Based Corpus FrancoToile: Developing a Design Method

Catherine Caws et Stewart Arneil

Volume 1, numéro 1, 2017

URI : https://id.erudit.org/iderudit/1083154ar
DOI : https://doi.org/10.5334/kula.3

Aller au sommaire du numéro

Éditeur(s)

University of Victoria Libraries

ISSN

2398-4112 (numérique)

Découvrir la revue

Citer cet article

Caws, C. & Arneil, S. (2017). Modes of Annotation in the Video-Based Corpus FrancoToile: Developing a Design Method. *KULA*, *1*(1), 1–12. https://doi.org/10.5334/kula.3

Résumé de l'article

In corpus linguistics, texts are typically annotated in order to focus the attention on: (a) the form of the text or words, and (b) the structure of sentences (that is, morphological and syntactic tagging). Yet, when dealing with language learning and the development of skills other than just linguistic ones, other types of annotations are needed. Annotating with either a specific learner or pedagogy in mind often engages the researcher in more complex issues than the ones just related to corpus linguistics. In this article, we report on the methods used to create a digital library of videos and annotated transcripts called FrancoToile (http://francotoile.uvic.ca). As a needs-driven corpus, FrancoToile includes annotations within the video transcripts in order to help users develop their cultural and linguistic literacies in French. These annotations must relate directly to the purpose of the system (the development of cultural and linguistic literacies) and to the specific skill or competency that we hope language learners will gain. We analyze learning needs, modify the software, and observe and engage with users on an ongoing basis to create a language tool that will better address users' needs. This approach of incorporating user feedback increases the usefulness of the annotated videos. We continue to seek means to encourage the involvement of users, both teachers and learners, in the process of corpora editing and content building.

KULA KNOWLEDGE CREATION, DISSEMINATION, AND PRESERVATION STUDIES

Caws, Catherine, and Stewart Arneil. 2017. Modes of Annotation in the Video-Based Corpus FrancoToile: Developing a Design Method. *KULA: knowledge creation, dissemination, and preservation studies* 1(1): 1. DOI: https://doi.org/10.5334/kula.3

RESEARCH ARTICLE

Modes of Annotation in the Video-Based Corpus FrancoToile: Developing a Design Method

Catherine Caws[1] and Stewart Arneil[2]

[1] Department of French, University of Victoria, Victoria, BC, CA

[2] Humanities Computing and Media Centre, University of Victoria, Victoria, BC, CA

Corresponding author: Catherine Caws, Professor (ccaws@uvic.ca)

In corpus linguistics, texts are typically annotated in order to focus the attention on: (a) the form of the text or words, and (b) the structure of sentences (that is, morphological and syntactic tagging). Yet, when dealing with language learning and the development of skills other than just linguistic ones, other types of annotations are needed. Annotating with either a specific learner or pedagogy in mind often engages the researcher in more complex issues than the ones just related to corpus linguistics. In this article, we report on the methods used to create a digital library of videos and annotated transcripts called FrancoToile (http://francotoile.uvic.ca). As a needs-driven corpus, FrancoToile includes annotations within the video transcripts in order to help users develop their cultural and linguistic literacies in French. These annotations must relate directly to the purpose of the system (the development of cultural and linguistic literacies) and to the specific skill or competency that we hope language learners will gain. We analyze learning needs, modify the software, and observe and engage with users on an ongoing basis to create a language tool that will better address users' needs. This approach of incorporating user feedback increases the usefulness of the annotated videos. We continue to seek means to encourage the involvement of users, both teachers and learners, in the process of corpora editing and content building.

Keywords: Corpora; needs-driven corpus; annotation; digital documents; technology-enhanced language learning

Introduction

The use of corpora (understood herein as a collection of spoken, written or visual material collected and stored electronically for future retrieval or analysis) in language-learning research has become more and more popular in the last decade, due in part to the development of systems facilitating their integration into educational settings and their easy retrieval by learners. Web-based corpora have the potential to offer many advantages to learners because they provide good evidence of language patterns: they show language in use, often use authentic material such as interviews from native speakers, and can easily be adapted and updated to reflect current linguistic and cultural norms (e.g., Bernardini 2004; Boulton and Landure 2016; Braun 2005; Meunier et al 2011). In short, as noted by Bernardini, language corpora 'offer an ideal instrument to observe and acquire socially-established form/meaning pairings' (2004, 17), and, as such, they allow users to acquire a more nuanced understanding of a language. Additionally, corpus annotation, defined by McEnery and Wilson as the 'explicit identification of parts of speech, sentence structure, word meaning, and so on within a running text' provides, they note, 'a great amount of "added value" for the corpus user and consequently extends the potential of corpora in teaching' (1997, 10).

Corpora are rarely intended for second language learning. As Braun (2005) states, corpora have not often been used as main language resources in language programs despite their pedagogical potential; even today, the content of language corpora is seldom tailored to fit specific learning or instructional goals. Large corpora such as the Bank of English (http://www.titania.bham.ac.uk/docs/about.asp), a collection of over 450 million samples of modern English, or the ARTFL-FRANTEXT database (https://artfl-project.

uchicago.edu/content/artfl-frantext), a corpus of over 160 million French words from texts spanning the twelfth through twentieth centuries, offer good examples of genre or situational variations (e.g., Grieve-Smith 2007). But their data formats, mostly text-based (Braun 2005; McEnery and Wilson 1997), are driven by a need to address linguistic variations for research purposes and are not designed for exploitation by users with varied learning needs or styles, such as second-language learners.

More recently, however, researchers have introduced other types of language corpora, such as video-based corpora or learner corpora (Granger, Gilquin, and Meunier 2015; Meunier 2011; Meunier et al. 2011), which are slowly making their way into the educational landscape. Such projects include Elisa,[1] a corpus of language interviews developed at the University of Tübingen that focuses on spoken English and contains different national varieties. This corpus was created specifically as a Second-Language Learning Application; hence, it can be defined as a Pedagogically Driven Corpus (PDC). At the Centre for English Corpus Linguistics (CECL) and other institutes in Europe, many projects use corpora as a tool for second-language learning (see Meunier et al. 2011). In sum, while corpus linguistics may not have originally been focused on enhancing pedagogical practices, we are now seeing more exploitation of electronic corpora that include linguistic and/or pedagogical uses (e.g., Granger, Gilquin, and Meunier 2015; Meunier et al. 2011).

When considering annotation systems, PDCs differ from corpora created for linguistic purposes. In corpus linguistics (CL), typical annotations are based on both morphological and syntactic tagging in order to focus the attention on: (a) the form of words, and (b) the structure of sentences (e.g., Bowker and Pearson 2002; Sinclair 1991). Yet, to learn languages and develop skills other than purely linguistic ones, such as cultural awareness, users rely on other types of annotations. Annotating for these users, with either a specific learner or pedagogy in mind, often engages the researcher in more complex issues than the ones just related to CL (even though we could argue that any annotation system, such as the one presented herein, is of a linguistic nature, whether sociolinguistic or pragmatic). Unlike corpora created for linguistic purposes, which often use computer-generated automatic tagging systems to tag large sets of texts, PDCs include small sets of texts and rarely use automatic tagging. In that regard, there exist small multimodal annotated corpora that can serve as good examples of this manual annotation approach, where specific types of annotations are used (such as annotations of gesture, gaze or body movements) with a view to analyzing classroom or online interactions (e.g., Amory and Kissilev 2016).

Annotations in PDC are informed by research and findings on 'focus on form,' an increasingly popular field in second-language (L2) acquisition (SLA) (e.g., Ellis 2001; Nassaji 2000; O'Rourke 2005; Williams 2001), particularly for developing strategies to help learners acquire lexical items and reuse them adequately. Pedagogically speaking, many empirical studies have concluded that an experiential approach to teaching L2 should be enriched by form-focused instructional methods (Nassaji 2000). From a theoretical standpoint, some of the findings revealed by research on form-focused instruction (FFI) draw on information-processing theories and other aspects of cognitive psychology. Skills such as noticing—that is, the process of becoming cognitively aware of a linguistic item—while not guaranteeing acquisition, would help learners to process specific forms into their short-term memory (e.g., Schmidt 1994) and allow for conscious reflection on language form (e.g., Doughty and Williams 1998) while raising awareness of meta-linguistic focus (O'Rourke 2005). FrancoToile draws in part on 'focus-on-form' research, in that the system developed is meant to help users notice specific linguistic and cultural items within the speech of native speakers. After offering a technical overview of the system that we have developed, the present paper presents the methodological principles that we have followed in order to organize the content of the built-in corpus and annotations system.

FrancoToile: Overview of the System Developed and Its Built-in Corpus

FrancoToile (http://francotoile.uvic.ca) is a digital library of videos and transcripts of speakers of different varieties of French (France, Québec, Haiti, Senegal, Mauritius Island, etc.). Videos are collected by researchers using a protocol approved by the Human Research Ethics Board at the University of Victoria, whereby native speakers are interviewed and recorded while answering specific questions or reflecting upon diverse cultural aspects of the Francophone world. Both cultural diversity and unity are emphasized by the fact that speakers come from many regions of the Francophone world but often reflect in a similar manner upon key aspects of their culture. On a pedagogical level, one of the objectives of the project is to help language learners understand and respect the unique flavors of Francophonie while also recognizing that diversity is an inherent trait of many people speaking the same language but living in geographically

[1] This corpus was available online until a few years ago at http://www.corpora4learning.net/elisa/. It is currently unavailable.

and socially distinct regions. Diversity can be represented or analyzed by listening to different speakers; a learner could listen to and compare the following speakers, for example, in order to notice the differences in their accents: http://francotoile.uvic.ca/player.xql?id=ancf1 and http://francotoile.uvic.ca/player.xql?id=sngl2.

FrancoToile allows users to view videos, show or hide one or more subtitle tracks/transcriptions, and bookmark specific frames in the video. The infrastructure can be used with any collection of videos and transcriptions or commentaries, but it was created in order to allow language learners (or any other users of the site) to explore key concepts of Francophone culture through the testimonials of French-speaking individuals who currently live or who have lived in a Francophone environment. We have used an iterative approach based on user feedback in the development and design of the repository (see section on Methodological Principles).

Technical Overview: System Mechanics and User Interface

The web application for FrancoToile is written in PHP (i.e., hypertext preprocessor), which generates standard web pages (xhtml, css and javascript) by communicating with an eXist XML database <http://exist.sourceforge.net/>. The eXist database indexes the XML tree of each of the documents, allowing for rapid responses to users' queries even as over time the transcriptions are marked up based on a more complex schema. It also treats the entire corpus, as well as each transcript, as searchable entities. These functions become increasingly important as the video library grows, and research on the materials becomes more complex.

The corpus is currently composed of about 80 videos and accompanying annotated transcripts. Each video lasts between one to five minutes. The actual video data are stored as QuickTime files and as Ogg files in the file system on the server (not within the database). QuickTime has the advantage of having a robust Application Program Interface and is a widely supported platform. This means that multiple users can be interacting with the system at the same time without any risk of it crashing or slowing down. Ogg files are the fall-back for environments that do not support QuickTime. The structure of our code allows the site to support additional player platforms if needed.

For each transcript, the database includes an XML file containing the following components: metadata and information about the interviewee; a transcript of the video marked up into sentence-length utterances; a list of time-stamps, tags relating each utterance with a start time-stamp and end time-stamp; and the URL of the QuickTime video file. To ensure consistent structure and reliable data manipulation of the transcripts, the XML conforms to the Text Encoding Initiative's TEI P5 standard and is validated by a relax-ng[2] schema that we created for the project. When the user requests a video, the raw XML from the database is processed through XSLT (Extensible Stylesheet Language Transformations) before being passed back to the PHP for display on the page. The PHP dynamically creates JavaScript based on the metadata for the video, instantiates a movie-player object, and tells it where the video file is. The JavaScript runs in the browser to determine the current time of the playing movie to the millisecond. This information is fed into the movie-player object, which updates the appropriate sections of the page with the events that correspond to the retrieved time (e.g., to update the display of transcript text based on the current time in the video). Our approach supports an unlimited number of text feeds, each with their own collection of time-stamps, so you could have a transcript, a translation, a commentary on body language, and some other commentary thread all running simultaneously with independent time-stamps.

The web application's user interface is bilingual French/English and contains three main pages: the browse page, the search page, and the about page. The browse page on the site displays a world map (as illustrated in **Figure 1** below) and allows users to choose a video clip to play. Once users select a location on the map, they are connected to a player page (as illustrated in **Figure 2** below).

The player page on the site contains the media player and five tabbed panels. A subtitle tab displays the transcript of the video and is dynamically updated based on the current time of the video. The subtitle tab also allows users to create a bookmark. The bookmarks tab displays the user's list of bookmarks for the current video. A collection of these can be saved to the database and is assigned a unique URL, which can then be distributed to others. The search tab searches the transcript of the current video and lists the timestamp and associated utterance (that is, a spoken word or phrase) for each result, along with a link to that point in video. The full transcript tab offers a transcript of all the utterances presented together. Finally, the download tab allows user to download TEI P5 XML or XHTML of all utterances. Other videos related to

[2] Regular Language for XML Next Generation. See http://en.wikipedia.org/wiki/RELAX_NG.

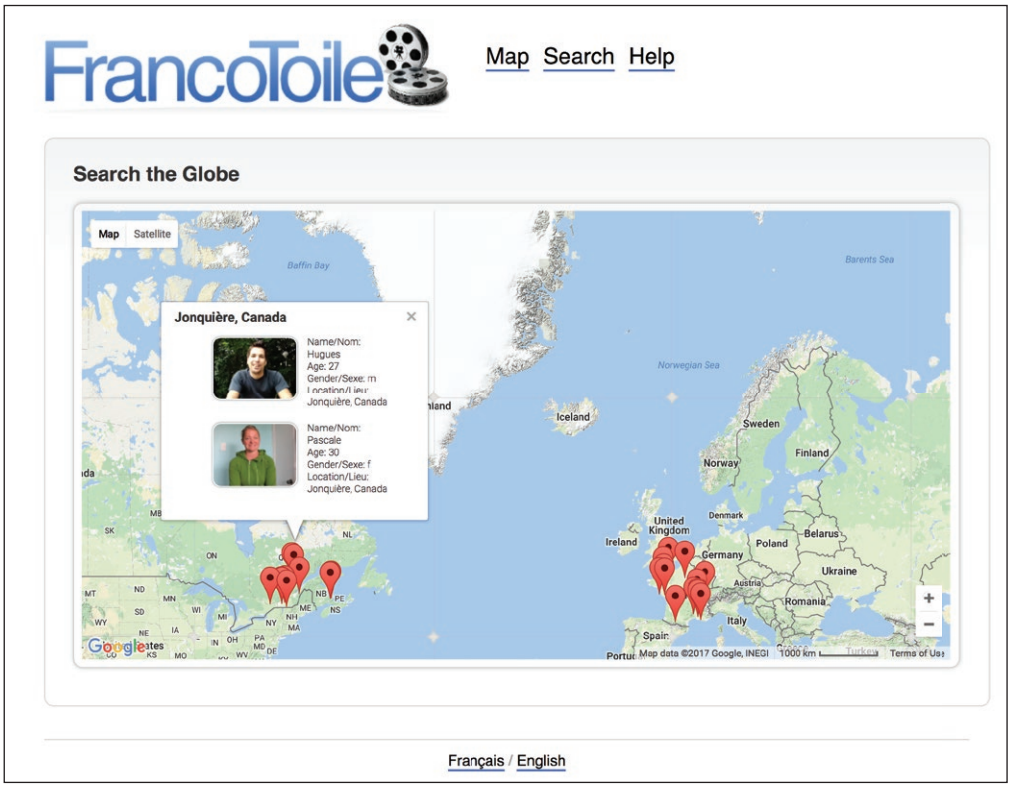

Figure 1: Browse page on FrancoToile website.

Figure 2: Player page on FrancoToile website.

the displayed videos are displayed on the right side of the page, a familiar page design for videos on sites such as YouTube.

While completely functional, the system is still under development, so annotations may have been entered in the XML files but have yet to be published on the site. Before annotations appear on the website, the team goes through a checking process and asks two other researchers or assistants to review the files and approve or add to them. In some cases, the annotation does not correspond to any of the categories that we have preselected. In such cases, either it is removed from the XML file, or we as a team decide to create a new category and include the annotation in the new category, then review other data files to see if any other comments should be moved or added to the new category.

Methodological Principles

Our project is set within an educational framework that uses a web-based tool to critically assess and understand cultural and language variations and similarities among speakers of French. Our process for developing the tool is based in part on the Analysis-Design-Development-Implementation-Evaluation (ADDIE) methodology (e.g., Strickland 2006; Colpaert 2004; Colpaert 2006). ADDIE is an instructional system design (ISD) particularly well suited to guiding developers in the creation and evaluation of language software or other language-related computer systems. As explained by Colpaert, one of the advantages of the ADDIE model is that 'each stage delivers output which serves as input for the next stage' (2006, 115). This process of recycling data is illustrated in **Figure 3** below. Once we applied this pedagogy and research-based approach to the creation of the system (as per Figure 3 below), we used the same model to design the annotation system.

FrancoToile is a needs-driven corpus that aims at helping users develop their cultural and linguistic literacies in French. The system includes annotations within video transcripts, which must relate both to the overall learning goal of the system (helping users develop their linguistic and cultural literacies) and to the specific skill or competency that we hope language learners will gain. In order to better assess learners' needs and goals, we have adopted the design-based research described above, using case studies to evaluate the system's needs (such as what types of annotations will help users develop the specific skills that they need to acquire). We analyze and recycle results of the evaluation into the development of the system. As explained in earlier papers (Caws 2009; Caws 2013; Caws and Hamel 2010), each implementation of the system into a learning environment functions as a case study with the goal of promoting students' development of cultural literacies. Part of the data collected during our interventions are of a qualitative nature and come from notes taken by users during their exploration of the system, transcripts of focus groups interviews, and

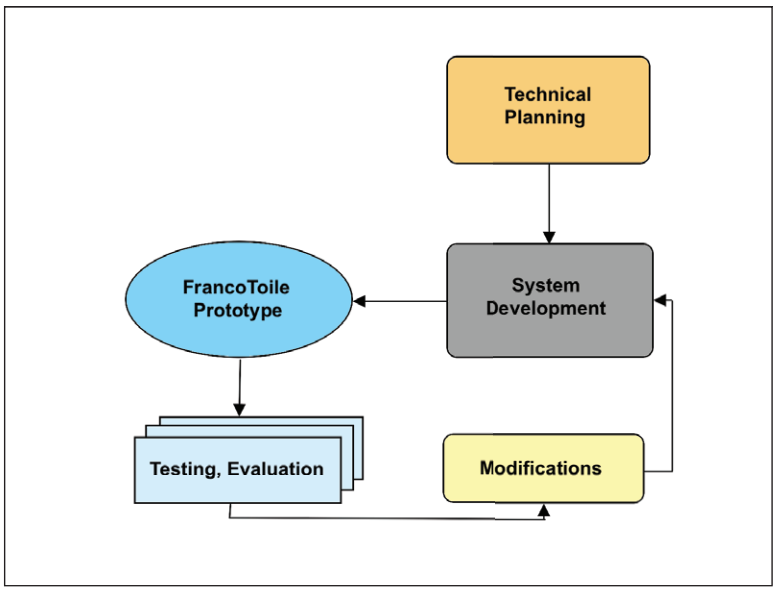

Figure 3: Authors' schema of the life cycle of tool development and implementation for FrancoToile.

comments made in online surveys. Adopting a discourse-analysis method, we extract trends and finds ways to recycle these data from users' feedback into concrete changes to the system or additions to its content. In the case of annotations, we pay particular attention to the type of information that users would like to see, would like to use, or would like to access. For instance, our first intervention (Caws 2009) revealed that users wished to have access to more information about linguistic features that differ from one Francophone region/country to another. As a consequence, as shown in more detail below, we have started to add more annotations relating to word connotations, pronunciations, and contrastive analysis of speech. To abide to our reviewing principles, the development team ensures that several of its members review annotations that have been recently added, in order to check for accuracy and authenticity. Once in the system, new annotations will become one focus (amongst others) of the next evaluation.

Categories of Annotations

With the goal of developing the cultural and linguistic awareness of FrancoToile users, we have designed categories of annotations to ensure consistency in the types of additional information that these notes provide. In addition, we hope that the categories produce an organized scheme and facilitate the manner in which users interact with the system. The categories that we have preselected are based on (1) our own observation of the kinds of questions learners tend to have or information they seek, and (2) on the feedback that we have collected from users through our various interventions while developing the tool. They include: word denotations and connotations; collocations and idiomatic expressions; links to visual documents; notes on pronunciation (contrastive or not); contrastive analysis of speech; and linguistic (lexical, morpho-syntactic or phonological) notes.

Word Denotations and Connotations

Vocabulary acquisition and expansion are often considered the learner's responsibility; yet, as noted by Godwin-Jones, 'An essential element of language learning is building one's personal store of words and expressions, a necessary component to improving competency in all areas of communication' (2010, 4). After years observing learners interact with digital documents, we noticed that online dictionaries (such as WordReference) or resources such as Wikipedia, where users tend to favour the English interface even if they can access a French article for a specific item, seemed to form the core resources used by learners. As a consequence, when we developed our system and transcribed videos, we tried to include a variety of ways to explain potentially challenging words while avoiding direct translations. Since we are trying to highlight cultural and lexical variations, mentioning connotations is particularly important. The development typically goes through a series of checks before finalizing the content of the annotations. One word can take many connotations, even denotations, depending on its geographic or social context of use (see **Figure 4** below).

To take into account polysemies related to specific contextual usage (such as register, location, or gender), the annotation is sometimes enriched by additional information in the form of lexical or semantic expansions of a word or concept. The annotations are created through discussion between researchers and speakers and in consultation with standard reference guides (dictionaries, glossaries, etc.). In some cases, we expand on the meaning of the word (semantic expansion) in order to warn/inform learners of specific usages of the word in certain contexts (to be avoided or, on the contrary, to be encouraged); in other cases, we propose other terms to express the same meaning and to inform learners of the variety of lexical usages.

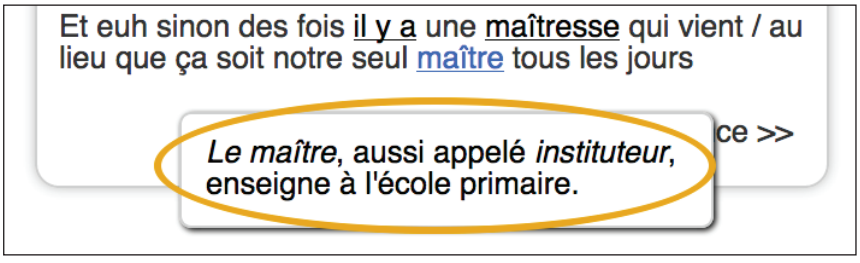

Figure 4: Annotation using simple denotation to explain a word.

Collocations and Idiomatic Expressions

A word takes its full meaning from the context in which it is used. Many isolated words, as described in a typical dictionary, will change meaning when used together with another word. In addition, words tend to often associate themselves to other lexical units and form a compound that is semantically and lexically fixed (see Polguère 2008). We note a clear correlation between form and meaning when studying the lexical items over time and place, as well as semantic variations that are clearly related to context of use (Sinclair 1996). In our corpus, we distinguish between collocation and idiomatic expression in the following manner: a collocation is a unit of two words that are semantically and lexically connected even if each word within the collocation has kept its original meaning. Some examples of collocations included in our corpus are *boîte de nuit* (night club), *se rendre compte* (to realize), and *vouloir dire* (to mean). According to the meaning-text theory (Polguère 2008), the relations between words that form a collocation are of diverse nature (lexical, semantic, syntactic, morpho-syntactic, etc.) and consequently will influence the entire meaning of the compound. Beyond collocations, we also find entire phrases that have gained specific meanings over the years, often due to cultural beliefs or customs that are inherent to a location or situation. These fixed phrases are defined as idiomatic expressions in our database and form a key aspect of our corpus because they often refer to essential cultural or historical information in regard to the French language.

Links to Visual Documents

The FrancoToile annotation system can associate a word with an image or link it to an external site (see **Figure 5** below) to provide an illustration of a place, person or thing mentioned by a speaker.

By assigning an image to an unknown word, we connect a linguistic term to a concrete image, a principle referred to as 'dual-coding' that can help learners better remember the new word (Godwin-Jones 2010). Images used in the system are typically images generated in house to avoid copyright issues or photos extracted from sites that are copyright-free for small images or offer high-quality images for a fee.[3]

Notes on Pronunciation

Accents and pronunciation constitute one of the primary focuses of our annotations. The original intention of the digital library was to expose learners of French to a variety of accents in order to dismiss the biased view that one accent is superior to another. It is often the case that learners believe that speakers from France are easier to understand than speakers from Québec; yet, little do they realize that speakers from the southeast of France have an accent that varies quite greatly from speakers from the northwest of France. Likewise, the accent of speakers from Montréal may be less noticeable compared to that of speakers

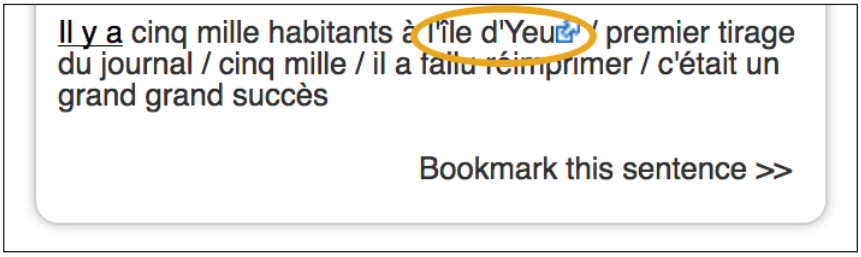

Figure 5: Annotation using hyperlink to include visual documents.

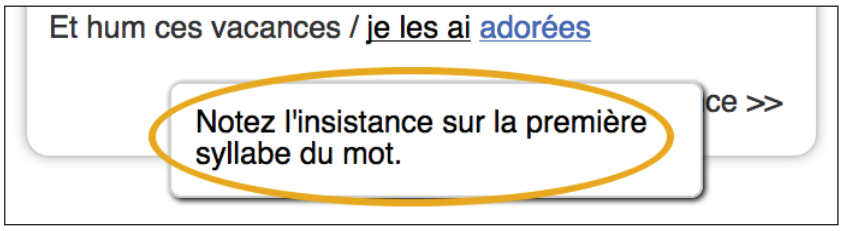

Figure 6: Annotation on pronunciation using descriptive non-technical language.

[3] See for instance <http://www.freedigitalphotos.net> or http://www.istockphoto.com.

from the Gaspésie region. In other words, the idea of one accent from France and one accent from Québec is a misconception that needed to be addressed. The annotations pertaining to pronunciation range from general description (avoiding linguistic jargon) to more specific explanation (using linguistic terminology). The differentiation is meant to address a variety of learning and teaching goals. **Figure 6** below gives an example of a non-linguistic note, while Figure 8, under 'Linguistic Annotation,' displays a note using phonetic terminology.

Contrastive Analysis of Speech

As part of our focus on raising learners' awareness of the effect of cultural variations on language, we include notes that pertain to lexical variations due to location. Such notes promote reflection on features such as socio-linguistic variations, languages in contact and their effect on linguistic creativity, and tolerance towards language variations. During our interviews with users of the system, this aspect of contrastive analysis was one of the most highly rated pieces of information. Learners expressed a vivid interest in understanding and capturing language variations (see **Figure 7** below) in order to ease their frustrations when communicating with French-speaking individuals from areas to which they were not accustomed.

Linguistic Notes

There are few tagged corpora that focus on variations in French and are accessible online as open sources to be used freely by learners and teachers alike. We submit annotations to a rigorous review process, therefore, to ensure their quality. Following our methodology principles, the corpus is (1) analyzed by a selection of consultants who are either native speakers from various French-speaking countries or former learners of French as an additional language. Any item that seems odd to one or more consultants will be further

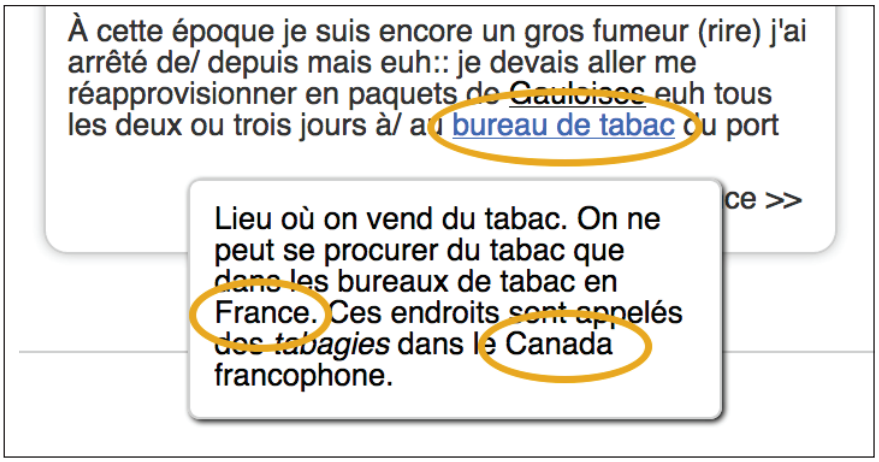

Figure 7: Annotation with contrastive analysis of a word.

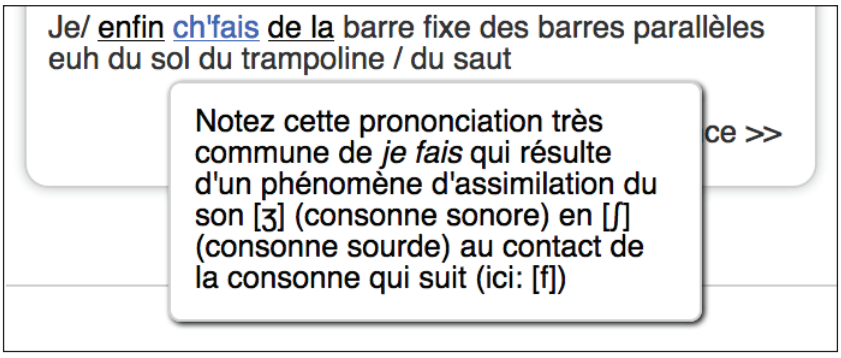

Figure 8: Example of linguistic annotation.

discussed and eventually annotated in the text. The annotation, which include notes on lexical, morpho-syntactic or phonological features, will be as precise as possible to allow either the learner or the instructor to expand on the note. Below is an example of a note regarding pronunciation:

Assessment of the Annotation System

While creating some challenges, adding annotations to our corpus offers several learning opportunities. Topic interest and prior knowledge are two factors that play a vital role in the use of annotations. According to a study by Erçetin (2010), students' prior knowledge of and interest in a topic seem to increase the likelihood that they will read and seek additional knowledge through the annotations. Being Internet-savvy, today's language learners are used to expanding their knowledge base by using hypermedia documents from the Internet. Annotations seem to be a natural process for most students. However, considering the variety of hypermedia available to students, it is wise to offer a diversity of modes of annotation, as shown in the previous section. This diversity will also synchronize with an increasingly diverse learner population, from the visual learner to the auditory one.

When using corpora for language learning and teaching, users can encounter serious problems due to the amount of data contained within a corpus and its data format (e.g., Braun 2005; Meunier 2002), but this is not the case for FrancoToile. Our corpus of video and texts started with 10 videos and annotated transcripts and now contains over 80 videos and transcripts (some still being annotated at time of writing). The small size of our corpus, which contains just over 100,000 tokens (i.e. words), should not present any problem to users; indeed, they can easily explore every facet of the database, and explore every occurrence of a word should they wish.

Another factor that has influenced our annotation methods is the fact that the corpus is pedagogically driven. We have adopted a discourse-based approach by annotating parts of the speech that have cultural, historical or social relevance. This is one the advantages presented by small corpora (e.g., Braun 2005; Henry & Roseberry 2001). Based on our experience as instructors and on the experience of L2 learners, we produce annotations only after having experts and language informants read and study *whole* transcripts, following Braun's point that 'In the exploitation of a pedagogically relevant corpus, learners and teachers will best be able to authenticate the corpus materials when they start by studying a general corpus description and some of the texts in their entirety' (2005, 8). In sum, being familiar with the texts at hand facilitates the annotation process.

In addition, being pedagogically driven, the corpus was created precisely to be exploited for second-language teaching and learning purposes. As explained earlier, the development of our system—based in part on the ADDIE model—includes a series of evaluations based on several implementations of the tool within the specific educational setting of language courses at the University of Victoria. These case studies include formative and summative evaluations in the form of online surveys, screen captures of interactions, note taking and focus group interviews (Caws 2009; Caws 2013). Data collected are analyzed using statistical measures for quantitative data and discourse analysis for qualitative data. Results of these analyses are recycled in the development of the tool and/or in the implementation of learning tasks. Testing follows the next implementation and so on and so forth.

Conclusions

Adopting the ADDIE method for the development of our system has presented both challenges (by making the design very user-dependent) and opportunities (by involving users in the system's development and engaging them in an experiential learning process). Indeed, the system that we are currently developing (and using in French language courses for specific learning goals) has allowed us to reflect more deeply upon the implications that creating a PDC may have for learners and instructors. Clearly, the evaluation part of our development cycle is one of the key factors in properly addressing teaching and learning needs and transforming the system into a tool that will be easy to use, effective, and efficient. While the system holds much promise for future development, we agree with Colpaert, who states:

> Software developers should only make the best possible technological choices for guaranteeing overall software quality. It is up to language teachers and language pedagogues, or others involved in the decision to develop software for language learning purposes, to bridge the gap and to learn how to specify accurately enough what they need. (2004, 15)

Time spent on analyzing learning needs and systems requirements is the essence of creating a language tool that will better address users' needs. Likewise, developing a method of annotation that derives from the same premise has been an essential component of the FrancoToile project. Yet, we still need to find means to facilitate annotation practices and techniques so that teachers and learners alike are even more involved in the process of corpora editing and content building. Pedagogically speaking, by engaging learners in the development process of the system, we take a 'post-critical' stand (e.g., Selber 2004)—that is, we include digital literacies within the discipline of language learning and contribute to the development of critical and functional skills.

In 1997, McEnery and Wilson, commenting on teaching and language corpora, advocated for a need to 'proceed in a clear and controlled fashion, so that we can establish pedagogical, research and resource goals' (12). Clearly, many initiatives and groups focused on the integration of corpora in language learning have emerged since then. However, while the creation of databases such as FrancoToile and Elisa are changing the landscape of language corpora and enhancing the use of PDC, more collaborative initiatives need to be launched in order to reduce duplication of efforts and create a true integrated community of research and development. Working in concert with teachers and learners is one venue that has not been fully explored; yet, it is promising, pedagogically speaking, as it will encourage a research-based teaching of languages and will engage learners in the co-construction of knowledge.

Acknowledgements

This research has been supported by the Social Sciences and Humanities Research Council of Canada.

Competing Interests

The authors have no competing interests to declare.

References

Amory, Michael, and Olesya Kissilev. 2016. "The Annotation of Gesture Designed for Classroom Interaction." In: *Proceedings of the International Conference on Language Resources and Evaluation: Multimodal Corpora: Computer Vision and Language Processing*, Portorož, 9–12. 23–28 May 2016. http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-MCC-2016-proceedings.pdf.

Bernardini, Silvia. 2004. "Corpora in the Classroom: An Overview and Some Reflections on Future Developments." In: *How to Use Corpora in Language Teaching*, edited by John McH. Sinclair, 15–36. Amsterdam: John Benjamins.

Boulton, Alex, and Corinne Landure. 2016. "Using Corpora in Language Teaching, Learning and Use." *Recherche et pratiques pédagogiques en langues de spécialité* 35(2). DOI: https://doi.org/10.4000/apliut.5433

Bowker, Lynne, and Jennifer Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. London and New York, NY: Routledge.

Braun, Sabine. 2005. "From Pedagogically Relevant Corpora to Authentic Language Learning Contents." *ReCALL* 17(1): 47–64. DOI: https://doi.org/10.1017/S0958344005000510

Caws, Catherine. 2009. "Contexte et culture en enseignement du FLS: De la création d'un corpus à son exploitation linguistique." *Mélanges CRAPEL* 31: 205–222.

Caws, Catherine. 2013. "Evaluating a Web-Based Video Corpus through an Analysis of User Interactions." *ReCALL*, 25(1): 84–104. DOI: https://doi.org/10.1017/S0958344012000262

Caws, Catherine, and Marie-Josée Hamel. 2010. "Usability Tests in CALL Development: Pilot Studies in the Context of the Dire autrement and FrancoToile." *Calico Journal* 27(3): 491–504. http://www.jstor.org/stable/calicojournal.27.3.491.

Colpaert, Jozef. 2004. Design of Online Interactive Language Courseware: Conceptualization, Specification and Prototyping. Research into the Impact of Linguistic-Didactic Functionality on Software Architecture. PhD diss., University of Antwerp.

Colpaert, Jozef. 2006. "Toward an Ontological Approach in Goal-Oriented Language Courseware Design and Its Implications for Technology-Independent Content Structuring." *Computer Assisted Language Learning* 19(2–3): 109–127. DOI: https://doi.org/10.1080/09588220600821461

Doughty, Catherine, and Jessica Williams. 1998. "Pedagogical Choices in Focus on Form." In: *Focus on Form in Classroom Second Language Acquisition*, edited by Catherine Doughty and Jessica Williams, 197–161. Cambridge: Cambridge University Press.

Ellis, Rod. 2001. "Introduction: Investigating Form-Focused Instruction." *Language Learning* 51: 1–46. DOI: https://doi.org/10.1111/j.1467-1770.2001.tb00013.x

Erçetin, Gülcan. 2010. "Effects of Topic Interest and Prior Knowledge on Text Recall and Annotation Use in Reading a Hypermedia Text in the L2." *ReCALL* 22(2): 228–246. DOI: https://doi.org/10.1017/S0958344010000091

Godwin-Jones, Robert. 2010. "Emerging Technologies: From Memory Palaces to Spacing Algorithms: Approaches to Second-Language Vocabulary Learning." *Language Learning & Technology* 14(2): 4–11. http://llt.msu.edu/vol14num2/emerging.pdf.

Granger, Sylviane, Gaëtanelle Gilquin, and Fanny Meunier, (eds.). 2015. *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139649414.001

Grieve-Smith, Angus B. 2007. "The Envelope of Variation in Multidimensional Register and Genre Analysis." In: *Corpus Linguistics Beyond the Word: Corpus Research from Phrase to Discourse*, edited by Eileen Fitzpatrick, 21–42. Amsterdam and New York, NY: Rodopi.

Henry, Alex, and Robert L. Roseberry. 2001. "Using a Small Corpus to Obtain Data for Teaching a Genre." In: *Small Corpus Studies and ELT: Theory and Practice*, edited by Mohsen Ghadessy, Alex Henry, and Robert L. Roseberry, 93–133. Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/scl.5

McEnery, Tony, and Andrew Wilson. 1997. "Teaching and Language Corpora (TALC)." *ReCALL* 9(1): 5–14. DOI: https://doi.org/10.1017/S0958344000004572

Meunier, Fanny. 2002. "The Pedagogical Value of Native and Learner Corpora in EFL Grammar Teaching." In: *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, edited by Sylviane Granger, Joseph Hung, and Stephanie Petch-Tyson, 119–141. Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/lllt.6

Meunier, Fanny. 2011. "Corpus Linguistics and Second/Foreign Language Learning: Exploring Multiple Paths." *Revista Brasileira de Linguística Aplicada* 11(2): 459–477. DOI: https://doi.org/10.1590/S1984-63982011000200008

Meunier, Fanny, Sylvie De Cock, Gaëtanelle Gilquin, and Magali Paquot. 2011. "Putting Corpora to Good Uses: A Guided Tour." *A Taste for Corpora: In Honour of Sylviane Granger*, edited by Fanny Meunier, Sylvie De Cock, Gaëtanelle Gilquin, and Magali Paquot, 1–6. Amsterdam: John Benjamins.

Nassaji, Hossein. 2000. "Towards Integrating Form-Focused Instruction and Communicative Interaction in the Second Language Classroom: Some Pedagogical Possibilities." *The Modern Language Journal* 84(2): 241–250. DOI: https://doi.org/10.1111/0026-7902.00065

O'Rourke, Breffni. 2005. "Form-focused Interaction in Online Tandem Learning." *CALICO Journal* 22(3): 433–466. http://www.jstor.org/stable/24147933.

Polguère, Alain. 2003. *Lexicologie et sémantique lexicale: Notions fondamentales*. Montréal, QC: Les Presses de l'Université de Montréal.

Schmidt, Richard. 1994. "Deconstructing Consciousness in Search of Useful Definitions for Applied Linguistics." *AILA Review* 11: 11–26.

Selber, Stuart A. 2004. *Multiliteracies for a Digital Age*. Carbondale, IL: Sourthern Illinois University Press.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, John. 1996. "The Search for Units of Meaning." *Textus* 9(1): 75–106.

Strickland, A. W. 2006. "ADDIE. Idaho State University College of Education, Science, Math & Technology Education." Accessed 17 July 2017. Archived on: https://web.archive.org/web/20060709154016, http://ed.isu.edu/addie/index.html.

Williams, Jessica. 2001. "Learner-Generated Attention to Form." *Language Learning* 51(Issue Supplement): 303–346. DOI: https://doi.org/10.1111/j.1467-1770.2001.tb00020.x

How to cite this article: Caws, Catherine, and Stewart Arneil. 2017. Modes of Annotation in the Video-Based Corpus FrancoToile: Developing a Design Method. *KULA: knowledge creation, dissemination, and preservation studies* 1(1): 1. DOI: https://doi.org/10.5334/kula.3

Submitted: 28 June 2016 **Accepted:** 06 August 2017 **Published:** 30 November 2017

]u[*KULA: knowledge creation, dissemination, and preservation studies* is a peer-reviewed open access journal published by Ubiquity Press

OPEN ACCESS