(C-159)

# THE ROLE OF CORPUS LINGUISTICS IN DEVELOPING INNOVATION IN DATA-DRIVEN LANGUAGE LEARNING.

*Pascual Pérez-Paredes*

*María Sánchez-Tornel*

*José M. Alcaraz Calero*

**(C-159) THE ROLE OF CORPUS LINGUISTICS IN DEVELOPING INNOVATION IN DATA-DRIVEN LANGUAGE LEARNING.**

*Pascual Pérez-Paredes[a], María Sánchez-Tornel[a], José M. Alcaraz Calero[b]*

**Afiliación Institucional:** [a] Departamento de Filología Inglesa. Universidad de Murcia ,
[b] Cloud and Security Lab. Hewlett-Packard Research Laboratories

## Indique uno o varios de los siete Temas de Interés Didáctico:

[X] Metodologías didácticas, elaboraciones de guías, planificaciones y materiales adaptados al EEES.

[ ] Actividades para el desarrollo de trabajo en grupos, seguimiento del aprendizaje colaborativo y experiencias en tutorías.

[X] Desarrollo de contenidos multimedia, espacios virtuales de enseñanza- aprendizaje y redes sociales.

[ ] Planificación e implantación de docencia en otros idiomas.

[ ] Sistemas de coordinación y estrategias de enseñanza-aprendizaje.

[x ] Desarrollo de las competencias profesionales mediante la experiencia en el aula y la investigación científica.

[ ] Evaluación de competencias.

## Resumen.

El *data-driven learning* o aprendizaje basado en datos se caracteriza por el uso de bases de datos de lengua en uso que los alumnos analizan para identificar patrones de uso. El aprendizaje a partir de datos encaja a la perfección dentro de los nuevos paradigmas de enseñanza y se ajusta a lo que se conoce como aprendizaje en el S.XXI o aprendizaje para la vida. En este nuevo escenario docentes y discentes deben adoptar nuevos roles. Los alumnos deben hacerse responsables de su propio aprendizaje y deben actuar como agentes activos en el proceso de aprendizaje y no como meros receptores de información. Por su parte, los profesores se convierten en guías o en facilitadores del proceso. Al hacer a los alumnos comportarse como investigadores, el aprendizaje a partir de datos es, por lo tanto, un ejemplo representativo de aprendizaje centrado en el alumno. Además, este enfoque promueve el aprendizaje inductivo, ya que el análisis de datos, la formulación de hipótesis y la extracción de  conclusiones son los tres pilares en los que éste se sustenta. A pesar de los beneficios derivados de esta metodología de trabajo basada en corpus lingüísticos, sus aplicaciones en el aula se han basado  tradicionalmente en la transferencia directa de los métodos y las herramientas que se usan para el análisis de la lengua en el ámbito investigador, lo que causa problemas para la implantación y expansión de esta metodología en contextos educativos.

El tipo de aprendizaje basado en datos que proponemos en este trabajo está basado en un enfoque novedoso en lo concerniente al modo en el que los datos lingüísticos son tratados por los investigadores/profesores y por los alumnos. El uso de un nuevo modelo que favorezca una transformación de datos en información significativa para los alumnos es la clave de nuestra forma de abordar la innovación educativa en el aprendizaje de lenguas. Los proyectos europeos, *SACODEYL* y *Corpora for Content and Language Integrated Learning*, son buena muestra de experiencias de innovación en el campo del aprendizaje para la vida.

**Keywords:** language learning and teaching, content, implementation, innovation, learning, distance learning, research

## Abstract.

Data Driven (language) Learning (DDL) is characterized by the use of language data in the language learning classroom so that students can analyse language and identify patterns of use. DDL fits well with contemporary learning paradigms and with the so-called 21$^{st}$ C learning or lifelong learning, which implies a new attitude on the part of students and teachers. In this new scenario, students need to take responsibility over their own learning and become active learners, and not mere recipients of information. Teachers turn themselves into guides and facilitators of the learning process. In making students work as researchers, DDL is therefore a representative example of learner-centred teaching. Moreover, this approach fosters inductive learning, as the process of analyzing data, formulating hypotheses and deriving conclusions is at the heart of this approach. However, classroom applications of traditional corpus linguistics have relied on heavy linguistic research paradigms, which according to different authors has problematized the use of this methodology.

In the context of our proposal, the data-driven culture that we want to foster is based on a totally new approach to the way in which language data are treated by researchers/teachers and learners. The use of a new data model which favours a more rapid transformation into information which is meaningful to learners is at the hub of our approach to innovation. Two European projects, SACODEYL and Corpora for Content Language Integrated Learning are examples of innovation in the field of lifelong learning.

## Text

### 1. Data-driven learning: an opportunity for lifelong learning in the field of language education

In 1991, Tim Johns published "Should you be Persuaded - two samples of data-driven learning materials", a pioneering work in corpus linguistics which offered a glimpse of the potential role that attested uses of language recorded in real communicative contexts could play in foreign language learning and teaching. These uses were subsumed under "Data-driven learning", DDL, a term difficult to translate into other languages, even difficult to translate into everyday, mainstream language teaching methodology.

DDL is characterized by the introduction of language data in the language learning classroom so that students can analyse language and identify patterns of use. Language data are most often extracted from linguistic corpora by means of concordancers. A concordancer is a computer programme used to search a corpus which produces concordance lines, our language data, as output (Figure 1).

Figure 1 Traditional concordance lines.

These concordance lines show how a node, the word or words which appear in the middle, is used in different contexts; each line extracted from a text which is representative of real language in different contexts of use. The underlying idea is that learners can grasp a better understanding of lexis and grammar by accumulating instances of use in the form of lines which are representative of the wide range of uses of a given node. Languages like PERL or LMS such as Moodle offer ready-to-use concordancers that can be easily integrated in websites or desktop applications.

While indirect applications of corpora to language teaching involves material writers or researchers accessing a corpus to inform foreign language syllabus design or materials development such as dictionaries, course books, grammars, etc., direct approaches i.e., DDL, entail an eminently hands-on exploitation of the corpus by the students and their teachers for language learning purposes (Römer, 2008). Teachers can opt for mediating between the students and the corpus by selecting relevant samples, concordance lines, to be used in class or, in a less controlled way, can make the learners operate directly with the corpus. In any case, the ultimate aim is to involve students in a learning scenario where they become "detectives" (Johns, 2002:108) or "linguistic researchers" (*ibid*) in search of relevant evidence in corpus data from which they can draw conclusions and answer questions about the language. This idea is closely related to Fligelstone's (1993) 'exploiting to teach' dimension. He sees three interrelated dimensions in the direct application of corpora to language teaching: 'teaching about', or teaching about corpus linguistics as a discipline, 'teaching to exploit', training students in the use of concordancers and the analysis of corpus data, and 'exploiting to teach', that is, teaching language with corpus-based resources.

The benefits of DDL are many in number and diverse in nature. From a purely pedagogic perspective, DDL fits well with contemporary learning paradigms and with the so-called 21$^{st}$ C learning or lifelong learning, which implies a new attitude on the part of students and teachers. In this new scenario, students need to take responsibility over their own learning and become active learners, and not mere recipients of information. Similarly, teachers turn into guides and facilitators of the learning process. In making students work as researchers, DDL is therefore a representative example of learner-centred teaching. Moreover, this approach fosters inductive learning, as the process of analyzing data, formulating hypotheses and deriving conclusions is at the heart of this approach. Besides, there is evidence that human beings have "evolved to be good at noticing regularities in nature, interpreting them and extrapolating to other cases" (Boulton, 2009:84). From a more language-oriented perspective, data-driven learning is a very versatile approach which can be adapted to different facets of language teaching. The fact that corpora offer huge amounts of authentic language helps overcome the problems posed by the use of decontextualised or invented examples and makes of corpora tools with an enormous "learning potential" (Römer, 2008:120), turning them into some sort of "tireless native-speaker" (Barnbrook 1996 in Römer, 2008:120) that is always available.

Although corpus data have been recognized as an excellent aid to teach the fuzzy boundary between "syntax and lexis" (Johns 2002:109), their potential goes far beyond. Corpus tools and methods in general and DDL and concordancing in particular have been employed to teach aspects as diverse as listening, speaking, reading or writing. Apart from their learning potential, DDL fosters learning and linguistic awareness and favours better recall and retention of the contents learned through exposure to data, given the high degree of task involvement demanded by this approach (Bernardini, 2000). It has also been argued that DDL presents important advantages over traditional rule-based approaches, as the latter "do not provide an accurate picture of language in general, which adheres to patterns, tendencies and generalisations of prototypical usage rather than rigid right or wrong" (Boulton, 2009:84).

## 2. DDL: just a good idea?

The field of corpus linguistics (CL) has opened new paths for the analysis of language that go beyond the view of the word as the hub of language analysis. Sinclair (1991, 2004), a key figure to understand modern lexicography (cf. COBUILD project) and the use of electronic corpora in modern linguistics, stated that the use of these research methods unveiled new units of meaning that just remained hidden until very recently. Sinclair (1991) argued that the units of meaning are phrasal in nature, not word-bound. Together with his analysis of collocation, Sinclair's main contribution to linguistics may well be the so-called idiom principle, that is, the tendency of language to be idiomatic, recursive and repetitive on a phrasal level. In the obituary that Michel Hoey wrote after John Sinclair's passing[1], he described this principle in the following terms: "when we speak or write, we make use of semi-preconstructed phrases, and these account for fluency in speech, with grammar being used as a fall-back resource when we run out of suitable semi-preconstructed phrases". This definition of the idiom principle encapsulates the far-reaching implications as well as the challenges of CL.

---

[1] http://www.guardian.co.uk/news/2007/may/03/guardianobituaries.obituaries

For decades, grammar and word-level vocabulary have been predominant in language teaching, which has driven to frustration to literally thousands of learners that have disdained the idiom principle, not for lack of interest, simply because it was not, and still is not, in the agenda of mainstream language teaching. The combination of language corpora and ad-hoc search tools is essential in making the results of the analysis of the new units of language available in the field of applied linguistics to language learning and teaching. Huston and Francis (2000) have proposed that the pattern should be the unit of linguistic analysis. While the proposal makes sense within the scope of CL, and most corpus linguists will almost instantly agree on the adoption of such unit as the basis for language research, we wonder whether this unit may serve as well as the new unit for language learning and teaching.

In a world where the use of computers for language learning is becoming normalized (Bax, 2003), the use of DDL is still in its infancy. This situation is surprising when we examine how the use of information technologies has impacted computer mediated communication (CMC), collaborative learning as well as computer assisted language learning (CALL). Despite the appropriateness of the approach, data-driven learning is far from being mainstream among language teaching practitioners. Back in the late 90s, McEnery and Wilson (1997) described the expansion of corpora in the UK up to that point as a process of percolation and the situation does not seem to have changed significantly since then. Although it has been gaining momentum in the last ten or fifteen years, the implementation of DDL techniques has, for the most part, been carried out by researchers and specialists close to the field Corpus Linguistics who have incorporated corpus tools and methods into their teaching, mainly in tertiary levels and with advanced students. Mukherjee (2004) draws attention to the scant impact that corpus methods have experienced so far on language teaching in general and in non-tertiary settings in particular. He argues that this situation is a result of the lack of familiarity of teachers with applied corpus linguistics and suggests that institutionalised workshops aiming at showing teachers the potential of corpus-based resources can serve to bridge the gap between language teaching and applied corpus linguistics. Boulton (2009) identifies a convergence of interrelated factors pertaining to learners, teachers, and resources that slow down the spread of DDL. In the lines of Mukherjee, he points at the limited knowledge of applied corpus linguistics of most teachers but he goes further and suggests that, even when teachers are aware of the potential of DDL, they disregard it because they consider it "dangerous" (Boulton, 2009:93).

Data-driven learning is dangerous in the sense that it clashes with traditional conceptions of the role of the teacher. They "have been trained to be the knower, the *fons et origo* of language and pedagogy in the classroom" (Boulton, 2009:93) and DDL advocates a learning environment where teachers are no longer the only owners of knowledge and are required to leave the starring role to the learners and step aside to become guides and facilitators. On a different level, learners themselves can find DDL 'less comfortable' than the traditional learning approaches they are used to. Although there is evidence suggesting that students can become skilled corpus users (Bernardini, 2000), feelings of frustration or dissatisfaction with corpus-based resources are not unusual (Kennedy & Miceli, 2001; Lavid, 2007; Yoon & Hirvela, 2004). Boulton, as many other specialists, claims that training in software use and language data analysis is key to reducing the negative response of students to DDL. The last source of obstacles to the spread of DDL comes from the side of resources. This encompasses both the "dearth of published materials" offering DDL activities (Boulton, 2009:97) and the accessibility to computer labs, corpora and freely available software packages for corpus exploitation. This concern with resources is common among the adopters of data-driven learning, who demand easy to use software that can contribute to the creation of a "DDL-friendly environment" (Römer, 2008:123).

**3. Lifelong learning: an opportunity for innovation in corpus linguistics**

Perez-Paredes (2010) has argued that DDL has not attracted the attention of language learning and teaching for a variety of reasons. His main argument is that DDL has relied almost exclusively on the transfer of CL research paradigm, tools and data to the language classroom, whether virtual or traditional. In the past, learning activities involving language corpora have been characterized by the use of research-oriented corpora, such as the British National Corpus, and the reliance on concordancing as a mediation tool between the corpus and the learner by means of research tools such as WordSmith or MonoConc, accompanied by the absence of an effort to make user interfaces less linguist-oriented. But most significantly, this approach to the exploitation of language data has been characterized by the very high cognitive demand which is put on learners when prompted to read and interpret concordance lines. In this context, innovation is to play an important role if DDL wants to be part of the language learning game in the forthcoming future.

Aceto, Dondi & Nascimbeni (2010) have pointed that one of the problems of innovation is that, in early stages, innovation is more disruptive than consensus-building. Their study looks at the uncertainties affecting future learning and identifies two multidimensional trends: convergence vs. context and innovation vs. inertia. After identifying four scenarios, or quadrants of uncertainty, they were the object of a Delphi consultation which involved 50 European experts. In summary, higher education was correlated with the innovation/context quadrant in which education becomes a catalyst of change and innovation. Distance education, on the contrary, was correlated with a scenario most likely to be characterized by global solutions and the concept of learning *fast foods*.

In the past five years, a group of European higher education institutions have stepped up to meet the challenge of developing tools and compile data that, while profiting from the research tradition and know-how of CL, could also pave the way to establish innovative DDL which differs from the research paradigm of CL. The first ground-breaking initiative promoting innovative DDL was the EU-funded research project SACODEYL (*System-Aided Compilation and Open Distribution of European Youth Language[2]*). The project aimed at bringing European L2 learners closer to the discourse of native speakers their age. The initiative adopted an ICT-based approach consisting in compiling multimodal corpora in seven European languages[3] and developing learning materials from them. Each corpus consists of 20-25 videotaped interviews with teenagers from seven EU countries which were transcribed and pedagogically annotated. In a later stage, different types of corpus-based exploratory and communicative activities were developed.

SACODEYL corpora do not only offer language data but also a large number of ready-made activities. The pedagogic relevance of the corpora and the materials was key to the consortium. It was ensured by selecting a language variety and a range of topics (daily routines, holidays, plans for the future, the EU, hobbies, school and education, etc.) which suit the needs of secondary education students. Figure 2 shows SACODEYL English corpus in browse mode:

---

[2] http://www.um.es/sacodeyl
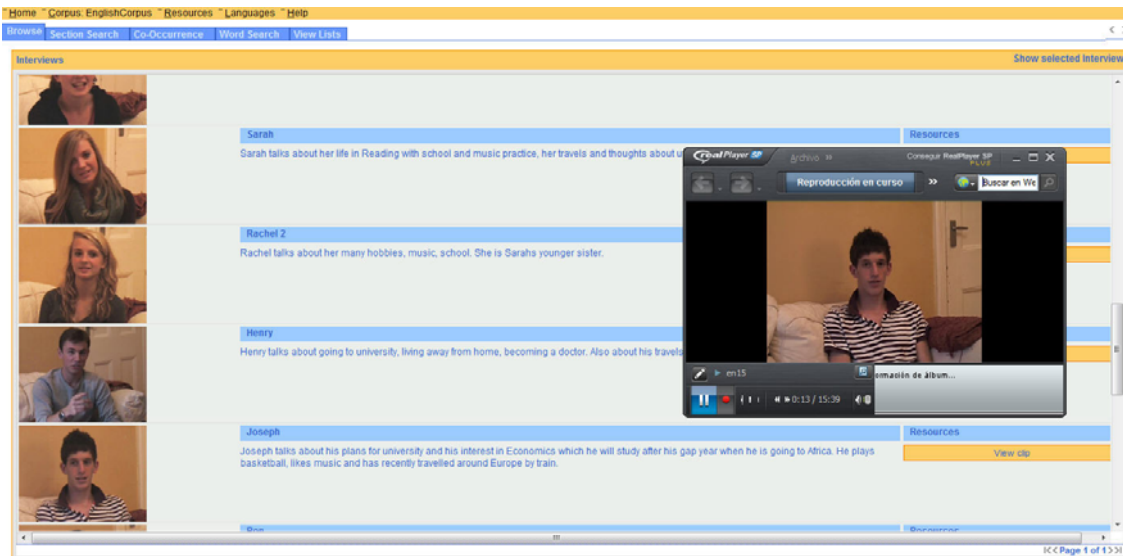[3] English, French, German, Italian, Lithuanian, Romanian, and Spanish.

Figure 2 SACODEYL English corpus in browse mode

Two aspects make SACODEYL a highly innovative initiative from a pedagogic perspective. First, it concentrates on youth language, a variety often ignored in traditional language learning materials. This way, contrary to the situation found when large, general corpora covering multiple registers and genres are brought to the classroom, the language found in the data is close to the learners' reality. Secondly, the topics covered in the interviews reflect the interests of teenagers which may constitute a potentially motivating factor. Figure 3 shows part of the section search interface for the Spanish corpus:
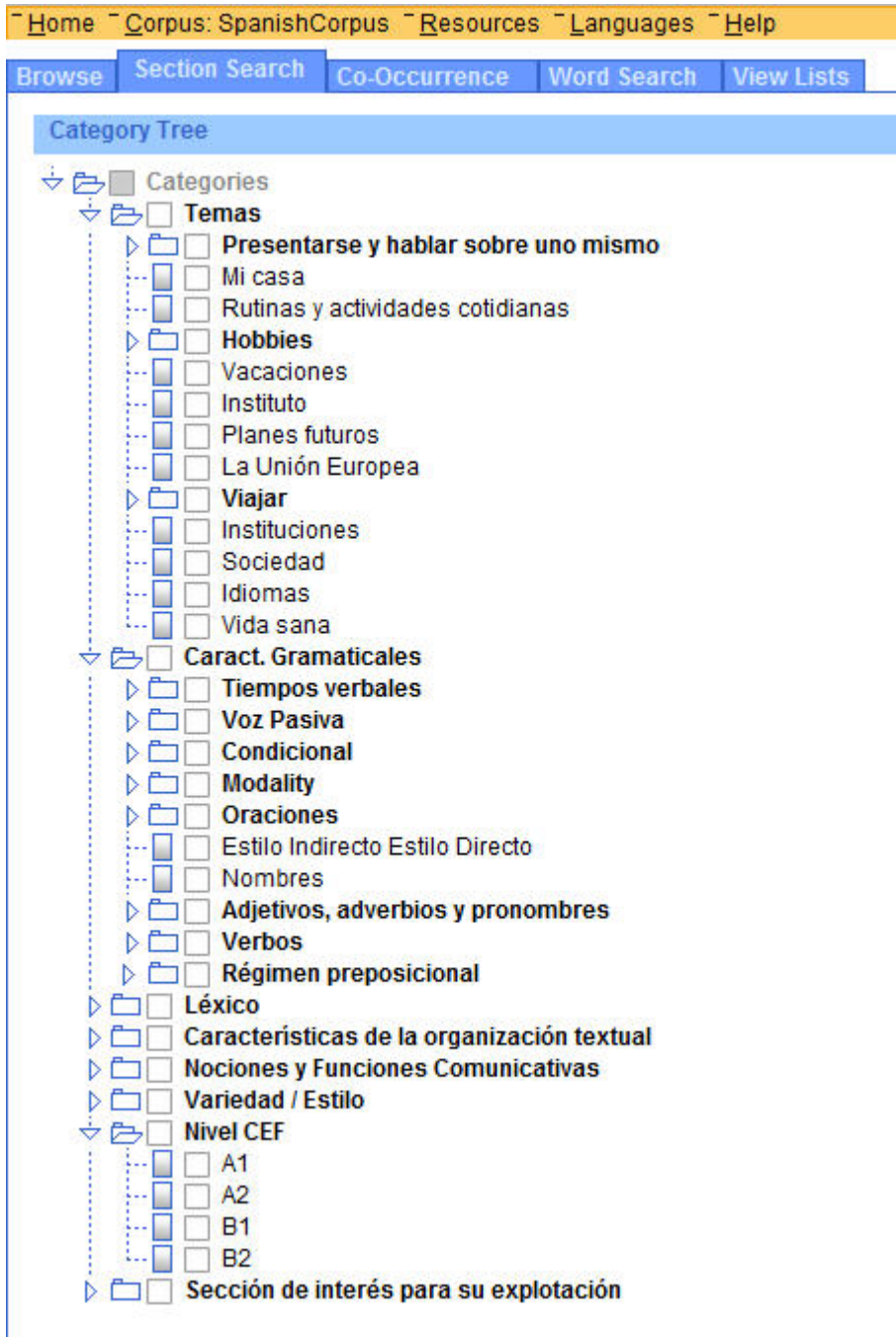
Figure 3 Part of the section search interface for the SACODEYL Spanish corpus

From a technical point of view, the development of a set of *ad hoc* tools for corpus compilation and exploitation also constitutes an element of originality (Pérez-Paredes & Alcaraz, 2009). This includes a transcriptor, a pedagogic annotator, a tool for searching the corpus in different modes (word, coocurrence, section, interview searches), and a materials repository. The tools are freely available so that teachers can either enrich their own corpora or customise SACODEYL language data according to their needs.

A second initiative that aims at implementing an innovative DDL approach is *Corpora for Content and Language Integrated Learning (BACKBONE[4])*. This project takes up the philosophy and rationale behind SACODEYL and proposes the compilation and subsequent pedagogical exploitation of 7 oral corpora enriched with learning materials suitable for CLIL settings (Sánchez-Tornel, Alcaraz & Pérez-Paredes, 2009). The project aims at fulfilling the need for materials for the teaching of pedagogically neglected languages and varieties. Thus, the sub-corpora contain interviews in lesser taught languages (Polish and Turkish), regional and socio-cultural varieties of more frequently taught languages (German, French, Spanish and English) and non-native speaker varieties of English as a lingua franca. The language gathered in the corpora adjusts to the needs of language learners in secondary, higher and vocational education settings.

The principle of pedagogic relevance remains invariable and, therefore, the interviewees belong to diverse professional domains to ensure the suitability of the corpora for students in different CLIL settings. The list of topics covered in the interviews is comprehensive and varied, including aspects such as economic issues, education and research, health and social security, rural and urban life, government and politics, social issues, institutions, etc. Technical innovation is a constant concern of the consortium and several improvements and functionalities have been implemented in the tools developed under the SACODEYL project. The main element of novelty in this respect is the development of a *Corpus Management Tool* which supports collaborative annotation. This new functionality allows several teachers (or annotators) to manage and annotate the same corpus simultaneously, which results in richer annotations.

Bambrick-Santoyo (2010) has outlined a theory of learner-cantered data-driven instruction which is founded on assessment, analysis, action, and a data-driven culture. In the context of our proposal, the data-driven culture that we want to foster is based on a totally new approach to the way in which language data are treated by researchers/teachers and learners. The use of a new data model which favours a more rapid transformation into information which is meaningful to learners is at the hub of our approach to innovation. Data in the web 2.0 era is multimodal, searchable, exportable and transformable. SACODEYL and BACKBONE are building the foundations of such approach and reflect the dimensions for educational innovation at the *innovative* or *most innovative* levels outlined by Law et al (2011), i.e., learning objectives, teacher's role, student's role, ICT use, connectedness, and multiplicity of learning outcomes. Regarding the learning objectives, our initiatives promote inquiry and communication skills. Besides both students and teachers adopt new roles; the former becoming proactive and getting involved in inquiry activities and the latter becoming facilitators and providing support during the above mentioned inquiry process. In an innovative context, the use of ICT is not enough. What makes a learning environment truly innovative, according to Law et al. (2011), is the fact that ICT tools include data analysis software and support both synchronous and asynchronous communication

---

[4] http://www.um.es/backbone

(distinctive features of the SACODEYL and BACKBONE tools). Connectedness is also present in these two initiatives, both in relation with the development of teaching materials and with respect to the possibility of collaborating with and feeding on agents from other countries. Finally, a multiplicity of learning outcomes is exhibited, since the learning tasks that can be developed from the data contained in the SACODEYL and BACKBONE corpora and require students to develop research and problem solving skills as well as other 21st C skills, thus fostering not only the acquisition of content but also the development of competencies and skills that will be instrumental for them in their future career.

Our ultimate goal is to take our know-how to the cloud and involve more European stakeholders in the process of developing data-driven learning experiences which are truly learner-oriented. SACODEYL has received so far 15.000 visits and counting; BACKBONE is expected to be a major breakthrough in the field, as professionals are expected to find in this resource an approach which truly appeals their needs.

## References

Aceto, S. Dondi, C.  & Nascimbeni, F. (2010). Vision for Learning in Europe in 2025.  *eLearningPapers*, 18. URL: http://www.elearningeuropa.info/files/media/media21858.pdf

Bambrick-Santoyo, P. (2010). Driven by data. A practical guide to improve instruction. San Francisco, CA: Jossey-Bass.

Barnbrook, G. (1996). *Language and Computers. A Practical Introduction to the Computer Analysis of Language*. Edinburgh: Edinburgh University Press.

Bax, S. (2003) "CALL - past, present and future", *System* 31, 1:13-28.

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

 Boulton, A. (2009). Data-driven learning: Reasonable fears and rational reassurance. *Indian Journal of Applied Linguistics, 35*, 81-106.

Fligelstone, S. (1993). Some Reflections on the Question of Teaching, from a Corpus Linguistics Perspective. *ICAME Journal ,17*, 97-109.

Hunston, S. & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English.* Amsterdam: John Benjamins.

Johns, T. (1991). Should you be persuaded---Two Examples of Data-Driven Learning Materials. *English Language Research Journal*,4,1-16. University of Birmingham.

Johns, T. (2002), Data-driven Learning: The Perpetual Challenge. In B. Ketteman, & G. Marko (Eds.). *Teaching and learning by doing corpus analysis* (pp. 107-117.). Amsterdam: Rodopi. Retrieved from http://www.ingentaconnect.com/content/rodopi/lang/2002/00000042/00000001/art00009

Kennedy, C., & Miceli, T. (2001). An evaluation of intermediate students' approaches to corpus investigation. *Language*

*Learning and Technology, 5*(3), 77-90. Retrieved from http://llt.msu.edu/vol5num3/kennedy/default.html

Lavid, J. (2007). Contrastive patterns of mental transitivity in English and Spanish: a student-centred corpus-based study. In E. Hidalgo, L. Quereda, & J. Santana (Eds.). *Corpora in the foreign language classroom* (pp. 237-252.). Amsterdam: Rodopi.

Law, N., Yuen, A. & Fox, R. (2011). *Educational innovations beyond technology. Nurturing leadership and establishing learning organizations.* New York: Springer

McEnery, T. & Wilson, A. (1997). Teaching and Language Corpora. ReCALL, *9,* 5-14.

McEnery, A.M. & Wilson, A. (1997). 'Corpora and language teaching', *RECALL Journal,* 9, 1*,* 5-14.

Mukherjee, J. (2004). Bridging the gap between Applied Corpus Linguistics and the reality of English language teaching in Germany. In U. Connor & T.A.Upton (Eds.). *Applied Corpus Linguistics. A Multidimensional Perspective* (pp.239-250). Amsterdam: Rodopi.

Pérez-Paredes & Alcaraz, 2009. Developing annotation solutions for online Data-driven learning. *ReCALL*, 21,1, 55-75.

Pérez-Paredes, P. 2010. Appropriation and integration issues in corpus methods and mainstream language education. In T. Harris & M. Moreno Jaén (Eds.), *Corpus Linguistics in Language Teaching*. Bern: Peter Lang.

Römer, U. (2008). Corpora and language teaching. In A. Lüdeling, & M. Kytö (Eds.). *Corpus Linguistics. An International Handbook (Volume 1)* (pp. 112-130.). Berlin: Mouton de Gruyter.

Sánchez Tornel, M., Alcaraz, J.M., & Pérez-Paredes, P. (2009). Collaborative annotation in implementing corpora for content and language integrated learning web services. Proceedings of the EuroCALL 2009 Conference. *New Trends in CALL: Working Together*. Universidad Politécnica de Valencia, 9-12 September, 2009. Macmillan ELT.

Sinclair, J 1991. *Corpus, Concordance, Collocation.* Oxford: OUP.

Sinclair, J. 2004. *Reading concordances.* London: Longman.

Yoon, H., & Hirvela, A. (2004). ESL student attitudes towards corpus use in L2 writing. *Journal of Second Language Writing*, *13*, 257-283.