*ReCALL* **21**(1): 55–75. 2009 © European Association for Computer Assisted Language Learning
doi:10.1017/S0958344009000093 Printed in the United Kingdom


# *Developing annotation solutions for online Data Driven Learning**

PASCUAL PÉREZ-PAREDES AND JOSE M. ALCARAZ-CALERO
*Universidad de Murcia, Departamento de Filología Inglesa, Facultad de Letras, Campus La Merced, 30071 – Murcia, Spain*
*(email: pascualf@um.es; jmalcaraz@um.es)*

**Abstract**

Although *annotation* is a widely-researched topic in Corpus Linguistics (CL), its potential role in Data Driven Learning (DDL) has not been addressed in depth by Foreign Language Teaching (FLT) practitioners. Furthermore, most of the research in the use of DDL methods pays little attention to annotation in the design and implementation of corpus-based/driven language teaching.

In this paper, we set out to examine the process of development of *SACODEYL Annotator*, an application that seeks to assist SACODEYL system users in annotating XML multilingual corpora. First, we discuss the role of annotation in DDL and the dominating paradigm in general corpus applications. In the context of the language classroom, we argue that it is essential that corpora should be pedagogically motivated (Braun, 2005 and 2007a). Then, we move on to deal with the analysis and design stages of our annotation solution by illustrating its main features. Some of these include a user friendly hierarchical and extensible taxonomy tree to facilitate the learner-oriented annotation of the corpora; real-time graphics representation of the annotated corpus matching the XML TEI-compliant (Text Encoding Initiative) standard, as well as an intuitive management of the different data sections and associated metadata.

SACODEYL (System Aided Compilation and Open Distribution of European Youth Language) is an EU funded MINERVA project which aims to develop an ICT-based system for the assisted compilation and open distribution of multimedia European teen talk in the context of language education. This research lays emphasis on the functionalities of the application within the SACODEYL context. However, our paper addresses similarly the needs of potential multimedia language corpus administrators in general on the lookout for powerful annotation assisting software. SACODEYL Annotator is free to use and can be downloaded from our website.

Keywords: Corpora, Data Driven Learning, Annotation, Software Design, Language Teaching, Teacher Support

## 1 Introduction

Despite its relevance, the role of annotation is very rarely discussed in Data-Driven Learning (DDL) proposals. Some of them exploit raw, unannotated corpora and use

* *System Aided Compilation and Open Distribution of European Youth Language* research funded by the European Commission under the Socrates-Minerva initiative (225836-CP-1-2005-1-ES-MINERVA).

existing applications to manage the language data, either with stand-alone solutions like WordSmith or Monoconc or with online resources such as Spaceless.[1] Other proposals use annotated corpora for language classroom activities. However, the very notion of annotation has remained bound to the linguistic research paradigm, which has led teachers to identify corpus applications in DDL with the exploitation of grammatical features and text genres. The usefulness of this approach is undisputable, but we were concerned by the fact that the efforts to implement DDL were mediated by a dominating morphosyntactic approach to language data. This becomes even more striking when one reviews mainstream publications in the field of foreign language teaching (FLT) and comes to realize that the morphology or the syntax of a language is driving pedagogy in none of them. So the issue is why *Data* in DDL should be restricted to morphological tagging. The challenge would be to try and turn *Data* into a pedagogical construct, just like any other CALL or published material.

In this paper we present the analytical framework that guided our research in providing annotation solutions for the implementation of the SACODEYL system, an effort to deliver the actual voices of European young people online. In particular, it is devoted to the design of the annotation application that has been developed for SACODEYL: SACODEYL Annotator.[2] The role of pedagogic annotation is central here as the resulting annotation will condition and shape the learning experiences of learners using our system. For reasons of space, areas of the system other than the annotation solution cannot be fully discussed here.

## 2 Annotation in corpus linguistics

Annotation in the context of linguistics can be seen as both the process and the resulting product of adding information to electronic texts. In the field of corpus linguistics (CL), annotation is primarily conceived of as an add-on, a way to describe the information that matches the needs of language researchers, linguists or corpus users. In this way, Leech (1991) equates annotation with *analysis*, while McEnery and Wilson (1996) describe corpus annotation in terms of *processing*. Within this scheme, the whole point of annotating corpora lies in the fact that annotation allows corpus users both refined information retrieval capabilities and subsequent treatment of the data. Once captured, this information may serve a very wide array of purposes, from language description to natural language processing.

There exist different approaches to the conceptualization of the annotation process. Annotation can be (a) automatic, semi-automatic or manual, depending on the degree of human intervention in the process, or (b) it can be done by one single annotator or a group of annotators. In either case, its *raison d'être* inevitably (c) reflects the different nature of the ultimate aim of the meta-information being added to the corpus. As an illustration, annotators' needs may range from an interest

[1] Available at http://www.spaceless.com/concordancer.php; Sabine Braun offers an excellent selection of tools and resources at http://www.corpora4learning.net/

[2] SACODEYL Annotator is free to use and can be downloaded from http://www.um.es/sacodeyl

in non polysemic ambiguity (Poesio & Artstein, 2005) to an interest in L2 speakers' errors (Abe & Tono, 2005). Within the CL research tradition, the efforts of annotation specialists have been geared towards grammatical *tagging*,[3] including morphological and syntactical information, (Garside, Leech & McEnery, 1997). Biber, Conrad and Reppen (1998) is, for example, a valuable introduction to corpus annotation issues, whereas Leech (1993) offers a more comprehensive treatment through a description of annotation procedures.

McEnery and Wilson (1996: 57) have pointed out that annotation poses important challenges to researchers, who predictably will need to find the right balance between "consensus-based-theory-neutrality" and problem solving issues in connection with the ultimate purpose of their annotating. Therefore, it follows that annotation in CL is well rooted, although not restricted, to traditional linguistic research paradigms. Teachers and language researchers have promoted the use of L1 representative corpora in the FL classroom. Examples of this interface are studies such as Santos Pereira (2004: 110), which reports on the use of a Portuguese language lexicon with "grammatical and quantitative information" as a "new approach to understanding […] meaning and semantic disambiguation, real differences between near-synonyms" and other features of language in actual usage. The tide of CL has reached the classroom shores, and now it has become more and more common to find otherwise specialised jargon such as 'concordance' and 'concordancer' (Flowerdew, 1993; Owen, 1996; Gavioli & Aston, 2001; Weber, 2001; Frankenberg-Garcia, 2005) or DDL (Mishan, 2004) not only in linguistics or CL, but also in FLT journals. Everything has gone the way it was predicted by Biber and Finnegan (1991), who expected that increased computer literacy among linguists and the overwhelming presence of computers in our daily work could contribute to the expansion of the field and its applications.

Although the situation is far from being one where CL has become mainstream in FLT, it is nonetheless true that major efforts have been made to voice the enormous potential of CL in the foreign language classroom (McCarthy & O'Dell (2006) or O'Keeffe, McCarthy & Carter (2007)). Let us consider some of these attempts.

## 3 Annotating corpora for the FL classroom

### *3.1 Corpora in the FL classroom*

In recent years various collected volumes have been devoted to the use of corpora in the language classroom (Sinclair, 2004; Braun, Kohn & Mukherjee, 2006; Hidalgo, Quereda & Santana, 2007). Although their scope is wide, and very often interdisciplinary, we can find in all of them some common ground as to the need for the FL community to develop specific corpora that suit the challenges of the language classroom. Notwithstanding, this is not a new original claim. Chambers (2007) takes

[3] Probably the most popular tagsets are the LOB, the ICE and the SKELETON. The AUTASYS, for instance, (Automatic Text Annotation System) tagger can use any of these. Makov's process model based on finite-state grammar and the use of frequency-driven probabilities has been particularly successful, according to Leech (1991).

up Leech and Candlin's (1986) pioneering work on CALL to remind us that corpora and CL methods that are not adapted to the specific classroom environment may be unsuccessful. In the same volume, Leech (1986) gave a longer-term perspective on the integration of corpora in CALL: ''The CALL and CBELT programs will be able to make use of corpora of graded and classified English language texts, and language databases in the form of a grammar and a lexicon''. Twenty-one years after this statement, and twenty-three after the Lancaster Symposium 'Computers in English Language Education and Research' where these ideas were first advanced, the CL and CALL community are still in the process of normalizing the use of corpora in the language classroom (Bax, 2003).

Chambers (2007: 12) has detected ''major obstacles'' for the popularisation of corpora, including attitudinal factors as well as gaps in teacher training and familiarization in secondary education with CL methodology. Braun (2007b) has explored corpus integration in the secondary school curriculum and has outlined the methodological challenges she encountered during her research. The author concludes that for corpora to be integrated in mainstream formal education it is necessary to move from DDL to needs-driven corpora.

It appears that, despite the benefits (Gavioli & Aston, 2001; Bernardini, 2004), CL methods have a long way to go before establishing themselves in the mainstream FL classroom. On this topic, Mauranen (2004: 99) points out that for a teaching method to become an important innovation, it has to ''make its way to the normal classroom where teachers and students can use it as part of their everyday routine, with not too much extra hassle''. Certainly, the reasons that underlie the *invisibility* of corpora in mainstream formal education lie outside the scope of this research. For the time being, the natural corpus playground continues to be tertiary education. However, some pioneering work is spearheading significant innovation.

### *3.2 Annotating with learning in view*

Within the context of advanced use of the English language, Braun (2007a: 43) has pointed out that a pedagogically motivated corpus may be instrumental in assisting learners in their discovery of ''interesting items'' in their materials and their further exploration ''with corpus-based methods of investigation''. Braun (2006a: 29 ff.) builds on previous *DIY corpus literature* and suggests that *ad hoc* spoken corpora in FLT may (a) provide a more systematic range of material than individual texts or scattered collections of activities and, if well-designed, (b) offer a wider range of idiolects than the average material. She states that thematic annotation, including topic keys and section titles, are particularly useful in the implementation of pedagogically motivated corpora. Thus, by adopting a classroom perspective, the morphological or syntactic features of the language are not necessarily driving the annotation of the corpus and, consequently, are not conditioning the type of learning experiences that a learner could engage in when working on these corpora.

As already stated, the annotation in a corpus describes the information that matches the needs of researchers, linguists or corpus users. It goes without saying that annotation driven by language description standards may have a very positive impact on the language classroom. However, it seems ironic to think that this

type of language research-oriented annotation, which heavily relies on traditional grammar paradigms, should be instrumental in a FLT communication-oriented methodological context in which the grammar-translation method or the structural method are no longer dominating the arena. Campbell *et al.* (2007) are a good example of how *ad hoc*, FLT-oriented corpora can be built around interests other than morphosyntactic units.

Braun (2005, 2006a) has discussed the creation of one topic-driven corpus: the English Language Interview Corpus as a Second-Language Application (ELISA), "a collection of video-based interviews with native speakers of different varieties of English, for example, US, England, Scotland, Ireland, Australia, and from different walks of life".[4] Taking one of these interviews[5] at random, we can find that the root element in the markup 〈session file = "horse_caravanning_ie.xml"〉 is a whole text. By enlarging the element, we obtain the annotated information, like title, metadata (creator, video file, language variety, list of speakers, description, speech rate information), topic and speaker name. The following is an excerpt from the code:

```
- <event start="0m0" end="1m24" video="horse_caravanning_ie" duration="1m24" wordcount="223">
  - <topic>
      <topic_title>What we do</topic_title>
      <topic_key>02 What we do</topic_key>
      <content_key />
    </topic>
  - <speaker name="Dieter">
      In the 60s, in the late 60s, I had worked in Germany for a while and I decided that I wanted to
      have my children reared in Ireland. So we came back from Germany, working for the Irish
      Tourist Board and started this enterprise
      <break />
      . It's lovely now with the sunshine, we don't always have it like this, but very often. We started
      with 12 and then 20 caravans, and now we have about 35. And it's been a basis of what which
      we can live as a family, raise our children in a nice environment. We work very hard for three
      months and then have a very relaxed time of it, nine months. And in that time then I took on as
      a hobby computers, and Mary took on tour-guiding. So we have various different aspects to
      what we do. The horse caravans is a very intensive work just for those three months, but it's
      very enjoyable because we mix in the family a quiet nine months where we are very much en
      famille with the children, you can concentrate on them much more than if we were nine-to-five
      workers. And then the intensity of the three months means that we can also have our children
      employed, and learning how to work, learning how to deal with people. So, good mixture, isn't
      it.
      <cut />
    </speaker>
  </event>
```

Fig. 1. XML code from ELISA

Finding XML in the code has ceased to be a surprise in CL or in CALL.[6] XML has been used successfully in different CALL applications for a variety of purposes and reasons. Cushion (2004) sees in XML a powerful technology to exchange language data and language structure and thus prolong the extension of data life. Ward (2002) believes that this technology facilitates the re-usability of resources and is a time-saving option for CALL projects where developers have to adhere to highly

[4] http://www.uni-tuebingen.de/elisa/html/elisa_info.html

[5] http://www.uni-tuebingen.de/elisa/html/horse_caravanning_ie.html

[6] It is not our intention here to discuss the benefits of Standard markup languages like SGML or the origins of the Text Encoding Initiative. For a thorough and authoritative review of these issues please visit http://www.tei-c.org/

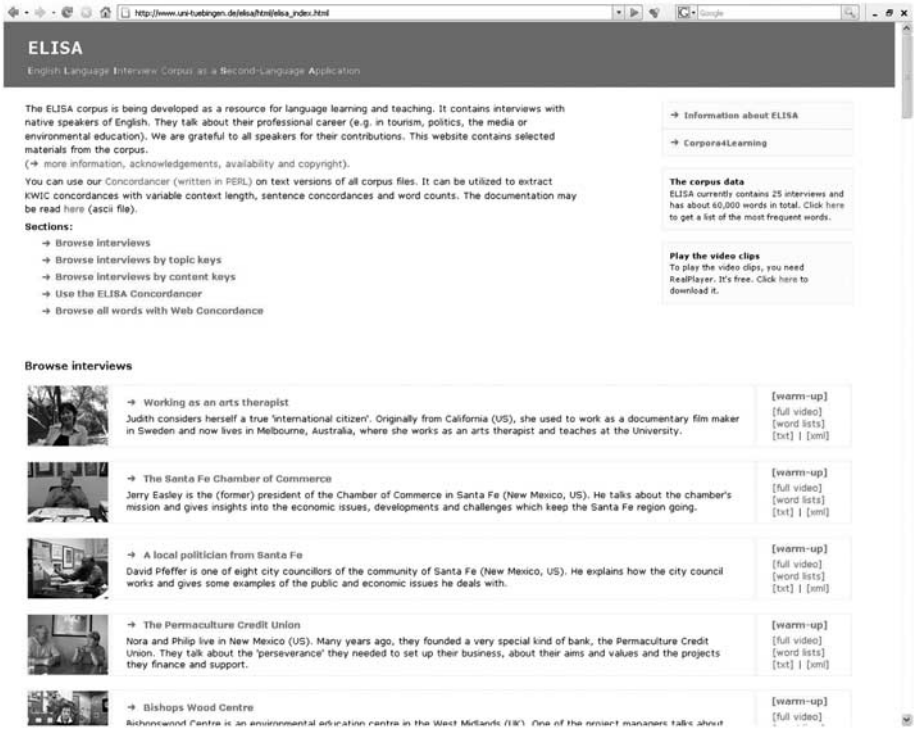


Fig. 2. ELISA web interface

structured and repetitive templates. Mishan and Strunz (2003) have laid emphasis on the possibilities of XML to customize language learning and, very significantly, to provide developers with the opportunity to design and implement applications that are driven by pedagogy rather than by technology.

The annotation in Figure 1 tells us that the information that is pedagogically relevant to learners is that which concerns the topic of the oral text as well as the speaker(s) taking part in the communication process. Visually put, one may argue that the tags here are telling us that pedagogy is driving the annotation, the markup text (enclosed in angle brackets) adding the information that will actually guide learners in their exploring and engaging in corpus-based activities. The annotators have a pedagogical use of the text in mind when approaching the annotation stage. The tags ⟨topic_title⟩, ⟨topic_key⟩ and ⟨content_key⟩ highlight the relevance of the communicative purpose of texts, that is, the topics and the contents that characterize them. Also, ELISA being a multimedia corpus, the markup also includes timestamping information on the related video. The screenshot in Figure 2 shows the ELISA users' interface and the integration of the annotation above.

Despite the simplicity of the markup, we can envisage this annotation approach as a powerful problem-oriented type of annotation where the XML is performing the role of separating the corpus, that is, the linguistic interface, from the learning, pedagogical interface. Figure 3 illustrates this approach.

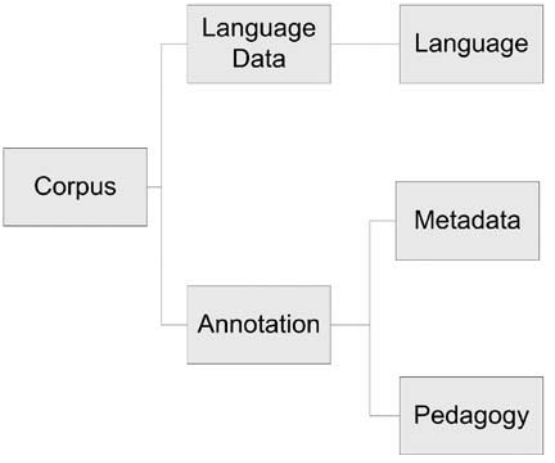


Fig. 3. Pedagogy as add-on

This is an example of what simple XML can do for CL in the language classroom. The resulting annotated corpus in Figure 3 can be seen as being integrative of language data and annotated pedagogy. The former is the liquid, the latter is the container. The interesting aspect here is that pedagogy can be annotated and, subsequently, accessed by corpus users. In the next section we review existing annotation tools and discuss their relevance to the field of pedagogical annotation.

### *3.3 Existing annotation applications*

There is a wealth of annotation solutions adapted to different contexts and users' requirements. In some cases, these tools are geared towards morphological and syntactical language research and, in particular, towards tagging. The tagging adscription can be performed in a manual, semi-automatic or automatic fashion. Examples of the latter include CLAWS (Garside, 1987), TreeTagger (Schmid, 1994) and FreeLing (Atserias, Casas, Comelles, González & Padró, 2006). In all of these applications, the end user is presented with a corpus/text which is tagged according to a language theory and a tagset. It is important that we understand that this tagset is given to the researcher, which demands a high degree of adaptation of the specifications of his study to the tagset characteristics. The degree of flexibility and adaptation of this tagset to the user-specific needs is limited, which may deter potential language educators from becoming involved in projects which are not primarily devoted to morphological research.

However, an increasing number of new tools based on XML-aware annotation have introduced powerful user features. Among these applications we can find AGTK (Bird, 2002), NITE (Reidsma, 2005), ACE (Maeda, 2006) and Callisto (Bayer, 2006). These tools allow for a more flexible representation of annotation, although the specification of user-driven annotation characteristics remains a complex issue. Thus, Callisto forces users to define a XML schema whereas ACE, NITE or Palinka (Orasan, 2003) offer users customization of their tagsets via configuration

of an external file. This fact hinders the utilization of these tools in a FLT or research environment with no ICT expert support. Dexter (Garretson, 2006) is a step forward in getting annotation tools to be used in a more generic and more flexible way in different contexts, including foreign language teachers who want to prepare teaching materials. Dexter codes the annotation of the linguistic corpus using its own XML solution. To start a simple annotation process with this tool, the annotators need to use a converter to process the text they want to annotate into a native XML format. A similar process has to be followed if we use MMAX2,[7] another stand-off solution suitable for multi-layer annotation and widely used in discourse analysis (Braun, Kohn & Mukherjee, 2006b).

In the context of a multinational, multilingual and multi-corpus project, however, the annotation solutions should cater for more complex needs as the language learning experiences of future users of our system beg for a high degree of sophistication.

## 4 Annotation challenges in DDL

In the previous section, we had borrowed the notion of problem-oriented tagging from McEnery and Wilson (1996: 56–7). This term is a special label for those cases where non-traditional, standard grammatical annotation is required so as to fulfil the aim and ultimate use of a corpus. It should be noted that the authors refer to this type of tagging as not being exhaustive and of not providing broad coverage, as well as lacking consensus-based theory neutrality. McEnery and Wilson's (1996) remark is better understood in the context of language analysis and, in particular, in the context of linguistic research, which clearly shows that corpus applications in FLT still need to gain a status in their own right.

While it is true that local problems may call for specific individual solutions, it is undeniable that such a generalization may not hold true for an annotation scheme simply because it is not morphological or morphosyntactic. There are, however, some important challenges for those interested in exploring new horizons in terms of enriching language corpora with pedagogy. These challenges can be classified under three different headings and affect the nature of the pedagogical annotation process: the design of the annotation scheme, the epistemology of the annotation, and the technology that will support the process.

The annotation of a pedagogically relevant corpus should comply with standard practices and, accordingly, good pedagogical annotation should arguably meet Leech's (1993) quality maxims, that is, it should be possible to remove the annotation from the text; if desired, the annotation could be extracted and should be based on guidelines everyone could reach; it should be made clear how and by whom the annotation was carried out and it should be based on widely agreed and theory-neutral principles.

Figure 3 shows conceptually how feasible it should be for anybody with basic computing skills to perform all of these basic operations. Any XML tagging can be easily extracted and removed with any XML editor and, of course, with find and

[7] http://mmax.eml-research.de

replace text tools. These *design* features will give our annotation a rigorous and well-planned structure. The issue of an underlying theory-neutral principle would require extensive treatment, however. As with other areas in applied linguistics, FLT lacks a unified theory that can account for the teaching/learning dichotomy in absolute terms. Ellis (2005) has drawn together the most important contributions from Second Language Acquisition (SLA) research in the area of FLT. Ellis has stressed the need for learner-centred teaching that focuses on form as well as on meaning and which fosters L2 competence based on both formulaic and rule-based declarative knowledge. Ellis also highlights the well-researched need to offer implicit in addition to explicit knowledge of the L2. Given this situation it would be, therefore, difficult for us to decide on an exclusive perspective that adheres to some areas of FLT methodology and neglects others, especially when we aim to offer a generic solution for pedagogy-driven annotation of natural language.

From a more epistemological perspective, Leech (1993) argues that the end user should be made aware that the corpus annotation is not infallible and that no annotation scheme has the *a priori* right to be considered as a standard. While accepting the implications of such a statement, and indeed because of it, we believe that the annotation of a DDL application should meet sound pedagogical presuppositions and foundations.

Antecedents in the DDL literature such as Aston (1997), Bernardini (2000, 2004), Meunier (2002) and Mukherjee (2006) call for DDL solutions that are representative of language situations that resemble real-life tasks and facilitate both inductive and deductive discovery learning. This raises the question whether annotation is instrumental here. We believe that for annotators to annotate such a complex interweaving of pedagogy-related facts, we should strive for a highly-structured approach that could be taken to different foreign language scenarios and which, at the same time, offers enough flexibility to meet the challenges of every specific annotation situation. For a multilingual initiative this means that the annotators of the Lithuanian corpus and the Spanish corpus, to cite two of the languages involved in SACODEYL, may decide on a different annotation structure for the sub-taxonomies of the general, main grammar taxonomy. This is the case because we are not looking at the annotation of every single grammatical feature in the corpus, but rather at those units which are suitable for pedagogical delivery and, accordingly, of interest for further DDL exploitation. A case in point is the pedagogical approach in most FLT textbooks, where the different learning units are not expected to cover or fully describe every single grammar or lexical point that appears in a unit. It is understood that the textbook performs a mediation role between (a) the teaching institution or learning/ teaching initiative and (b) the learner. The role of pedagogically-driven search in this context is of the utmost importance and accordingly the annotation of corpora for the language classroom should bear this subsequent use in mind.

As far as technology is concerned, we conclude that we need a user-friendly tool with multilingual support as our SACODEYL multinational teams of end-users are not computational linguists themselves. Furthermore, we wanted to develop a standard-compliant tool that could be instrumental in more generic contexts and thus favour the re-utilisation and dissemination of the application.

## 5 Developing annotation solutions

In the previous paragraphs we have implicitly discussed the analysis stage of Colpaert's (2004) Analysis-Design-Development-Implementation-Evaluation (ADDIE) Model. As we have shown, the challenges for pedagogical annotation abound. There is one, however, that will determine the design and implementation of the application: the type of use that we want to put our annotation to, that is, the annotated corpora that we will use so as to feed our DDL system. This can be seen in Figure 4.

To do so, we need to embark on a software design, development and implementation engineering process that will allow us to develop an annotation application that is instrumental for our purposes. In SACODEYL, the annotated corpora will be queried by a search tool which will return information that will read the pedagogically-enriched corpora uploaded to our servers.

Most of the CALL literature concerning application design and implementation (Plass, 1998; Levy, 1997) adopts a well-known input-output engineering approach. This is in line with mainstream software design principles and is instrumental in getting the application development started (Larman, 2002 and 2003). Although geared towards language-course development, Colpaert's (2004) Analysis-Design-Development-Implementation-Evaluation (ADDIE) model is an important reference for us. Ward (2006) has stressed the fact that Colpaert's design phase is probably the most crucial of all the stages. This phase is divided into conceptualisation, specification and prototyping. Table 1 shows the main aims of each stage.

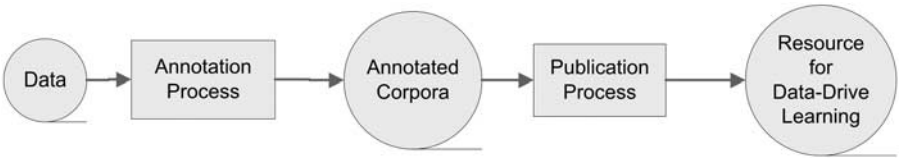


Fig. 4. Deploying online DDL

Table 1 *Stages in Colpaert's (2004) design phase*

| Conceptualisation | Specification | Prototyping |
|---|---|---|
| Concept development. Identifying personas or learner types. | Back-end (the system structure) and the front-end (the user interface) of the system. | Actual development of discrete elements of the system. |
| Realisation of practical goals Description of scenarios of system usage and translating the scenarios into system tasks. Usefulness criteria. usability, usage, user satisfaction, user-friendliness and didactic efficiency. | Description of system components and their interaction. | |

This conceptualization should include all the facts, considerations, principles, assumptions and requirements to be observed before the implementation of the application itself is started. In software engineering literature the specification of requirements is a central, well-documented stage (Larman, 2002). In the light of this, it is essential to avoid any trace of ambiguity in the specification of the requirements. Once these are well-established, we can move on to decide on the type of application architecture that best fulfils them. As a result of this design process, we may obtain eventually (a) the application architecture, (b) an application prototype or even (c) the application itself.

### *5.1 Requirements specification*

In 3 above, we have discussed the challenges that we faced when approaching the main features of an annotation solution for online DDL. Those challenges can be regarded now as shaping the design and implementation of our application. Therefore, our annotation tool can be understood as the *output* of the software engineering process that started with the analysis stage that gave rise to the specification of the requirements of pedagogical annotation.

In order to analyze thoroughly the conceptualization of SACODEYL Annotator, we have opted for a multi-layer framework analysis. To do so, we have used the models by Lee and Xue (1999) and Jain *et al*. (2003). These models integrate different areas that must be carefully reflected upon when dealing with the conceptualization and specification of the features of the tool engineering process. Every dimension in Figure 5 is instrumental in identifying the functional specification of an application.

We have already argued the need for a new bottom-up epistemological approach that shifts the focus of annotation from linguistic research to pedagogy. This new focus also entails a new type of user or users in the annotation debate. In the linguistic-research paradigm, linguists or computational linguists are the end users of tools and technologies such as CLAWS (Garside, 1987) or TreeTagger (Schmid, 1995).

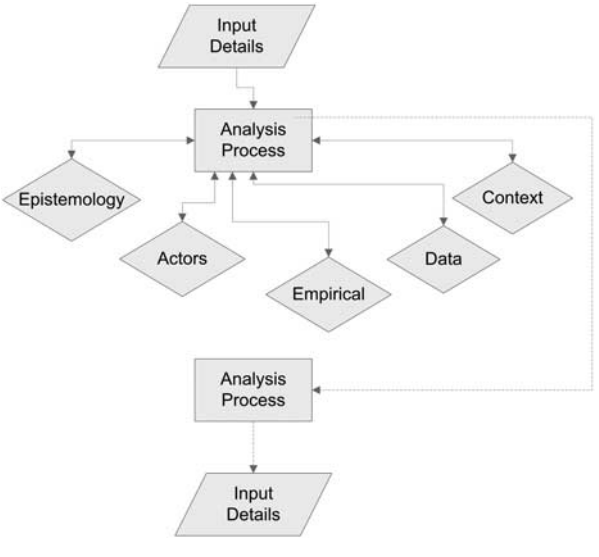


Fig. 5. Dimensions in the analysis process

In DDL, however, the actors are not so easily targeted. Teachers, students and CALL and CL professionals play a role in the making of this type of learning. This will have a highly significant technological effect on the way these new applications are to be designed.

Each type of end-user will demand different non-functional specifications. Language engineers or computational linguists want flexible applications which usually offer complex interfaces, subject-specific and extensible. However, a non-expert annotator such as a language teacher, or even a learner, wants intuitive, easy-to-use interfaces that remain stable and whose learning curve is low. While emphasizing the roles and specifications of the input/ output of the annotation application, we have constantly paid special attention to the needs of the end-users of the application.

In our case, pedagogy and DDL are the specific purposes for the annotation of corpora. Accordingly, such a tool should incorporate, among other things, the possibility for users to annotate discourse topics or the adequacy of a particular text or piece of text for further exploitation in a FLT context. As regards the type and quantity of data to be used as input for the tool, the main consideration here is to find the right balance between data and processing so as to achieve real design efficiency (Banker *et al.*, 1998). It is essential that the tool developers fully understand the profile of the potential users, whether these are linguists, language teachers or language learners.

Table 2 shows the specifications that our annotation approach should meet. Based on these specifications, we are in an optimal position to accomplish the design stage of the engineering process of our own annotation tool.

While actors and data are seen as cross-sectional dimensions, the epistemology, content and technology dimensions match the annotation challenges discussed previously.

### *5.2 Choosing the application life cycle*

In software design, there are numerous approaches to building efficient applications. These approaches, and their methodologies, are commonly referred to as software life cycles. Prototyped life cycles are widely used in developing software products that undergo important changes or modifications in their functionalities. This approach is also used when new technologies are introduced and there is a high degree of uncertainty regarding their impact or acceptance. A prototyped life cycle is advisable if the specifications for the application are not well defined or remain open to further modifications. In this sense, a prototype is a *partial* product, a provisional artefact which is not released for the end user's testing. This approach is very cost-effective and understandably is widely established in advanced software development. By developing and evaluating the prototype, more thorough specifications of the tool can be drafted. The drawback here is that the definition and tool building are repeated twice: once for the prototype and a second time for the final product.

A spiral life cycle is anchored to a different view of software engineering. Each spiral is a cycle, and in every cycle the developers fix and improve the functionalities of the application, as illustrated in Figure 6.

This approach is of use when the end user and/ or client is not fully aware of the output he expects from the application. This is very often the case in multi-organisation projects where expectations vary and, obviously, developers cannot

Table 2 *SACODEYL Annotator specifications*

| Multi-layer framework analysis elements | Epistemology Actors+Data | Context Actors+Data | Technology Actors+Data |
|---|---|---|---|
| Challenges in pedagogical annotation | Annotation/ General CL principles | Pedagogic principles | Supporting technologies |
| Text and annotation can be separated | | Based on widely agreed and theory-neutral principles | It must facilitate the separation of language data, data structure and annotation |
| Annotation can be extracted | | Annotators should be given enough freedom to perform annotation that favours discovery learning | It must support multilingual annotation |
| Text can be extracted | | Annotators should be given enough freedom to perform annotation that favours inductive learning | It should be extensible |
| Annotation must incorporate metadata that identify relevant textual features | | Annotators should be given enough freedom to perform annotation that favours deductive learning | It should be reusable |
| Annotation must incorporate metadata that identify relevant actors in the transcription and annotation process | | It must facilitate the integration of learning resources | It should be standard-compliant |
| Based on guidelines | | | It should be cross-platform |

wait for end users to evaluate the application once the development cycle is over. The illustration in Figure 6 represents this approach: each loop is a development cycle divided into four squares which stand for different stages, namely, planning, design, implementation and evaluation. Each loop increases the maturity of the application.

## 6 An annotation solution for online DDL: SACODEYL Annotator

### *6.1 SACODEYL Annotator*

SACODEYL Annotator is one of the applications developed for SACODEYL (System Aided Compilation and Open Distribution of European Youth Language), a Minerva project funded by the European Commission. SACODEYL Annotator is

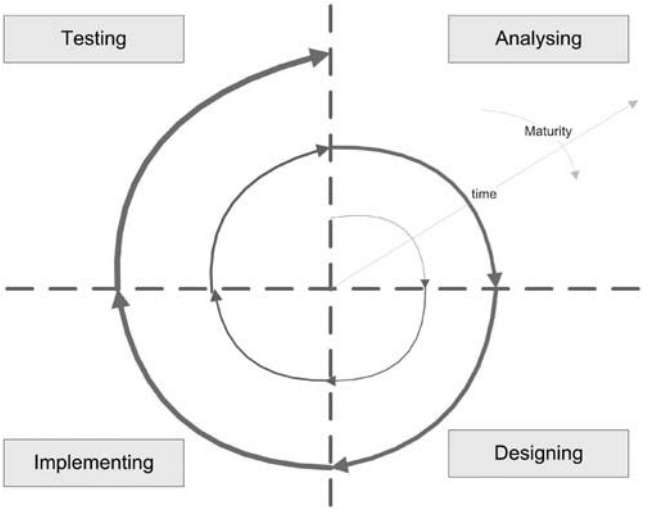


Fig. 6. SACODEYL Annotator spiral life cycle

the product of the analysis and design stages discussed in 3 and 4 and meets the specifications outlined in Table 2. Its main functionality is to provide a software-aided annotation solution for users who need to add structured information to electronic texts. To do so, we have adopted an open and extensible standard for the definition of XML documents: the Text Encoding Initiative (Burnard, 1995).[8] This encoding standard provides our tool with extensibility, interoperability and standardization, three characteristics which we consider of the utmost importance for the re-usability of our annotated corpora (Ward, 2002; Cushion, 2004).

Among other features, the application allows users to define and work with an open taxonomy set of data. This feature makes the application sufficiently generic to claim that it can potentially be used in a wide range of disciplines and fields, from corpus linguistics to forensic analysis of speech. However, the application as-it-is has been implemented to annotate a corpus pedagogically.

When using the application for the first time, users are prompted to create and add a new corpus container (Figure 7). Once this is done, a user will need to create a taxonomy that specifies the set of tags that will be used in the annotation of texts.

In the screenshot in Figure 7, the user has to define a group of TEI-header elements that we have considered necessary for the purposes of SACODEYL. Once a corpus file is created, the user can modify the tree structure. Figure 8 shows a part of the taxonomy tree that has been defined for the Spanish SACODEYL corpus.

This interface allows users to nest and organize the tags in a hierarchical way quite simply. Also, the annotation and the hierarchy tree annotated on a corpus can

[8] Information on the latest TEI developments can be found at http://sourceforge.net/projects/tei/ and http://www.tei-c.org/P5/. SACODEYL Annotator has been implemented following TEI-TP5 directives.

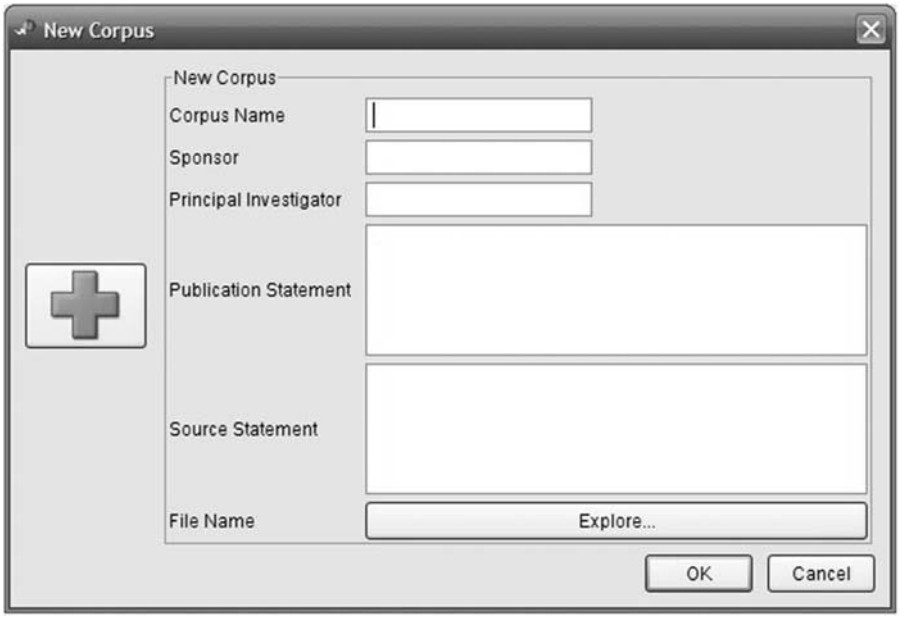


Fig. 7. Adding a new corpus to SACODEYL Annotator

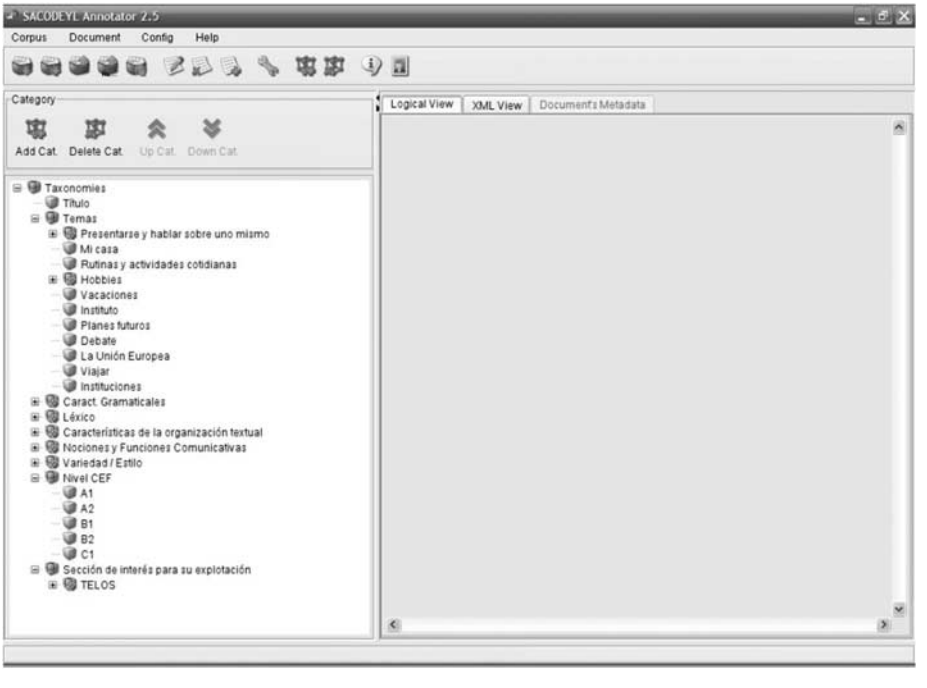


Fig. 8. Annotation taxonomy tree of the SACODEYL Spanish corpus as displayed in SACODEYL Annotator

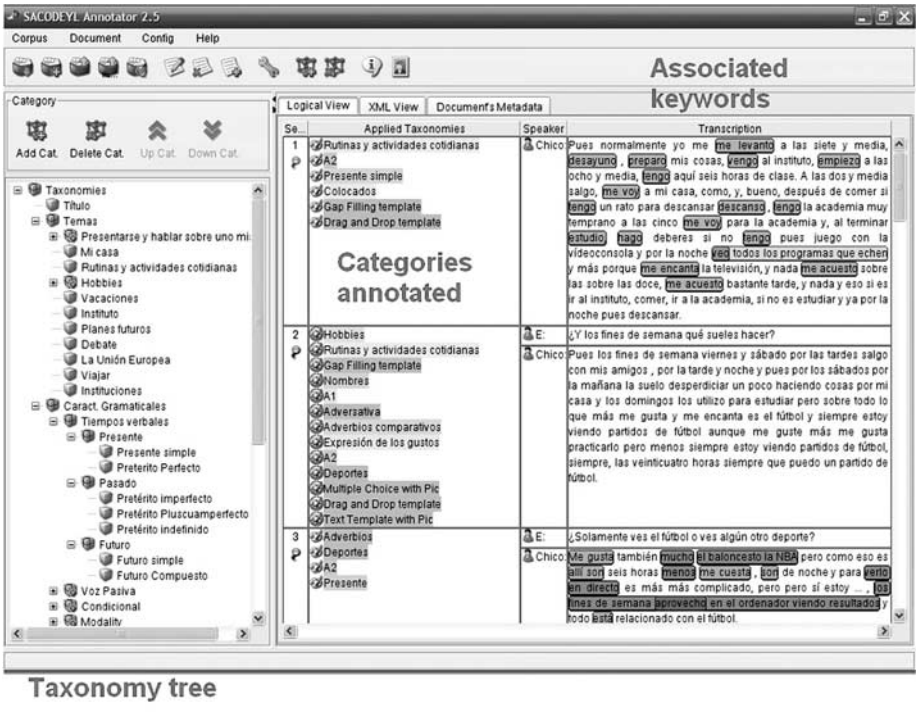


Fig. 9. SACODEYL Annotator work areas

always be modified. Let us try to exemplify this point. A user has already annotated the lexical features of a text or a group of texts. Later on, she needs to add new information to the corpus, for instance she wants to annotate oral discourse phenomena that are not present in the taxonomy. This can be accomplished by creating a new 'child' in the corresponding 'parent', something which will allow this teacher or linguist to keep the information previously annotated while extending the range of information added to the XML file.

Note that it is the user that decides on the hierarchy of this element, which means that this new feature, we call it *category*, can be nested where the annotator deems right. In terms of the generic use of the tool this means that the open annotation we advocate is not conditioned by the discipline or field of research, but rather by the final use or output that we want to obtain.

The application allows users to assign *keywords* to annotated categories. In SACODEYL Annotator, a *keyword* is a stretch of language (a word, more than one word or a whole paragraph) that the annotator associates to a category. In Figure 9 we can see three annotated sections of the Spanish corpus. The general taxonomy tree is on the right hand side, while on the left of the interface we can find the text that is being annotated. In Sections 1 and 3 of Figure 9, we can see the keywords associated to the categories that have been annotated. Keyword views and assignment are optional, and thus section 2 in Figure 9 is annotated but keyword view is disabled.

SACODEYL Annotator has been developed to meet the needs of a very wide range of users, and as a consequence no *a priori* knowledge of CL is needed in order to start

annotation right away. The tool is very easy to use: tag assignment is performed through drag and drop and keyword assignment is performed through select and click basic operations. To facilitate this process, the application filters out the information shown on screen and so users can decide which highlighted keywords they want to see or hide. Secure deleting of the annotation is also provided. The tool is so intuitive that even learners with no CL background might use it to navigate the annotation.

SACODEYL Annotator supports international coding standards such as ANSI, ASCII, ISO-7866 or Unicode (Needleman, 2000), and thus it can be used to annotate texts in any language. The tool has been designed to run on different operating systems and the only requirement is that users have installed Java Runtime Version 6 (JRE 6) v1.6.0 or higher. The tool uses Java Networking Launching Protocol (JNLP) technology and updates automatically if the machine has an Internet connection. JNLP has proven to be an excellent option in the context of SACODEYL as all the members of the consortium have consistently used the latest development of the application.

### *6.2 Evaluation*

So far, the life cycle of our tool has undergone 25 'loops'. These changes have affected areas such as end-user interface design issues, interface multilingual support, refinement of logical and physical views or metadata treatment. Table 3 summarises the aforementioned loops.

Most of the changes to the tool were prompted by the different users of the tool during its life cycle. On the one hand we find transcribers and annotators, on the other the developers. The former were very concerned with the functionalities of the tools (loops 2, 4 and 14 are good examples), while the latter were more focused on bug-fixing and improving the maturity of the tool (for example, loops 13 and 23). Team-work and very fluent communication gave us the chance to discuss the pros and cons of implementing new changes, which has resulted in improved performance and a high degree of satisfaction among the users of the tool in SACODEYL.

## 7 Conclusion and further work

SACODEYL Annotator is an annotation solution for those interested in the implementation of pedagogically motivated corpora in the language classroom. It is part of a larger SACODEYL system and, accordingly, its usefulness is better appreciated when used together with the SACODEYL search tools. Far from a weakness, this is a common feature of XML annotation tools which are largely used to create robust, highly-structured searchable files.

From a theoretical perspective, SACODEYL Annotator addresses an important issue raised by McEnery and Wilson (1997: 8): ''The needs of the researchers primarily involved with the development of corpora have been subtly different from those which would generally be relevant to classroom research''. Not only is classroom research different from linguistic research, but also applied uses of corpora differ widely from those in the scope of CL. While it is a fact that there is a need for specific uses of corpora in the classroom, it is nonetheless true that the efforts to implement pedagogically motivated corpora are in their infancy.

Table 3 *SACODEYL Annotator loops*

| Loop | Ver | Change demanded by | Main cause for the change |
|---|---|---|---|
| 1 | 0.1 | | Original release |
| 2 | 0.2 | Annotators | Section title management incorporated |
| 3 | 0.3 | All | Tool bar incorporated |
| 4 | 0.4 | Annotators | Colour customization associated to categories added |
| 5 | 0.5 | Interview transcribers | Relaxation of the format requirements for the texts |
| 6 | 0.6 | Annotators | Bug when deleting a category and naming a new one with the same ID fixed |
| 7 | 0.7 | All | Materialization of the section added |
| 8 | 0.8 | Annotators | Filter choice for the rendering of keywords in the logical view incorporated |
| 9 | 0.9 | All | New help section added |
| 10 | 1.0 | All | Internationalization and new encoding formats support for the treatment of texts added |
| 11 | 1.1 | All | Rendering of oral-discourse-related annotation such as unclear passages, cuts, breaks, foreign words, etc. added |
| 12 | 1.2 | All | Auto-install and auto-update added |
| 13 | 1.3 | Developers | Debugging mechanism to detect possible errors during the usage of the tool added |
| 14 | 1.4 | Annotators | Categories reordering in the taxonomy tree added |
| 15 | 1.5 | All | XML View/XML TEI document view incorporated |
| 16 | 1.6 | Annotators | Improvement of rendering of the text and annotation trough incorporation of cache mechanism |
| 17 | 1.7 | Annotators | Deletion of only one keyword added |
| 18 | 1.8 | All | Edition capabilities in XML View added |
| 19 | 1.9 | All | Improved algorithm to include the renderization of the overlapping annotations added |
| 20 | 2.0 | All | Incorporation of the Resources management |
| 21 | 2.1 | All | Improved generic uses for the tool: more textual types supported (interviews, dialogues, monologues, etc.) |
| 22 | 2.2 | All | Metadata management added |
| 23 | 2.3 | Developers | Compliance with TEI v0.9 |
| 24 | 2.4 | All | Internationalization support in the user interface added |
| 25 | 2.5 | Developers | Multi-platform support (Windows, Linux, Mac OS, etc.) added |
| 26 | 2.6 | Transcribers | Section comment enabled |
| 27 | 2.7 | Developers | Auto-import mechanism from previous versions of TEI added |

SACODEYL Annotator is a first technological step towards XML TEI-compliant annotation of pedagogy in FLT, and certainly a solution to meet the needs of the type of teachers and students-as-researchers that our project actively seek to promote in the foreign language classroom.

It must be stressed, however, that having served the purpose of our multi-national team SACODEYL Annotator is expected to become a useful tool for the community

of FLT language corpus users in particular, and for those interested in creating multi-layered manually-annotated XML TEI-compliant files, in general. SACODEYL Annotator has proved to be a useful tool in the process of building a mediation effort to bring CL methods to secondary school learners. There is room for improvement in the generalization of the tool to other, less specific fields, including the adaptation of the tool to multi-modal annotation, synchronous and asynchronous collaborative team annotation, export features and the integration of publishing capabilities.

## References

Abe, M. and Tono, Y. (2005) Variations in L2 spoken and written English: investigating patterns of grammatical errors across proficiency levels. In: *Corpus Linguistics Conference* http://www.corpus.bham.ac.uk/PCLC/CL2005proceedings_AbeTono.doc

Atserias, J. Casas, E. B., Comelles, M. González, L. and Padró, M. (2006) FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC'06). Genoa, Italy.

Banker, R. D., Davis, G. B., and Slaughter, S. A. (1998) Software development practices, software complexity, and software maintenance performance: a field study. *Management Science*, **44**(4): 433–450.

Bax, S. (2003) CALL – past, present and future. *System*, **31**(3): 13–28.

Bernardini, S. (2000) *Competence, capacity, corpora. A study in corpus-aided language learning*. Bologna: CLUEB.

Bernardini, S. (2004) Corpora in the classroom: An overview and some reflections on future developments. In: Sinclair, J. McH. (ed.), *How to Use Corpora in Language Teaching*. Amsterdam; Philadelphia: J. Benjamins, 15–36.

Biber, D. and Finnegan, E. (1991) On the exploitation of computerized corpora in variation studies. In: Aijmer, K. and Altenberg, B. (eds.), *English corpus linguistics. Studies in honour of Jan Svartvik*. London: Longman, 204–220.

Biber, D., Conrad, S. and Reppen, R. (1998) *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Braun, S. (2005) From pedagogically relevant corpora to authentic language learning contents. *ReCALL*, **17**(1): 47–64.

Braun, S., Kohn, K. and Mukherjee, J. (eds.) (2006) *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Frankfurt/M: Peter Lang.

Braun, S. (2006a) ELISA – a pedagogically enriched corpus for language learning purposes. In: Braun, S., Kohn, K. and Mukherjee, J. (eds.), *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Frankfurt/M: Peter Lang, 25–47.

Braun, S., Kohn, K. and Mukherjee, J. (2006b) Multi-Level Annotation of Linguistic Data with MMAX2. In: Braun, S. Kohn, K. and Mukherjee, J. (eds.), *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods* (English Corpus Linguistics, Vol. 3). Frankfurt: Peter Lang, 197–214.

Braun, S. (2007a) Designing and exploiting small multimedia corpora for autonomous learning and teaching. In: Hidalgo, E., Quereda, L. and Santana, J. (eds.), *Corpora in the Foreign Language Classroom. Selected papers from TaLC 2004*. Amsterdam: Rodopi, 31–46.

Braun, S. (2007b) Integrating corpus work into secondary education: from data-driven learning to needs-driven corpora. *ReCALL*, **19**(3): 307–328.

Burnard, L. (1995) The Text Encoding Initiative: an overview. In: Leech, G., Myers, G. and Thomas, J. (eds.), *Spoken English on Computer: Transcription, Markup and Applications*. London: Longman, 69–81.

Campbell, D. F., Mcdonnell, C., Meinardi, M. and Richardson, B. (2007) The need for a speech corpus. *ReCALL*, **19**(1): 3–20.

Chambers, A. (2007) Popularising corpus consultation by language learners and teachers. In: Hidalgo, E., Quereda, L. and Santana, J. (eds.), *Corpora in the Foreign Language Classroom. Selected papers from TaLC 2004*. Amsterdam: Rodopi, 3–16.

Colpaert, J. (2004) *Design of Online Interactive Language Courseware. Conceptualization, Specification and Prototyping, Research into the Impact of Linguistic-didactic Functionality on Software Architecture*. Antwerp: University of Antwerp.

Cushion, S. (2004) Increasing accessibility by pooling digital resources. *ReCALL*, **16**(1): 41–50.

Ellis, R. (2005) Principles of instructed language learning. *System*, **33**(2): 209–224.

Flowerdew, J. (1993) An educational, or process, approach to the teaching of professional genres. *ELT*, **47**(4): 305–316.

Frankenberg-Garcia, A. (2005) Pedagogical uses of monolingual and parallel concordances. *ELT*, **59**(3): 189–198.

Garside, R. (1987) The CLAWS word-tagging system. In: Garside, R., Leech, F. and Sampson, G. (eds.), *The Computational Analysis of English*. London: Longman, 30–41.

Garside, R., Leech, G. and McEnery, A. (eds.) (1997) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.

Gavioli, L. and Aston, G. (2001) Enriching reality: language corpora in language pedagogy. *ELT*, **55**(3): 238–246.

Hidalgo, E., Quereda, L., and Santana, J. (2007) Corpora in the Foreign Language Classroom. In: *TALC 2004. Proceedings*. Amsterdam: Rodopi.

Jain, H., Vitharana, P. and Zahedi, F. M. (2003) An assessment model for requirements identification in component-based software development. *Special Interest Group on Management Information Systems*, **34**(4): 48–63.

Larman, C. (2002) *Applying UML and Patterns: An Introduction to Object-Oriented Analysis and Design and the Unified Process*. Indiana: Prentice Hall PTR.

Larman, C. (2003) *Agile and Iterative Development: A Manager's Guide*. New York: Addison-Wesley Professional.

Lee, J. and Xue, N. L. (1999) Analyzing user requirements by use cases: a goal-driven approach. *IEEE Software*, **16**(4): 92–101.

Leech, G. and Candlin, C. N. (eds.) (1986) *Computers in English Language Teaching and Research*. London: Longman.

Leech, G. (1986) Automatic grammatical analysis and its educational applications. In: Leech, G. and Candlin, C. (eds.), *Computers in English Language Teaching and Research*, 205–215.

Leech, G. (1991) The State of the Art in Corpus Linguistics. In: Aijmer, K. and Altenberg, B. (eds.), *English corpus linguistics. Studies in honour of Jan Svartvik*. London: Longman, 8–29.

Leech, G. (1993) Corpus Annotation Schemes. *Literary and Linguistic Computing*, **8**(4): 275–281.

Levy, M. (1997) Theory-driven CALL and the development process. *Computer Assisted Language Learning*, **10**(1): 41–56.

Mauranen, A. (2004) Spoken – general: Spoken corpus for an ordinary learner. In: Sinclair, J. McH. (ed.), *How to Use Corpora in Language Teaching*, 89–105.

McCarthy, M. and O'Dell, F. (2006) *English Collocations in Use Intermediate*. Cambridge: Cambridge University Press.

McEnery, A. M. and Wilson, A. (1996) *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McEnery, A. M. and Wilson, A. (1997) Corpora and language teaching. *ReCALL*, **9**(1): 5–14.

Meunier, F. (2002) The pedagogical value of native and learner corpora in EFL grammar teaching. In: Granger, S., Hung, J. and Petch-Tyson, S. (eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: Benjamins, 119–142.

Mishan, F. (2004) Authenticating corpora for language learning: a problem and its resolution. *ELT*, **58**(3): 219–227.

Mishan, M. and Strunz, B. (2003) An application of XML to the creation of an interactive resource for authentic language learning tasks. *ReCALL*, **15**(2): 237–250.

Mukherjee, J. (2006) Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods. In: Braun, S., Kohn, K. and Mukherjee, J. (eds.), *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Frankfurt/M: Peter Lang, 1–4.

Needleman, M. (2000) The Unicode Standard. *Serial Review*, **26**(2): 51–54.

O'Keeffe, A., McCarthy, M. and Carter, R. (2007) *From Corpus to Classroom*. Cambridge: Cambridge University Press.

Owen, C. (1996) Do concordances require to be consulted? *ELT J*, **50**(3): 219–224.

Plass, J. L. (1998) Design and evaluation of the user interface of foreign language multimedia software: a cognitive approach. *Language Learning & Technology*, **2**(1): 35–45.

Poesio, M. and Artstein, R. (2005) Annotating (Anaphoric) Ambiguity. *Corpus Linguistics Conference* 2005: http://ron.artstein.org/publications/anaphoric-ambiguity.pdf

Santos Pereira, L. (2004) *Spoken – an example:* The use of concordancing in the teaching of Portuguese. In: Sinclair, J. McH. (ed.), *How to Use Corpora in Language Teaching*, 109–122.

Schmid, H. (1995) *TreeTagger – a language independent part-of-speech tagger*. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Sinclair, J. (2004) *How to Use Corpora in Language Teaching*. Amsterdam; Philadelphia: J. Benjamins.

Ward, M. (2002) Reusable XML technologies and the development of language learning materials. *ReCALL*, **14**(2): 285–294.

Ward, M. (2006) Using Software Design Methods in CALL. *Computer Assisted Language Learning*, **19**(2–3): 129–147.

Weber, J. J. (2001) A concordance- and genre-informed approach to ESP essay writing. *ELT*, **55**(1): 14–20.