THE *CPDB*: A LEARNING AND TEACHING CORPUS-BASED METHODOLOGICAL TOOL

- Laso Martín, Natàlia Judith[1]
  njlaso@ub.edu
- Comelles Pujadas, Elisabet[1]
  elicomelles@ub.edu
- Verdaguer Clavera, Isabel[1]
  i.verdaguer@ub.edu
- Forcadell Guinjoan, Montserrat[1]
  forcadell@ub.edu
- Celaya Villanueva, María Luz[1]
  mluzcelaya@ub.edu

(1) Universitat de Barcelona
Departament de Filologia Anglesa i Alemanya/Facultat de Filologia
Gran Via Corts Catalanes, 585 08007 Barcelona

1. **ABSTRACT:** The *Grup de Recerca en Lexicologia i Lingüística de Corpus* group (GReLiC) at the UB has recently developed a database of English clause patterns: the *Clause Pattern Database (CPDB).* Among the attested pedagogical benefits of the *CPDB* (Comelles *et al.* 2012), it is worth noting that a) it provides users with authentic language, b) it triggers class discussions on the analysis of language behaviour and thus provides a framework for critical reflection and collaborative learning, and c) it promotes the use of new technologies in the linguistics classroom.

2. **RESUM:** El Grup de Recerca en Lexicologia i Lingüística de corpus (GReLiC) de la UB ha desenvolupat recentment una base de dades de patrons sintàctics en anglès: la *Clause Pattern Database (CPDB).* Entre els beneficis pedagògics de la *CPDB* (Comelles *et al*. 2012), cal destacar els següents: a) ofereix als usuaris exemples reals de llengua real, b) promou debats a classe sobre l'anàlisi del comportament de la llengua

anglesa i, per tant, proporciona un marc per a la reflexió crítica i l'aprenentatge col·laboratiu, i c) promou l'ús de les noves tecnologies a l'aula de lingüística.

3. **KEYWORDS:** clause pattern database, corpus-based methodological tools, ESL & EFL teaching and learning, new technologies, collaborative learning. **/ PARAULES CLAU:** base de dades de patrons sintàctics, eines metodològiques basades en corpus, ensenyament-aprenentatge de l'anglès com a segona llengua i llengua estrangera, noves tecnologies, aprenentatge col·laboratiu.

4. **DEVELOPMENT:**

a) Introduction

The use of corpus-based methodological tools and computer technologies (NTICs) has now proliferated in the teaching and learning of second and foreign languages (Granger 2003, Sinclair 2004, Bernardini 2004, Conrad 2005, Laso & Giménez 2007, Granger & Meunier 2008, Aijmer 2009, Comelles *et al*. 2010, Bennet 2010, MacDonald *et al*. 2011). If "corpora have changed our views on language and language use", as Aijmer (2009: 1) puts it, they must also have an obvious effect on second language acquisition.

Corpus studies have provided empirical evidence on the interrelatedness of lexis and syntax, which is a crucial feature in the lexically-oriented approaches to language. Syntactic phenomena are projected from lexical entries and clause patterns depend on the presence of specific lexical verbs. Corpus analysis can thus throw light on the different complementation patterns of polysemous verbs, or show how different syntactic structures can be used to indicate differences in meaning and style.

One of the advantages of the use of corpora often mentioned is that the learner can become language observer and researcher (Johns 1991; Bernardini 2004), and thus be aware of the many complexities of real language in use. By analysing concordance lines, learners can discover the characteristics of the target language. By means of this active process of exploration, the process of language acquisition is also strengthened, as language learners

ISSN: 2385-6203

can be exposed to the target language independently from the teacher, increasing their autonomy. In many cases corpus exploration can also provide the learner with informed answers on aspects not found in grammar or reference books.

It is true, however, that there are still many unanswered questions as regards the use of corpora in the classroom and that, enthusiastic as corpus linguists may be about the effectiveness of the use of corpus in the foreign language classroom, not many teachers are willing to incorporate corpora in the EFL classroom or have the skill or the appropriate information and training to exploit corpora with their students. One problem with the use of corpora in the foreign language classroom is that future language teachers are rarely trained to use corpora, so they lack the skills to use them in their teaching. Consequently, learners are not exposed to corpus analysis and methodology. However, as Grannath (2009: 47) says: "if training in how to use corpora is integrated into university level courses such as syntax, written proficiency and translation, in time it could become just as natural to consult a corpus as to look up an item in a dictionary or grammar book".

Another problem with corpora is that learners can be exposed to too much disorganised material. In a way, this is a positive aspect of corpus linguistics since corpus concordances can give the learner the real fuzzy picture of language in use. However, it may also be useful to organise teaching materials, at least to a certain extent. So, if necessary, learning materials can be carefully controlled either by editing or limiting in some way the concordance lines to be analysed, by selecting examples for activities in the classroom, or by means of corpus-based methodological tools that allow teachers to focus the students' attention on specific linguistic aspects.

Corpora and corpus-based tools can be applied not only to the learning of a language but, as Grannath (2009) recommends, to the analysis of language by university students of linguistics. The present study shows a corpus-based database of English clause patterns which has been used for some years in the Descriptive Grammar classroom of third-year Spanish students of English Studies, as part of a teaching innovation project. In addition to

showing students a methodology of linguistic research which will allow them to gain insights into the language, we also hope to raise students' awareness of the complexity of language and of the benefits of using corpus data for teaching purposes.

b) Methodology: designing the CPDB

With the aim of integrating computer technologies in the undergraduate course Descriptive Grammar of English II (DGE II), the *Grup de Recerca en Lexicologia i Lingüística de Corpus* (GReLic)[1] has developed a database of English clause patterns. This subject is concerned with verb complementation and approaches linguistic analysis from a lexicogrammatical perspective. Thus, a database such as the Clause Pattern Database (CPDB), which displays corpus-based examples, analysed in terms of verb subcategorisation (i.e., phrasal categories and syntactic functions), as well as tree diagrams which illustrate the dependencies established between a lexical verb and its complements, seemed to be a useful tool to supplement course materials.

The GReLic group, interested in teaching innovation research and directly involved in the teaching of courses focused on English Lexicology and Morphology as well as Descriptive Grammar of English (DGE), has compiled the *Whodunnit Corpus*, a collection of 48 best selling mystery novels (of approximately 8 million tokens), which would serve as a methodological tool for the creation of corpus-based teaching materials to be used in the linguistics classroom. This genre was motivated by the fact that DGE students were expected to be familiar with the work by contemporary authors, such as Dan Brown, Michael Crichton, John Grisham and Patricia Cornwell, and thus this could draw their attention more easily. In addition, the use of linguistic data obtained from a corpus (rather than made up examples) is considered to be more illustrative of authentic language in use.

[1] The support of the Spanish Ministerio de Ciencia e Innovación and FEDER (Reference FFI2011-28947) as well as the Programa de Millora i Innovació Docent (References 2011PID-UB/37, 2012PID-UB/133) is acknowledged.

The CPDB has proven to be a very useful corpus-based methodological tool for the accurate description and analysis of lexical verbs. Bearing in mind that verb complementation is one of the main contents dealt with in the course Descriptive Grammar of English II (DGE II), the GReLic group made a selection of the most frequently used verbs in the *Whodunnit Corpus*. The resulting list (217 verbs) was further explored so as to group them according to the five canonical patterns discussed by Huddleston & Pullum (2002); that is, SV, SVO, SVC, SVOO, SVOC. Once explored, a total of 11 patterns (i.e., SVOblique, SVA, SVOOblique, SVOA, SVX-Complement, SVOX-Complement, SVS-Complement, SVAA, SVOS-Complement, SVObliqueOblique, SVObliqueX-Complement) were incorporated into the pre-existing list. Currently, the CPDB consists of 714 corpus-based registers. As can be seen in Fig. 1, each register shows the following fields:

a. Corpus-based example
b. Clause pattern, which can be selected from a scroll-down menu
c. Lemma (i.e., lexical verb)
d. Phrasal categories of the dependencies of the lexical verb

ISSN: 2385-6203

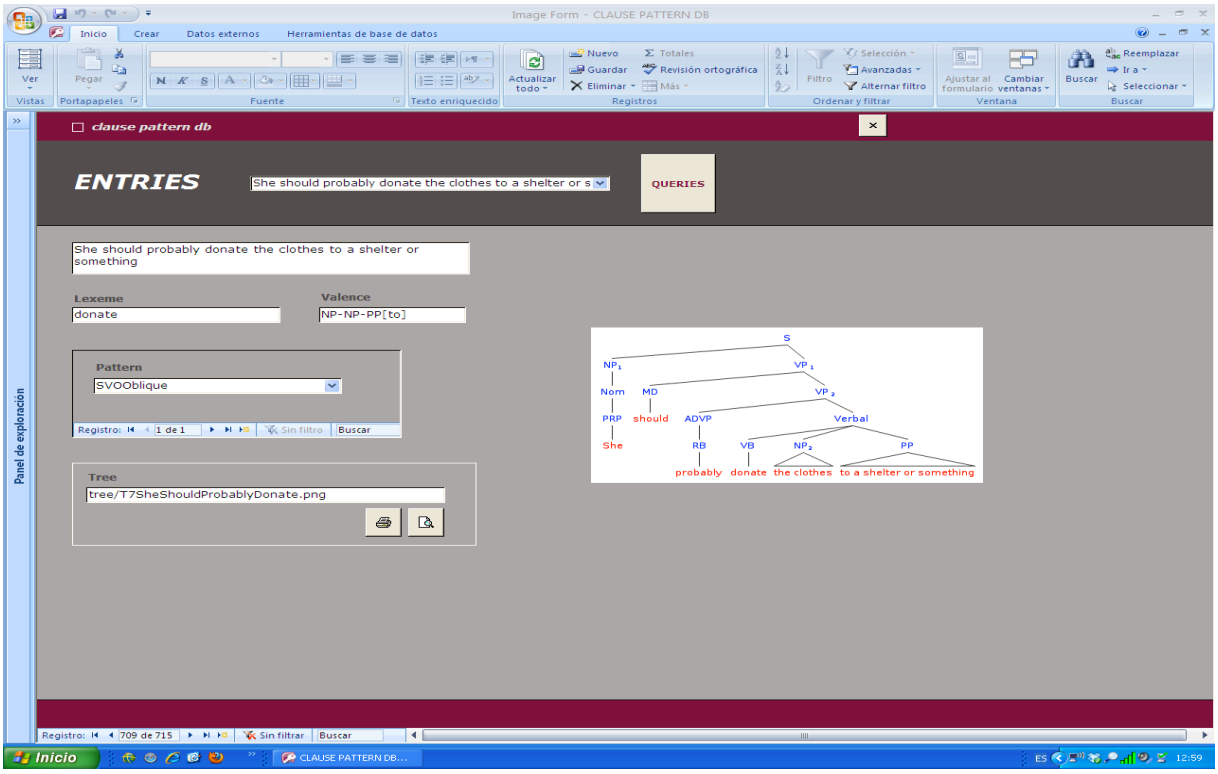

Figure 1: Full entry. Database entry corresponding to the example *She should probably donate the clothes to a shelter or something*.

Once this lexicogrammatical information was introduced in each of the registers, the research group decided to supplement the CPDB with the incorporation of tree diagrams linked to each of the entries in the database. Tree diagramming is also an important issue in the undergraduate course of DGE II, since trees are used as methodological tools that illustrate the dependencies established between a verb and its complements.

At this stage, the group was faced with the further challenge of searching for a) a syntactic parser which would provide us with automatic constituent analyses, and b) an automatic tree diagram-generator, which would translate the syntactic parser analyses into a dependency tree diagram. Several factors, such as the userfriendliness of the tools as well as their publicly available access, were very much taken into account when considering the various tools at our disposal. Finally, we selected the Charniak parser (Charniak & Johnson

2005) for the syntactic parsing and the *phpSyntax Tree* (http://www.ironcreek.net/phpsyntaxtree) as a tree diagram generator.

The Charniak parser uses a regularized MaxEntreranker to elect the best parse from the 50-best parses provided by a generative parsing model and results into automatic bracketed constituent analyses. As a consequence, the automatic parser output had to be revised and adapted to the type of linguistic analysis conducted in our linguistics course. Some linguistic labels, such as 'Nominal' and 'Verbal' were added to the syntactic analyses. The concept of 'Verbal', for instance, was considered to be relevant so as to distinguish complements from adjuncts, whereas the concept of 'Nominal' also played an important role in discriminating both pre and postmodifiers from the head of a noun phrase. Some other worth noting changes relate to the inclusion of phrasal tags. The online version of the Charniak parser only displays part of speech (PoS) tags, but in DGE II syntactic analyses also account for phrasal categories, so they seemed worth including.

Once the editing of automatic parser analyses was agreed on, the resulting bracket constituent analyses were transformed into tree diagrams with the assistance of the phpsyntaxtree tool. This transformation was fairly automatic as it only required copying the parser output (once modified, as described above) into the phrase box available in the phpsyntaxtree interface. The resulting tree diagrams went through a thorough revision process, after which they were saved as .jpg images and later linked to their corresponding register in the database.

c) Clause pattern database: a lexicogrammatical tool

The process described in the previous section has resulted into a lexicogrammatical tool: the *Clause Pattern Database* (*CPDB*). This database was first developed by means of Microsoft Access, but we're currently porting it to an online platform in order to make it available to a larger number of users and grant a more flexible access to the information included. The aims of such a tool are twofold: to allow the addition of new entries, and to perform queries in the same database. Thus, when entering the database, users decide

whether they want to enter information in a new entry, edit an already existing one or make a query in order to retrieve information stored in the database. Those users who want to enter new information and/or edit information in already existing entries must click on "Entries", whereas those who want to make a search have to click on "Queries".

If users opt for adding a new entry, they must open a new register where they will type in the new sentence, the lexeme of the main verb of the sentence and the valence of the verb by stating the phrasal categories of the complements. In addition, in the centre of the screen there is a scroll-down menu from which users will have to select the pattern exemplified in that sentence. Finally, they are also given the opportunity to upload the image of the corresponding tree diagram. The resulting entry will contain, therefore, all the information, as shown in Fig. 1. If users need to edit an already existing entry, they can easily do so by selecting the example from a scroll-down menu which contains all the examples stored in the database. A new screen will be displayed with all the information contained in the entry and users will be allowed to change what they may find necessary.

On the other hand, users who might be interested in searching for information stored in the database will have to click on "Queries" and a new screen will open, allowing them to search by "Lexeme" or "Pattern" (Fig. 2). When searching by "Lexeme", all examples contained in the database are displayed enabling the user to select one of them and decide whether to look for the several patterns linked to this lexeme or its valence. When searching for the patterns performed by a specific lexeme, a list of the patterns and their corresponding examples are displayed in the database, as exemplified in Fig. 3, where all patterns and corresponding sentences to the lexeme *get* are shown. Alternatively, only a list of the patterns which belong to the selected lexeme can be obtained by clicking on the "List" button (see Fig. 4).

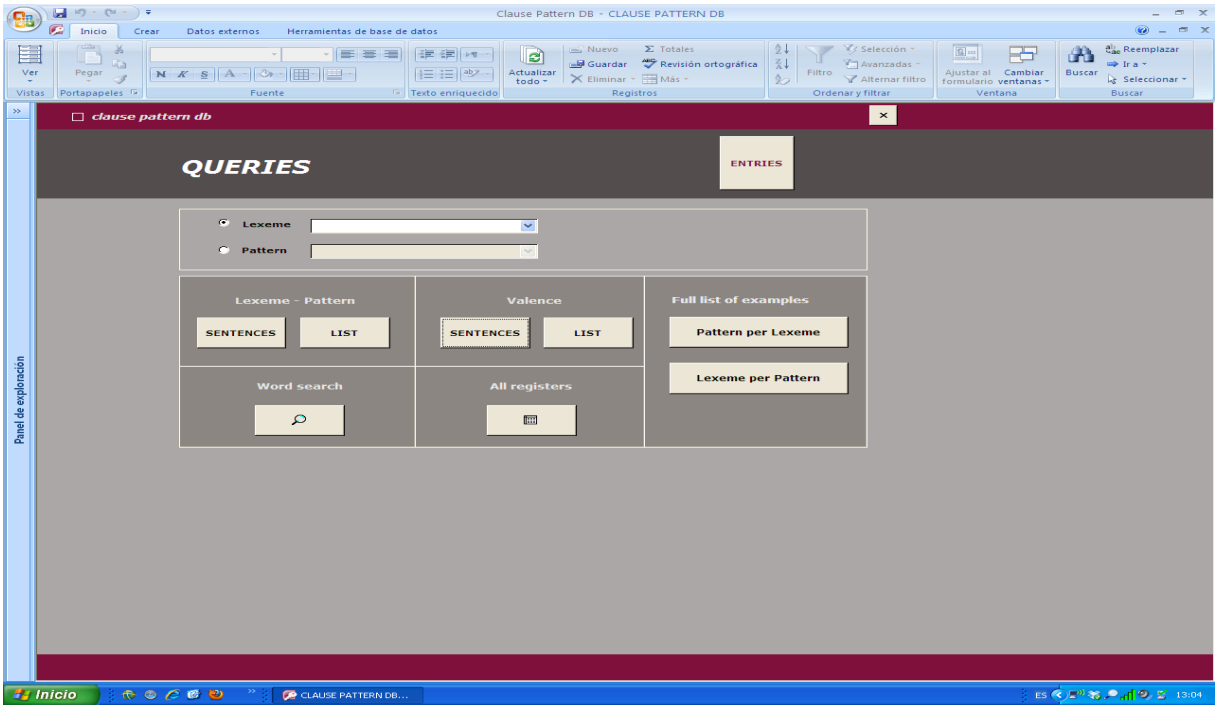

Figure 2: Queries. Screen displayed when clicking on "Queries", which shows the different type of queries available in the database.

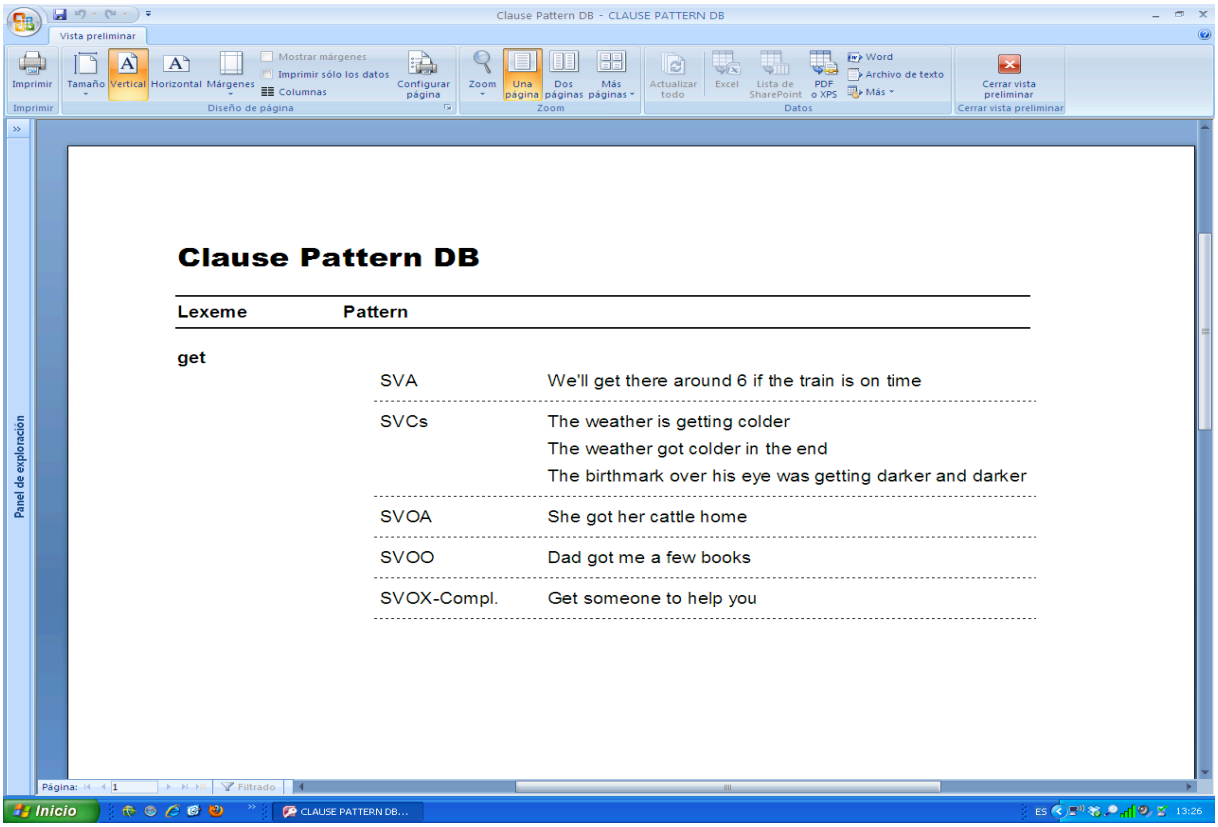

Figure 3: *get* and its corresponding patterns and examples. List of patterns and their corresponding sentences that is displayed when searching for the information about the lexeme *get*.

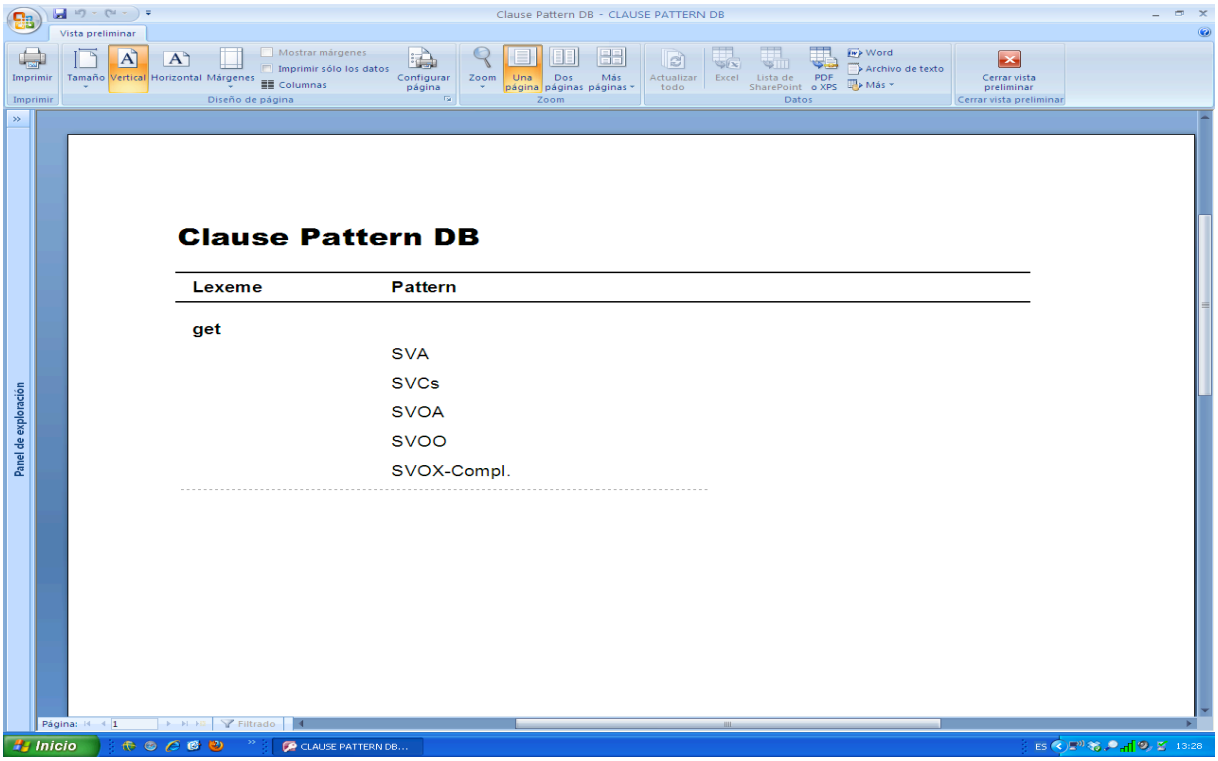

Figure 4: *get* and its patterns. List of patterns performed by the lexeme *get*.

Likewise, if the focus of users' research is the valence of a specific verb, similar queries can be performed in order to obtain a list of the corresponding valences and sentences (Fig. 5) or only a list of the corresponding valences, disregarding the sentences linked to each of them (Fig. 6).

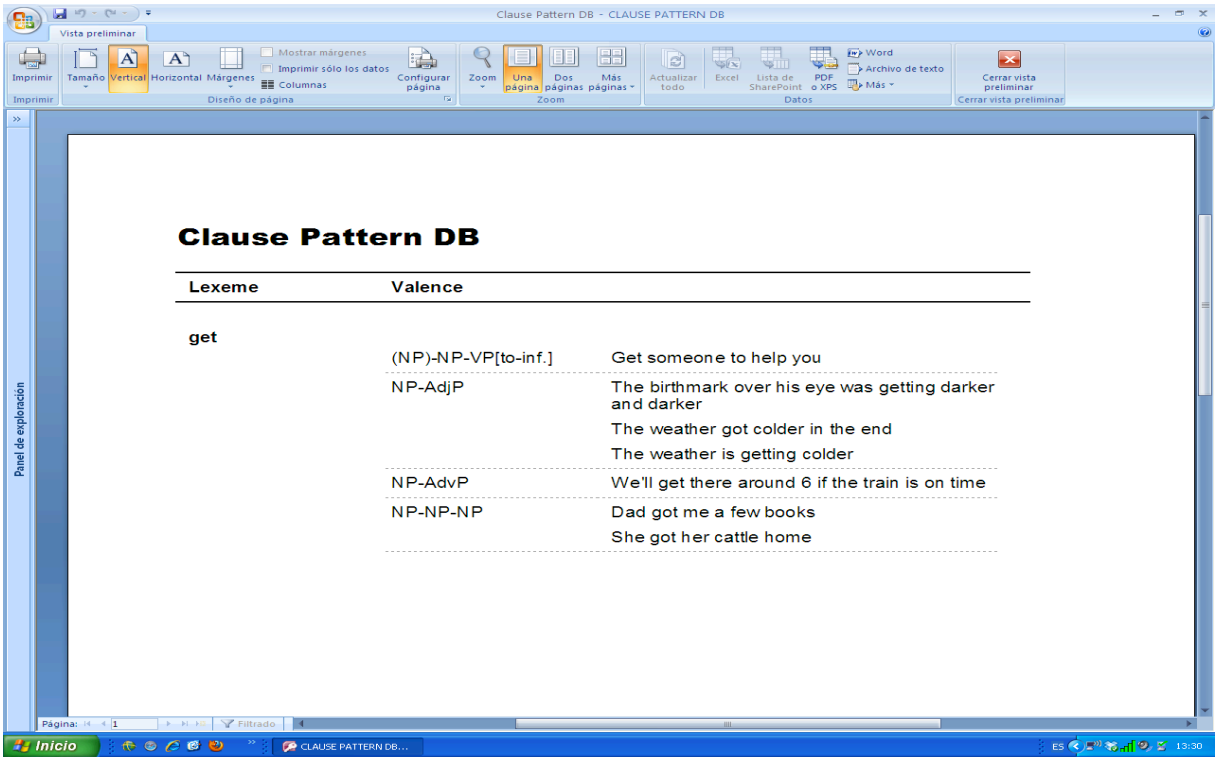

Figure 5: *get* and its corresponding valences and examples. List of valences and their corresponding examples when searching for the lexeme *get*.

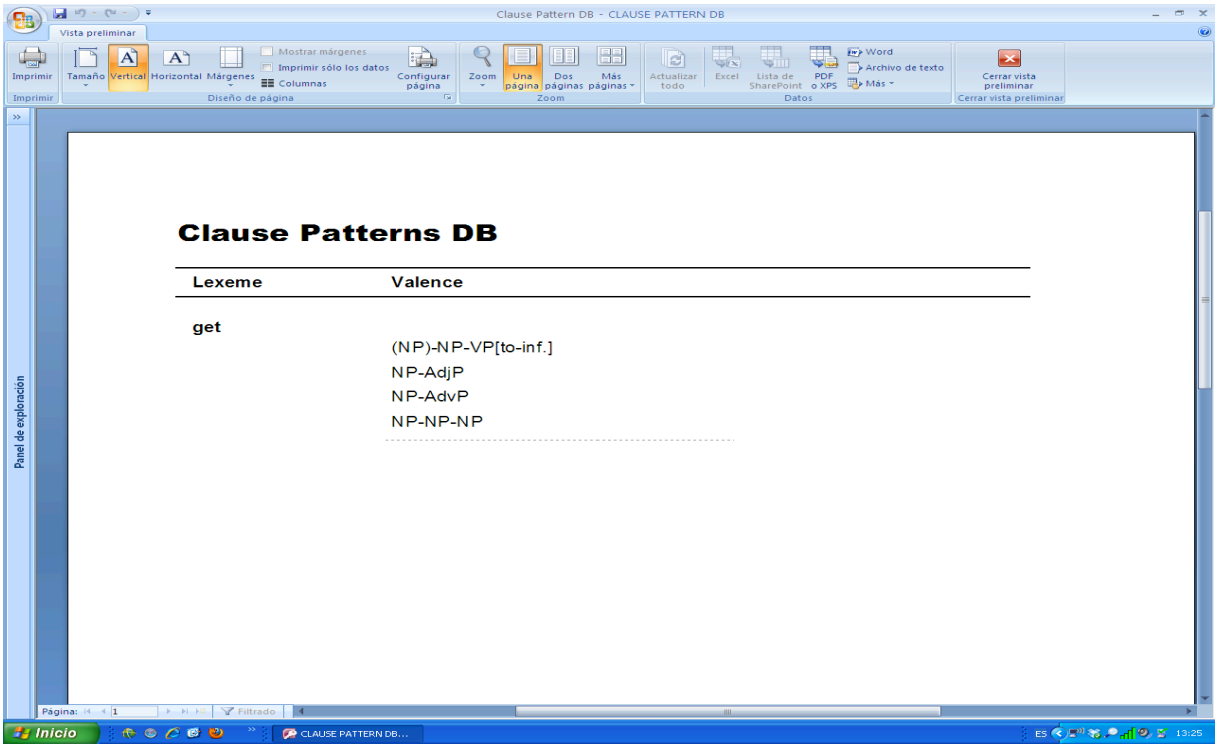

Figure 6: *get* and its corresponding valences. List of possible valences retrieved from the database when searching for the lexeme *get*.

Finally, users can also obtain a full list of all the lexemes stored in the database in different formats: a list organised by patterns and the lexemes performing them or a list by lexemes and their corresponding patterns.

d) Conclusion

The study presented in this paper explores the use of a corpus-based tool (CPDB) in a linguistics class within the framework of an innovation project in the Degree of English Studies (University of Barcelona). After a brief revision of the advantages and disadvantages of using corpora in the classroom, this paper has explained how and why our tool was designed and the way it works and is used by teachers.

As both designers of the tools presented here and teachers of the subjects where these tools are directly applied to the teaching of the subject contents, the researchers have first-hand

experience of the effectiveness of resorting to the use of such tools in a university teaching context. More specifically, our project has allowed us to realise that having the sentences linked to their specific tree diagrams helps the teacher explain dependency structure much clearly, since our experience tells us that bracketing does not suffice for a correct labeling of sentence structure. Students need to be able to discriminate between obligatory and non-obligatory constituents, and this is nicely shown by tree layout. Thus, the CPDB has proved effective in helping students (non-native speakers of English) understand both the theoretical and the analytical perspective of the topic of the subject (i.e., the English sentence and its patterns).

Besides this clear pedagogical application, the benefits of the CPDB have also been proved at the methodological level: collaborative work, both between peers and students and teacher, has been enhanced; this is rather an infrequent way of working in large university classes, where sessions are usually teacher-fronted. The flexibility of the CPDB allowing both to be enlarged and to receive queries from both teachers and students makes it quite a powerful tool not only to actively teach the contents involving the direct participation of the students but also to be used as a helpful source of data for the answering of questions on language use and also for the creation of teaching materials that, at some point, the instructor might want to tailor to specific needs.

Thus, developing a database of clause patterns has greatly contributed to the automatisation of teaching and learning activities related to the specific contents on the syllabus, since it has helped to systematise and organise the information provided by a self-compiled corpus. It has also promoted collaborative work between instructors when discussing and interpreting corpus-based examples. Both outcomes of our project have encouraged us to further investigate the use of NTICs in the university classroom by incorporating new information to be applied in other related subject classrooms.

5. REFERENCES

[1] Aijmer, K. (Ed.) (2009). Corpora and language teaching, Amsterdam/Philadelphia, John Benjamins.

[2] Bennet, G. R. (2010). Using CORPORA in the language learning classroom: Corpus linguistics for teachers. University of Michigan, Michigan.

[3] Bernardini, S. (2004). Corpora in the classroom. An overview and some reflections on future development, in: Sinclair, J. M. (Ed.), How to use corpora in language teaching. John Benjamins, Amsterdam, pp. 15-38.

[4] Charniak, E., Johnson, M. (2005). Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. Proceedings of the 43rd annual meeting of the association for computational linguistics. Association for Computational Linguistics, Michigan (USA), pp. 173-180.

[5] Comelles, E., Laso, N. J., Verdaguer, I., Gimenez, E. (2010). Clause pattern DB: a corpus-based tool, in: Moskowich-Spiegel Fandiño, I., Crespo García, B., Lareo Martín, I., Lojo Sandino, P. (Eds.), Language windowing through corpora. Visualización del lenguaje a través de corpus. Universidad da Coruña, Coruña, pp. 215-234.

[6] Conrad, S. (2005). Corpus linguistics and L2 teaching, in: Hinkel, E. (Ed.), Handbook of research in second language teaching and learning. Lawrence Erlbaum Associates, New York, pp. 393-409.

[7] Grannath, S. (2009). Who benefits from learning how to use corpora, in Aijmer, K. (Ed.) Corpora and language teaching. Amsterdam/Philadelphia, John Benjamins, pp. 47-65.

[8] Granger, S. (2003). The International Corpus of Learner English: A new resource for foreign language learning and teaching and second language acquisition research. TESOL Quarterly 37, 538-46.

[9] Granger, S., Meunier, F. (Eds.) (2008). Phraseology in foreign language learning and teaching. John Benjamins, Amsterdam/Philadelphia.

[10] Huddleston, R. & Pullum, G.K. (2002). The Cambridge Grammar of the English Language. Cambridge: CUP.

[11] Johns, T. (1991). Should you be persuaded: Two examples of data-driven learning, in Johns, T. & King, P. (Eds). Classroom Concordancing. Birmingham, CELS University of Birmingham, pp. 1-16.

[12] Laso, N. J., Giménez, E. (2007). Bridging the gap between corpus research and language teaching, in: Periñán, C. (Ed.), Revisiting language learning resources. Cambridge Scholars Publishing, Newcastle, pp. 49-64.

[13] MacDonald, P., Murcia, S., Boquera, M., Botella, A., Cardona, L., García, R., Mediero, E., O'Donnell, M., Robles, A., Stuart, K. (2011). Error coding in the TREACLE project, in: Carrió Pastor, M. L., Candel Mora, M. A. (Eds.), Actas del 3 congreso internacional de lingüística de corpus, tecnologías de la información y las comunicaciones. Universidad Politécnica de Valencia, Valencia, pp.725-740.

[14] phpSyntaxTree http://www.ironcreek.net/phpsyntaxtree/ (last visited: 05 May 2013).

[15] Sinclair, J. M. (Ed.) (2004). How to use corpora in language teaching. John Benjamins, Amsterdam.