IAC: A dynamic corpus interface

Toni Badia, Carme Colominas

Universitat Pompeu Fabra

Judith Domingo

Barcelona Media

Abstract: Corpora in translation studies are essential not only for research but for training as well. Interfaces for accessing corpora are often not user-friendly enough to satisfy the real needs of translation students and researchers. Moreover, interfaces differ from each other in the layout and type of searches that they allow for. Conscious of these limitations, we have developed IAC (Corpora Access Interface), a non-dependant corpus interface for monolingual and parallel corpora that allows users, without programming grounding, to create searching interfaces between a given corpus and the underlying search tools1. IAC includes also user-controlled access that allows to distinguish between private and public corpora. Once the corpus is uploaded and the interface is created, IAC indexes the corpus and a user-friendly searching interface is automatically created allowing for 3 types of searches: simple, expert and frequency-based.

1. Motivation

As it becomes evident through the significant number of publications and conferences devoted to the field of corpus-based contrastive and translation studies during the last decade, analyses/studies/investigation are conducted in a wide range of subfields and with many different aims. Depending on the research aim a given corpus type can be more relevant than another. The question of which type of corpus is suitable for which type of study has been discussed in many papers, but following Johansson (1998), we can claim that most of the research done in contrastive linguistic and translation studies requires parallel corpora and some other:

1. multilingual corpora of original texts and their translations (for contrastive studies and translation studies)
2. multilingual corpora of original texts which are matched by criteria such as genre, time and composition etc. (for contrastive studies)
3. monolingual corpora consisting of original and translated texts (for translation studies)

Besides research activities, corpora have been increasingly involved in learning activities. For this so called data-driving learning, besides monolingual and bilingual corpora also learner corpora can be involved and integrated into foreign language and translation teaching. In the context of second language acquisition, learner corpora are used to compare native and non-native varieties of the same language (Granger 1998b: 47). For its part, in the context of translation, Learner Corpora, which are composed of translations produced by trainee translators annotated with errors[2], aim to inform translation training.

With regard to the availability of all these corpora types, nowadays we can claim that "the limited availability of corpora for cross-linguistic research" referred by Granger (2003) is fortunately much minor: many studies are actually carried in other languages than English, as large and representative corpora are available for the major European languages. Besides the British National Corpus[3] (100 million words), we have, for example, the IDS (Institut für Deutsche Sprache) corpus[4] for German with 1 billion words, the CREA[5] (Corpus de Referencia del Español Actual) for Spanish with over 200 million words and the CORIS/CODIS[6] (Dynamic Corpus or Written Italian) for Italian with 100 million words.

In a parallel way, in recent years the arduous and expensive task of building large corpora has found in the world wide web as a source of linguistic data (Kilgarriff and Grefenstette, 2003) real new chances. Exploiting the web as a corpus is becoming a real alternative to the traditional building of large corpora, as can be checked for instance by the Internet corpora compiled at the Centre for Translation Studies of Leeds (Sharoff, 2006), the OPUS collection of parallel corpora, or the CUCWeb project developed by the GLiCom (Grup de Lingüística Computacional, UPF). Besides academic work, the potentiality of the web as a corpus is also being exploited by independent initiatives with remarkable results for the professional translators like by the Linguee project[7].

But apart from the question of the corpus availability, researchers need to access the compiled data, and depending on the very object of investigation a simple concordancer will be enough or otherwise more sophisticated annotation and retrieval software tools will be needed. For example, if the research focuses on particular forms (preposition, conjunctions…) or on keywords, a tool such WordSmithTools, which makes it possible to extract all the occurrences of a given form and to visualize them in context will be enough. Otherwise, if the research relies, for example, on data like frequencies over lemmata or POS sequences, the corresponding corpus needs to be lemmatised and linguistically annotated. Furthermore, annotated corpora have to be searched through an interface that normally has been built according to the annotation type used. The multiplicity of interfaces derived from this fact carries enorm economic as well as qualitative costs. On the one side, new funds have to be raised to build new interfaces, but on the other, the user has to be faced with several interfaces and query languages depending on the corpus. Interfaces differ not only on their layout, but on the types of queries they allow for, and this even affects the exploitation possibilities and especially those that imply comparing the results obtained from several corpora. In order to get familiar with different interfaces and query languages and, what is worse, to face the differences in creating concordances (by form, lemma or part of speech (POS), in gathering statistical information, etc, between corpora, the user needs to spent much time and efforts. As a result, the usefulness of resources, even when they exist, becomes far form evident, especially by users that are not trained in query formalisms as is often the case in the context of translation. This is often a crucial point in the contrastive and translation studies field due to the fact that most researchers and trainers in the field come actually from the humanities.

The problems derived from the multiplicity of interfaces and query languages can be solved by the creation of uniform and user-friendly interfaces that allow the location of very different corpora, such as IAC, which will be presented in the following sections.

2. Corpus Query Methods

As described in the previous section, there is a need of homogenizing the access to corpora as systems like databases, concordancers or corpus query systems (like CQP or EMDROS) require a long learning process and they cannot be used with any corpus.

There have been initiatives like CQP Web, Sketch Engine or Jaguar that have offered tools to search in corpus through friendly platforms.

CQP Web is a web-based corpus analysis system that is compatible with any corpus, but is especially useful for large corpora, corpora with word-level annotation (such as part-of-speech tagging), and corpora with rich text-level metadata. The purpose of this tool is not to allow users to upload their own corpora but to offer a platform for any corpora. CQPWeb only offers simple searches that can be restricted by metadata. It also offers quantitative results that can be presented by lemma or pos (depending on corpus annotations) and incorporates other useful functions, such as a query history and a subcorpora creator in order to create your own piece of corpus and a user control that restricts the access to the corpora.

Another corpus tool is Sketch Engine that takes as input a corpus of any language and the corresponding grammar patterns and generates word sketches for the words of that language. It offers different search types, a standard search called simple where the query matches the lemma and the word introduced for searching or searches for lemma or word combined with part of speech for key words out of context. It also offers a query based on Corpus Query Language inputting complex queries that requires some training from the user. Moreover, SketchEngine also allows for frequency distribution and also generates a thesaurus and ‘sketch differences’, which specify similarities and differences between near-synonyms.

Finally, Jaguar is a software that offers basic administration for corpora and corpora extraction from the web developed by Rogelio Nazar at IULA (Institute for Applied Linguistics). There are two ways to introduce a corpus: one is to update user files (plain text, html, xml, pdf, doc or ps) and Jaguar automatically converts them to the adequate format and another option is to build a corpus from the web. This procedure works with key words introduced by the user, and other parameters like language and documents format. Once the corpus is uploaded, it is possible to get concordances and n-grams ordered by frequency. The search interface is limited to forms or regular expressions in order to find pattern concordances or pieces of words. In some cases, if the corpus is lemmatized and tagged with IULA part-of-speech tagset, more complex searches with these attributes can be made. The resulting lists can be reordered to identify interesting relations between lexical units and it calculates distribution measures (t-test, chi-square, and similarity coefficients, mutual information and mutual information cubed).

As we have seen, some efforts have been made to develop general interfaces for corpus although there are still some shortcomings: i.e. limitation to monolingual corpora, tough configuration, limitations with attributes (such as pos tagset, etc.). Aware of this situation, we have developed IAC trying to contribute to corpus query tools.

3. IAC

IAC is a multilingual tool to create interfaces for corpus without previous knowledge of programming. IAC tries to fill a gap under corpus researchers as they usually need external support (design and programming) to build an interface for their corpora. As researchers are not autonomous in building their own interfaces, more budget is needed and delays in the research are frequent to build interfaces. Moreover, IAC also goes beyond simple concordancers as it allows for complex searches and works with monolingual and aligned corpora facilitating a versatile tool to search in corpus.

Not only IAC is useful for researchers but also is interesting for instructors as from a single interface they offer access to all their corpora to students avoiding to waste time training them in the multiplicity of interfaces.

As a summary, IAC integrates:

- A tool to upload new corpus without programming knowledge
- A platform to host the corpora
- A platform to search in corpus, offering textual and quantitative results

See next sections for detailed examples of IAC.

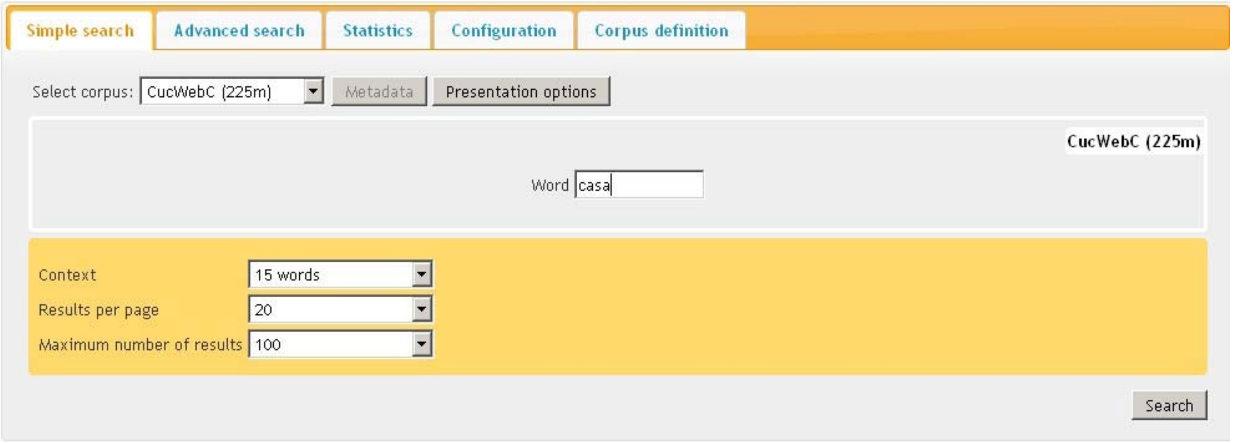
3.1. Monolingual corpus

Each monolingual corpus uploaded in IAC has 3 interfaces for queries: simple, advanced and statistics.

3.1.1 Simple search

IAC simple searches allow for key words out of context. Depending on corpus annotations, searches can be restricted by lemma, pos or other attributes.

To show the simple search, we use a Catalan annotated corpus, called CucWeb, of 208 million of words.



Select corpus: CucWebC (225m)

Word casa

Context 15 words

Results per page 20

Maximum number of results 100

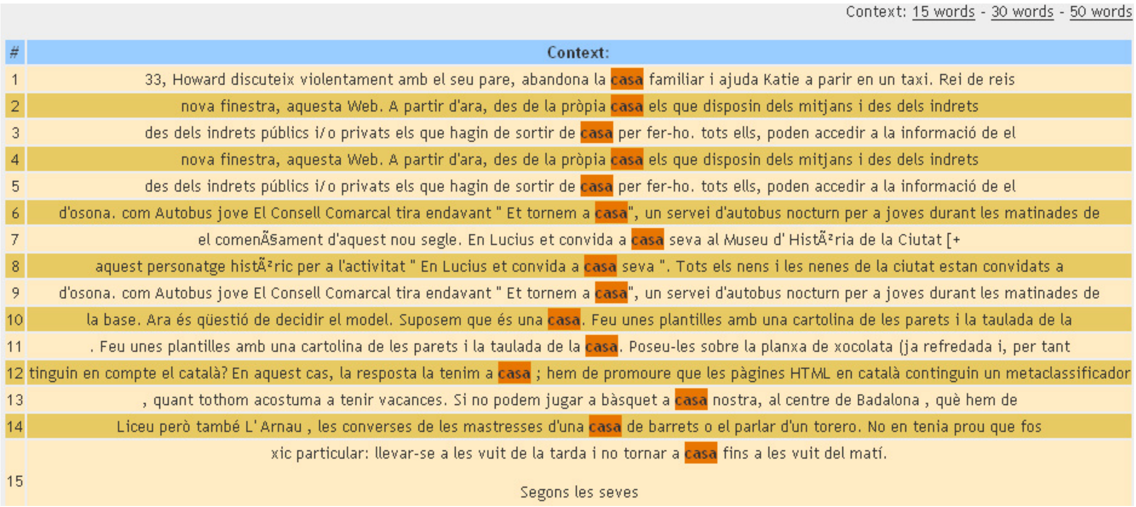
Figure 1. Simple search for word *casa*

CucWeb's simple search interface allows to introduce ONE word by form. Moreover you can customize the appearance of the results page:

Context: sentence / 15 words / 30 words / 50 words

Results per page: 20 / 50 / 100

Maximum number of results: 100 / 500 / 1000 / 10000 As we can see in fig. 1, we search for the word "casa" (*house*) with context *15 words* context, 20 results per page and a maximum number of results (the bigger, the longer processing time)



Context: 15 words - 30 words - 50 words

#	Context:
1	33, Howard discuteix violentament amb el seu pare, abandona la casa familiar i ajuda Katie a parir en un taxi. Rei de reis
2	nova finestra, aquesta Web. A partir d'ara, des de la pròpia casa els que disposin dels mitjans i des dels indrets
3	des dels indrets públics i/o privats els que hagin de sortir de casa per fer-ho. tots ells, poden accedir a la informació de el
4	nova finestra, aquesta Web. A partir d'ara, des de la pròpia casa els que disposin dels mitjans i des dels indrets
5	des dels indrets públics i/o privats els que hagin de sortir de casa per fer-ho. tots ells, poden accedir a la informació de el
6	d'osona. com Autobus jove El Consell Comarcal tira endavant " Et tornem a casa", un servei d'autobus nocturn per a joves durant les matinades de
7	el comenÃ§ament d'aquest nou segle. En Lucius et convida a casa seva al Museu d' HistÃ²ria de la Ciutat [+
8	aquest personatge histÃ²ric per a l'activitat " En Lucius et convida a casa seva ". Tots els nens i les nenes de la ciutat estan convidats a
9	d'osona. com Autobus jove El Consell Comarcal tira endavant " Et tornem a casa", un servei d'autobus nocturn per a joves durant les matinades de
10	la base. Ara és qüestió de decidir el model. Suposem que és una casa. Feu unes plantilles amb una cartolina de les parets i la taulada de la
11	. Feu unes plantilles amb una cartolina de les parets i la taulada de la casa. Poseu-les sobre la planxa de xocolata (ja refredada i, per tant
12	tinguin en compte el català? En aquest cas, la resposta la tenim a casa ; hem de promoure que les pàgines HTML en català continguin un metaclassificador
13	, quant tothom acostuma a tenir vacances. Si no podem jugar a bàsquet a casa nostra, al centre de Badalona , què hem de
14	Liceu però també L' Arnau , les converses de les mastresses d'una casa de barrets o el parlar d'un torero. No en tenia prou que fos
15	xic particular: llevar-se a les vuit de la tarda i no tornar a casa fins a les vuit del matí. Segons les seves

Figure 2. Results for simple search

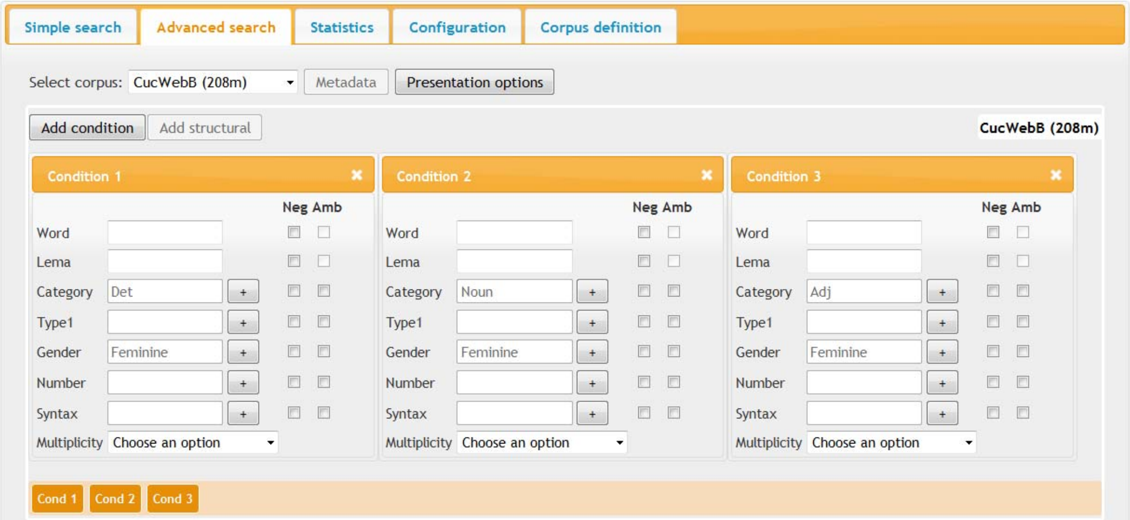
With the simple search, we obtain textual results that can be downloaded in a text file for further processing.

3.1.2 Advanced search

The main difference between simple and advanced search is that the advanced search allows for key words in context. This interface is useful to get examples of collocations. Moreover, if the corpus is annotated with linguistic information, each element of the search can be restricted by these attributes.

Let's see an example, in the Cucweb corpus:

Let's imagine we want to search a feminine noun phrase formed by a determiner, a noun and an adjective, we define the search as 3 conditions (see fig. 4): [determiner feminine] + [noun feminine] + [adjective feminine] and we get textual results that match with our search (See fig. 5).

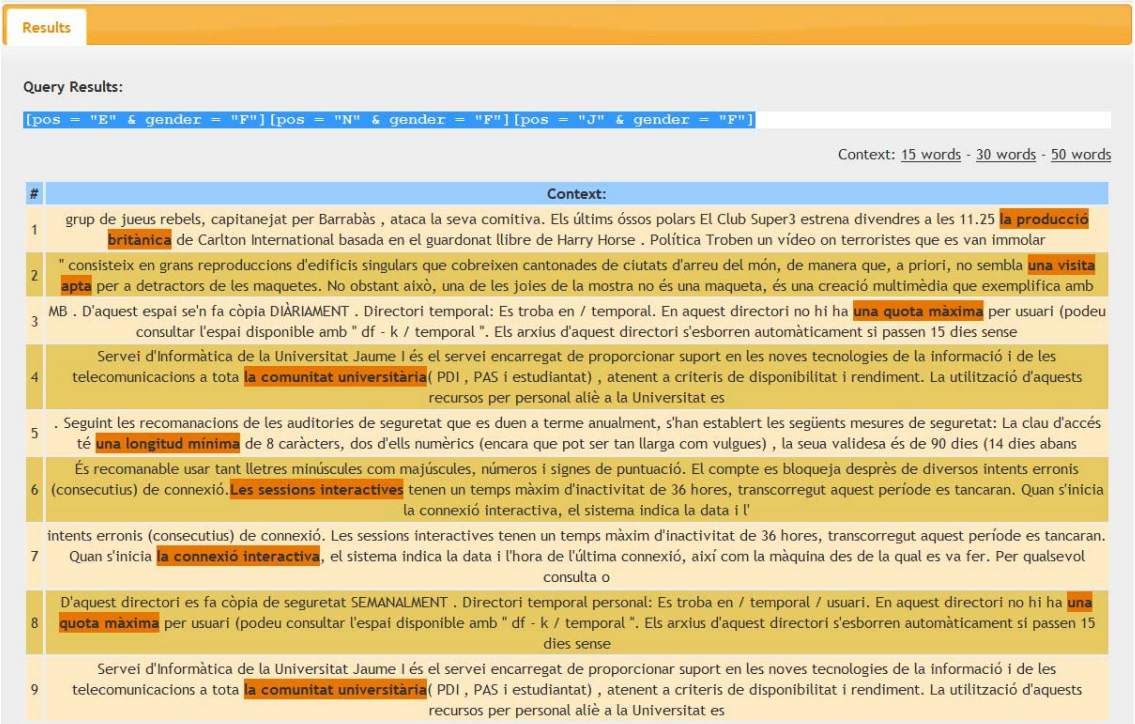


Select corpus: CucWebB (208m)

CucWebB (208m)

	Condition 1	Condition 2	Condition 3
Word			
Lema			
Category	Det	Noun	Adj
Type1			
Gender	Feminine	Feminine	Feminine
Number			
Syntax			
Multiplicity	Choose an option	Choose an option	Choose an option

Figure 4. Advanced search of feminine noun phrases



Query Results:

```
[pos = "E" & gender = "F"][pos = "N" & gender = "F"][pos = "J" & gender = "F"]
```

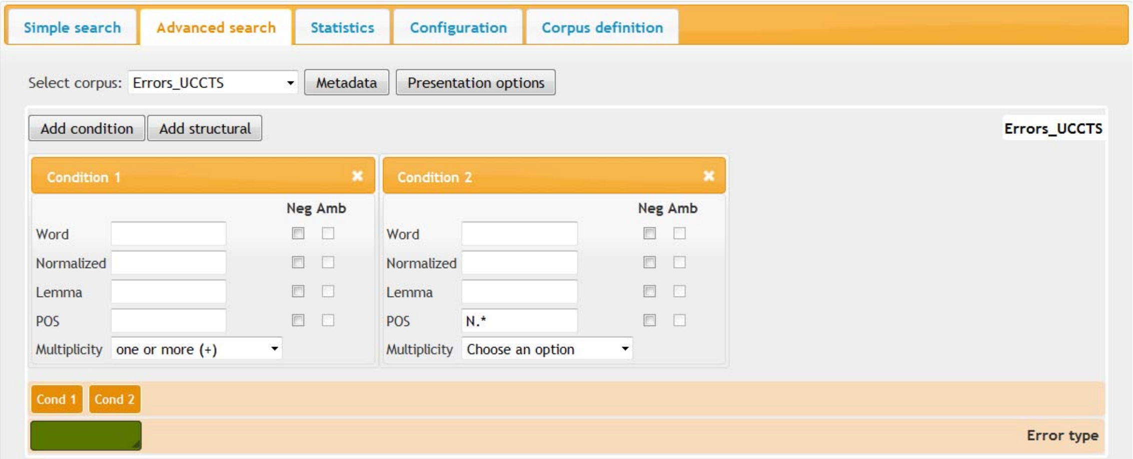
Context: 15 words - 30 words - 50 words

#	Context:
1	grup de jueus rebels, capitanejat per Barrabàs , ataca la seva comitiva. Els últims óssos polars El Club Super3 estrena divendres a les 11.25 **la producció britànica** de Carlton International basada en el guardonat llibre de Harry Horse . Política Troben un vídeo on terroristes que es van immolar
2	" consisteix en grans reproduccions d'edificis singulars que cobreixen cantonades de ciutats d'arreu del món, de manera que, a priori, no sembla **una visita apta** per a detractors de les maquetes. No obstant això, una de les joies de la mostra no és una maqueta, és una creació multimèdia que exemplifica amb
3	MB . D'aquest espai se'n fa còpia DIÀRIAMENT . Directori temporal: Es troba en / temporal. En aquest directori no hi ha **una quota màxima** per usuari (podeu consultar l'espai disponible amb " df - k / temporal ". Els arxius d'aquest directori s'esborren automàticament si passen 15 dies sense
4	Servei d'Informàtica de la Universitat Jaume I és el servei encarregat de proporcionar suport en les noves tecnologies de la informació i de les telecomunicacions a tota **la comunitat universitària**(PDI , PAS i estudiantat) , atenent a criteris de disponibilitat i rendiment. La utilització d'aquests recursos per personal aliè a la Universitat es
5	. Seguint les recomanacions de les auditories de seguretat que es duen a terme anualment, s'han establert les següents mesures de seguretat: La clau d'accés té **una longitud mínima** de 8 caràcters, dos d'ells numèrics (encara que pot ser tan llarga com vulgues) , la seua validesa és de 90 dies (14 dies abans
6	És recomanable usar tant lletres minúscules com majúscules, números i signes de puntuació. El compte es bloqueja desprès de diversos intents erronis (consecutius) de connexió.**Les sessions interactives** tenen un temps màxim d'inactivitat de 36 hores, transcorregut aquest període es tancaran. Quan s'inicia la connexió interactiva, el sistema indica la data i l'
7	intents erronis (consecutius) de connexió. Les sessions interactives tenen un temps màxim d'inactivitat de 36 hores, transcorregut aquest període es tancaran. Quan s'inicia **la connexió interactiva**, el sistema indica la data i l'hora de l'última connexió, així com la màquina des de la qual es va fer. Per qualsevol consulta o
8	D'aquest directori es fa còpia de seguretat SEMANALMENT . Directori temporal personal: Es troba en / temporal / usuari. En aquest directori no hi ha **una quota màxima** per usuari (podeu consultar l'espai disponible amb " df - k / temporal ". Els arxius d'aquest directori s'esborren automàticament si passen 15 dies sense
9	Servei d'Informàtica de la Universitat Jaume I és el servei encarregat de proporcionar suport en les noves tecnologies de la informació i de les telecomunicacions a tota **la comunitat universitària**(PDI , PAS i estudiantat) , atenent a criteris de disponibilitat i rendiment. La utilització d'aquests recursos per personal aliè a la Universitat es

Figure 5. Results of advanced search of feminine noun phrases

If our corpus has annotations at phrase level (such as errors in a learner corpus or syntactic information), we can also combine restrictions at word level and at phrase level. For example (See fig. 6), if we want to find errors in a learner corpus formed by one or more elements plus a noun (Results page can be customized with the same values as simple search).

We just need two conditions and a phrase attribute for our search.



Select corpus: Errors_UCCTS

Condition 1

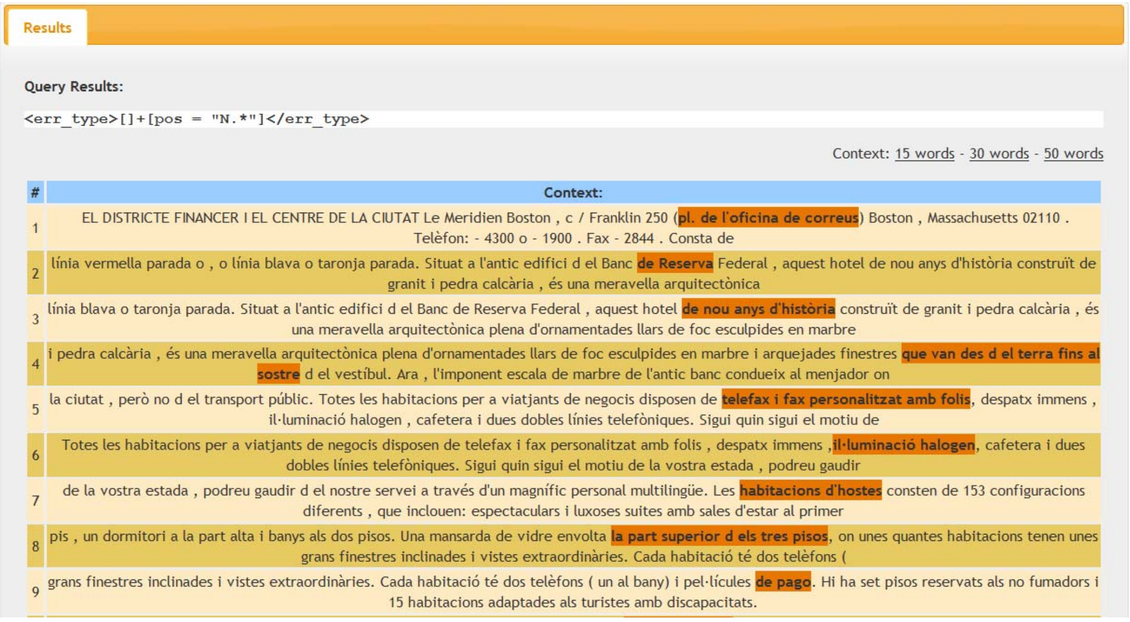
		Neg	Amb
Word		□	□
Normalized		□	□
Lemma		□	□
POS		□	□
Multiplicity	one or more (+)		

Condition 2

		Neg	Amb
Word		□	□
Normalized		□	□
Lemma		□	□
POS	N.*	□	□
Multiplicity	Choose an option		

Figure 6. Advanced search with attributes at phrase level

We also get textual results (see fig. 7).



Query Results:

```
<err_type>[]+[pos = "N.*"]</err_type>
```

Context: 15 words - 30 words - 50 words

#	Context:
1	EL DISTRICTE FINANCER I EL CENTRE DE LA CIUTAT Le Meridien Boston , c / Franklin 250 (**pl. de l'oficina de correus**) Boston , Massachusetts 02110 . Telèfon: - 4300 o - 1900 . Fax - 2844 . Consta de
2	línia vermella parada o , o línia blava o taronja parada. Situat a l'antic edifici d el Banc **de Reserva** Federal , aquest hotel de nou anys d'història construït de granit i pedra calcària , és una meravella arquitectònica
3	línia blava o taronja parada. Situat a l'antic edifici d el Banc de Reserva Federal , aquest hotel **de nou anys d'història** construït de granit i pedra calcària , és una meravella arquitectònica plena d'ornamentades llars de foc esculpides en marbre
4	i pedra calcària , és una meravella arquitectònica plena d'ornamentades llars de foc esculpides en marbre i arquejades finestres **que van des d el terra fins al sostre** d el vestíbul. Ara , l'imponent escala de marbre de l'antic banc condueix al menjador on
5	la ciutat , però no d el transport públic. Totes les habitacions per a viatjants de negocis disposen de **telefax i fax personalitzat amb folis**, despatx immens , il·luminació halogen , cafetera i dues dobles línies telefòniques. Sigui quin sigui el motiu de
6	Totes les habitacions per a viatjants de negocis disposen de telefax i fax personalitzat amb folis , despatx immens ,**il·luminació halogen**, cafetera i dues dobles línies telefòniques. Sigui quin sigui el motiu de la vostra estada , podreu gaudir
7	de la vostra estada , podreu gaudir d el nostre servei a través d'un magnífic personal multilingüe. Les **habitacions d'hostes** consten de 153 configuracions diferents , que inclouen: espectaculars i luxoses suites amb sales d'estar al primer
8	pis , un dormitori a la part alta i banys als dos pisos. Una mansarda de vidre envolta **la part superior d els tres pisos**, on unes quantes habitacions tenen unes grans finestres inclinades i vistes extraordinàries. Cada habitació té dos telèfons (
9	grans finestres inclinades i vistes extraordinàries. Cada habitació té dos telèfons (un al bany) i pel·lícules **de pago**. Hi ha set pisos reservats als no fumadors i 15 habitacions adaptades als turistes amb discapacitats.

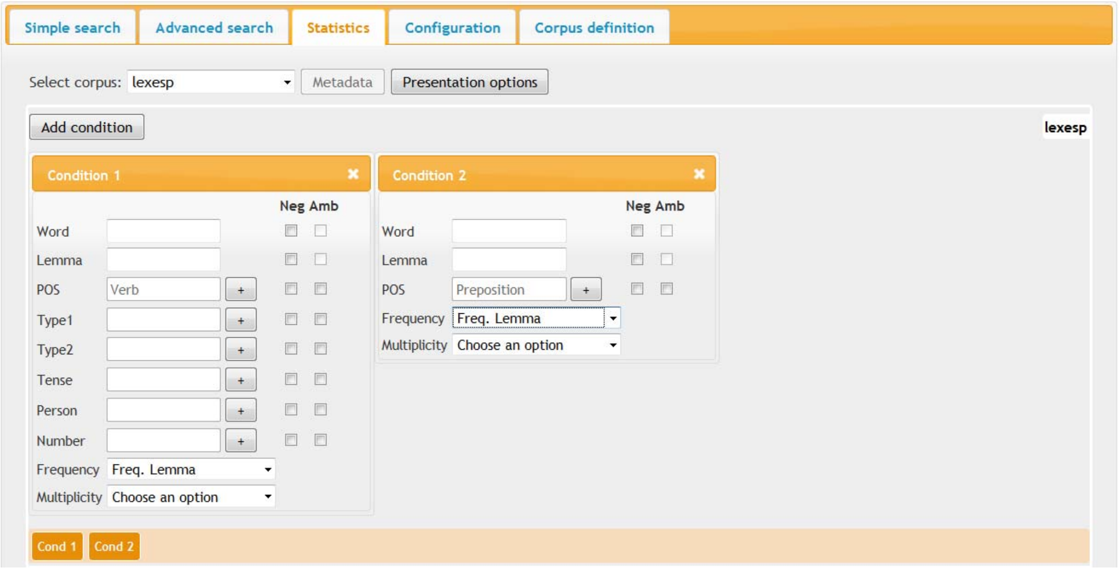
Figure 7. Results of advanced search with attributes at phrase level

3.1.3 Statistics

Statistics interface has the same appearance as advance search although only searches restricted at word level are allowed. The main difference is that the results obtained are quantitative that can be sorted by all the attributes of the corpus. Getting quantitative results is useful for collocation or pattern extraction depending on how results are sorted, i.e. pos, lemma, etc. Moreover, if textual results are needed, they can be obtained from the statistical results page.

As an example, we want to get quantitative results of the prepositions that follow the verb *pensar (to think)* in Spanish in order to get the prepositional subcategorization pattern.

For our search (see fig. 8), we define a verb in condition 1 followed by a preposition in condition 2 and we apply also for results grouped by lemma.



Select corpus: lexesp

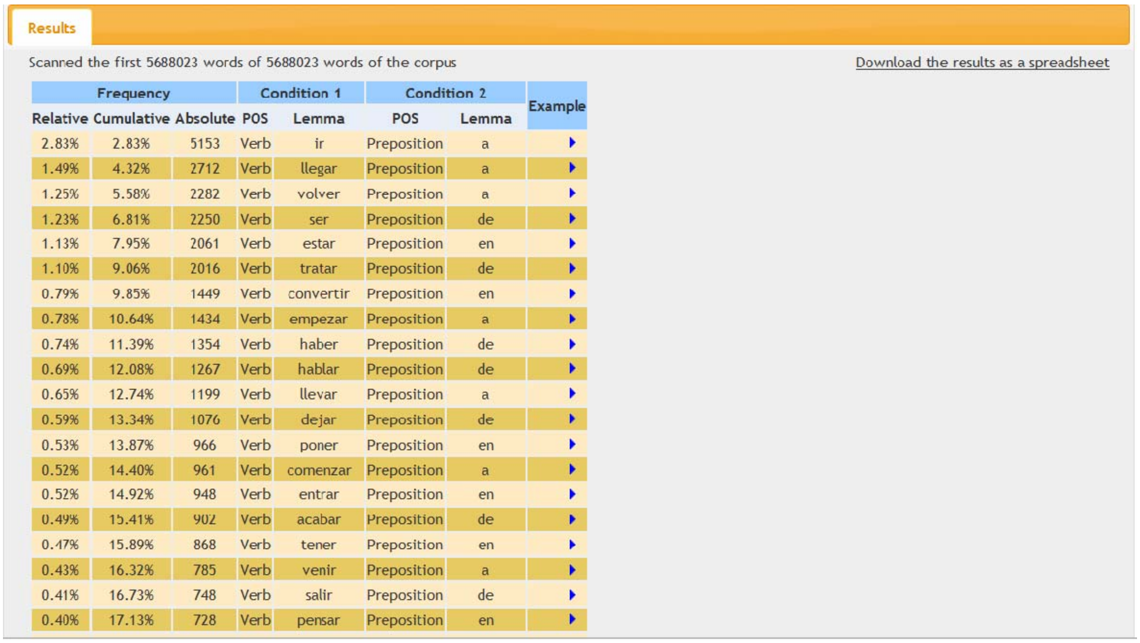
Condition 1: POS: Verb; Frequency: Freq. Lemma; Multiplicity: Choose an option

Condition 2: POS: Preposition; Frequency: Freq. Lemma; Multiplicity: Choose an option

Figure 8. Quantitative search for the verb *pensar* + *preposition*

See the query results in fig. 9, we obtain the lemma *pensar* (to think) with all the available prepositions in the corpus and its frequency.

If we want to access the textual results of the verb *pensar* + a specific preposition, we click on the right blue arrow and we get a search of pensar + the preposition with textual results.



Scanned the first 5688023 words of 5688023 words of the corpus

Frequency			Condition 1		Condition 2		Example
Relative	Cumulative	Absolute	POS	Lemma	POS	Lemma	
2.83%	2.83%	5153	Verb	ir	Preposition	a	▶
1.49%	4.32%	2712	Verb	llegar	Preposition	a	▶
1.25%	5.58%	2282	Verb	volver	Preposition	a	▶
1.23%	6.81%	2250	Verb	ser	Preposition	de	▶
1.13%	7.95%	2061	Verb	estar	Preposition	en	▶
1.10%	9.06%	2016	Verb	tratar	Preposition	de	▶
0.79%	9.85%	1449	Verb	convertir	Preposition	en	▶
0.78%	10.64%	1434	Verb	empezar	Preposition	a	▶
0.74%	11.39%	1354	Verb	haber	Preposition	de	▶
0.69%	12.08%	1267	Verb	hablar	Preposition	de	▶
0.65%	12.74%	1199	Verb	llevar	Preposition	a	▶
0.59%	13.34%	1076	Verb	dejar	Preposition	de	▶
0.53%	13.87%	966	Verb	poner	Preposition	en	▶
0.52%	14.40%	961	Verb	comenzar	Preposition	a	▶
0.52%	14.92%	948	Verb	entrar	Preposition	en	▶
0.49%	15.41%	902	Verb	acabar	Preposition	de	▶
0.47%	15.89%	868	Verb	tener	Preposition	en	▶
0.43%	16.32%	785	Verb	venir	Preposition	a	▶
0.41%	16.73%	748	Verb	salir	Preposition	de	▶
0.40%	17.13%	728	Verb	pensar	Preposition	en	▶

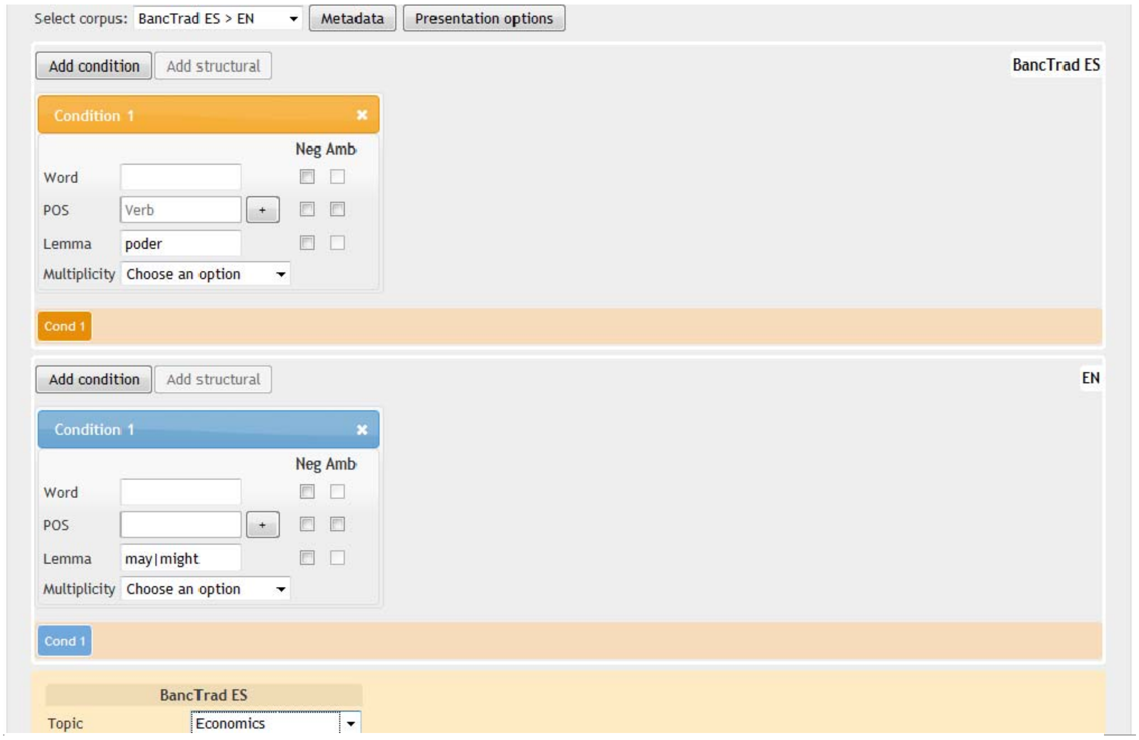
Figure 9. Results of quantitative search for the verb *pensar* + *preposition*

3.1.4 Aligned corpora

The interfaces for aligned corpora are: simple and advanced. As for monolingual corpora, simple search allows for queries for key words out of context and advanced search allows for queries for key word in context.

The main difference between monolingual and aligned interfaces is that queries can be done on the source and the target corpus in both searches (simple and advanced).

For example, in a Spanish > English corpus, we want to study the verb *poder* translated as *may* or *might* in Economics texts. In the advanced search, we need in the source language one condition: *poder* restricted as a verb (as a noun means *power*) and one condition in the target language with a disjunction may or might in the lemma restricted in Economics texts at metadata section (See fig. 10). (Results page can be customized with the same values as the monolingual corpus)



Select corpus: BancTrad ES > EN

BancTrad ES

Condition 1

		Neg	Amb
Word		□	□
POS	Verb	□	□
Lemma	poder	□	□
Multiplicity	Choose an option		

EN

Condition 1

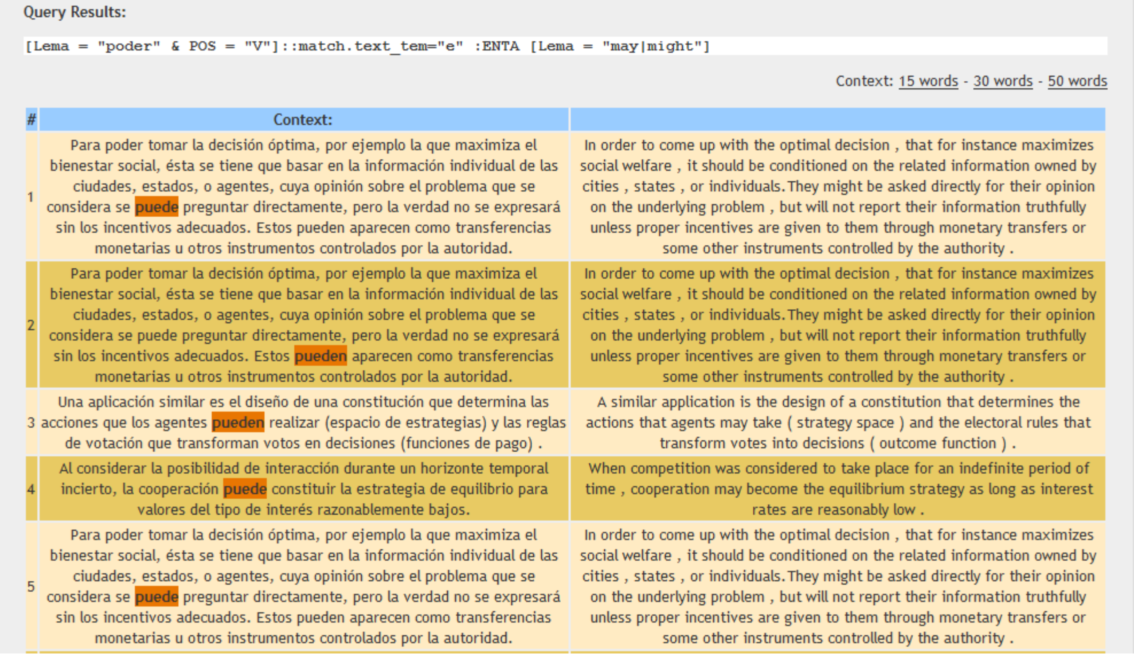
		Neg	Amb
Word		□	□
POS		□	□
Lemma	may|might	□	□
Multiplicity	Choose an option		

BancTrad ES

Topic: Economics

Figure 10. Results of advanced search in aligned corpus Spanish > English

We get the source text with the query match highlighted aligned with the target text (see fig. 11).



Query Results:

```
[Lema = "poder" & POS = "V"]::match.text_tem="e" :ENTA [Lema = "may|might"]
```

Context: 15 words - 30 words - 50 words

#	Context:	
1	Para poder tomar la decisión óptima, por ejemplo la que maximiza el bienestar social, ésta se tiene que basar en la información individual de las ciudades, estados, o agentes, cuya opinión sobre el problema que se considera se puede preguntar directamente, pero la verdad no se expresará sin los incentivos adecuados. Estos pueden aparecen como transferencias monetarias u otros instrumentos controlados por la autoridad.	In order to come up with the optimal decision , that for instance maximizes social welfare , it should be conditioned on the related information owned by cities , states , or individuals. They might be asked directly for their opinion on the underlying problem , but will not report their information truthfully unless proper incentives are given to them through monetary transfers or some other instruments controlled by the authority .
2	Para poder tomar la decisión óptima, por ejemplo la que maximiza el bienestar social, ésta se tiene que basar en la información individual de las ciudades, estados, o agentes, cuya opinión sobre el problema que se considera se puede preguntar directamente, pero la verdad no se expresará sin los incentivos adecuados. Estos pueden aparecen como transferencias monetarias u otros instrumentos controlados por la autoridad.	In order to come up with the optimal decision , that for instance maximizes social welfare , it should be conditioned on the related information owned by cities , states , or individuals. They might be asked directly for their opinion on the underlying problem , but will not report their information truthfully unless proper incentives are given to them through monetary transfers or some other instruments controlled by the authority .
3	Una aplicación similar es el diseño de una constitución que determina las acciones que los agentes pueden realizar (espacio de estrategias) y las reglas de votación que transforman votos en decisiones (funciones de pago) .	A similar application is the design of a constitution that determines the actions that agents may take (strategy space) and the electoral rules that transform votes into decisions (outcome function) .
4	Al considerar la posibilidad de interacción durante un horizonte temporal incierto, la cooperación puede constituir la estrategia de equilibrio para valores del tipo de interés razonablemente bajos.	When competition was considered to take place for an indefinite period of time , cooperation may become the equilibrium strategy as long as interest rates are reasonably low .
5	Para poder tomar la decisión óptima, por ejemplo la que maximiza el bienestar social, ésta se tiene que basar en la información individual de las ciudades, estados, o agentes, cuya opinión sobre el problema que se considera se puede preguntar directamente, pero la verdad no se expresará sin los incentivos adecuados. Estos pueden aparecen como transferencias monetarias u otros instrumentos controlados por la autoridad.	In order to come up with the optimal decision , that for instance maximizes social welfare , it should be conditioned on the related information owned by cities , states , or individuals. They might be asked directly for their opinion on the underlying problem , but will not report their information truthfully unless proper incentives are given to them through monetary transfers or some other instruments controlled by the authority .

Figure 11. Results of advanced search in aligned corpus Spanish > English

4. Public Corpora currently available in IAC

IAC allocates at this very moment the following public corpora:

Cucweb (Corpus d'Ús del Català a la Web) is a monolingual Catalan webcorpus that consists of more than 125,000 documents (more than 200 million words) extracted from the Web under the *.es* domain. In addition, it has been tagged with morphosyntactic information, which allows queries by lemma, part of speech, and syntactic function.

Compartrad/Comparor is a comparable Catalan corpus containing two subcorpora comparable in size, topic and type of text. One of the subcorpus (Comparor) contains texts written originally in Catalan and the other one (Compartrad) contains texts translated (mainly from English) into Catalan. Each subcorpus is about 200.000 words and the documents included in these both subcorpora are in the field of art and proceeds from art exhibition catalogues as well as from art books. This corpus is annotated with POS.

BancTrad is a paralel corpus aligned at sentence level with the following languages pairs: en<->ca, en<->es, fr<->ca, fr<->es, de-<->ca, de<->es. Altogether
it is about 5 Million words. With regard to the sources, the documents proceed from translation teachers, work done in translation classes (translations
done by students under the supervision of the professor), publishing houses and the Internet. Selected translations and originals are annotated with lemma and POS as well as with some extralinguistic information. This extralinguistic tagging allows queries to be filtered by:

Subject matter (economics, science, politics, etc.)

Text type (normative, descriptive, literary, etc.)

Register (colloquial, standard, learned, etc.)

Degree of specialisation (low, intermediate, high)

LEXESP (Corpus de Referencia del Español) is a reference Spanish corpus that is about 5 Million Words and it contains written documents from the period between 1978 and 1995. Regarding to the text type and topic this corpus aims to be balanced: 40% (about 3 Million words) of the documents proceed from novels, 10% from popular science magazines, and 40% from many different Spanish newspapers.

LTC is an annotated corpus composed of translations produced by trainee translators. It comprises originals in English aligned at sentence level with their respective Catalan translations. The LTC currently contains about 145.000 words in Catalan and 114.00 words in English from student translations that have been annotated with translation errors.

Besides all these public corpora, IAC contains other private corpora, but our intention is progressively to extend the number of available corpora.

5. Uploading a corpus to IAC

5.1 Corpus format

Although IAC allows for uploading any type of corpus, it should be in an adequate format. In this section we describe the specifications to upload a corpus in IAC[1].
Firstly, it must be a verticalized text, this means that there must be a word per line (see fig. x). In case, the corpus does not have annotations, it can be uploaded verticalized after interface configuration.

In order to introduce annotations, IAC requires tabular format for attributes at word level (ie. pos or lemma). This means that a word may have more than one attribute and all of them should be separated by a tab.

[1] IAC requires the same format as CWB as IAC uses CWB engine.

```
The          Det    sg
boy          Noun   sg
buys         Verb   sg
pencils      Noun   pl
```

Figure 13. Verticalized and pos tagged text

In Fig. 13, the word *houses* has the lemma *house*, pos *Noun* and number *singular*, all these features at word level are separated by tabs. Not only is interesting to search by tags at word level but also to search at upper levels, such as phrase level, as many phenomena occur over the word level (ie. syntactic information). In these types of attributes, the information should be introduced with xml tags.

Fig. 14 shows an example of syntactic attributes over word level tagged with xml tags.

```
<func=”subj”>
The          Det    sg
boy          Noun   sg
</func>
buys         Verb   sg
<func=”DO”>
pencils      Noun   pl
</func>
```

Figure 14. Verticalized, pos and attributes at phrase level text

And finally, the last type of attributes that can be introduced in IAC are metadata, attributes at text level. Their structure is similar to xml format (see fig 15) but we can add as many attributes as we need.

```
<metadata title = “Demo” year=“2010”>
<func=”subj”>
The    Det    sg
boy    Noun   sg
</func>
buys   Verb   sg
<func=”DO”>
pencils      Noun   pl
</func>
</metadata>
```

Figure 15. Verticalized, pos, attributes at phrase level and metadata annotated text

5.2 Configuration tool

Once the corpus is in the adequate format, the interface is designed through a friendly – user graphical tool integrated in IAC.

As a description of the corpus, the owner must introduce: the name of the corpus, the id (file name), the type of corpus (translated or original) and the group it belongs to in order to restrict the corpus access to a group of users.

The next step is the attributes definition. The interface distinguishes between 3 types of attributes (See fig. 15):

Positionals (orange box): Attributes at word level.

Structurals (green box): Attributes over the word level.

Metadata (light brown box): Attributes at text level.

5.2.1 Positional attributes:

Each attribute separated by tab (word, pos, number, etc.) has to be described by the ID and Name (label that will appear at the interface). There’s also an optional Info field to add an information tip for the user when it goes over the attribute. The configuration interface will have as many orange boxes as many columns the corpus has.

We can also choose between a text box in the interface, ie. to introduce the lemma or a drag and drop menu to select the category of the pos. If we select the Obligatory box, these attributes will appear in the simple and the advanced search. In case, we don’t select the check box, the attribute will only appear in the advanced search.

5.2.2 Structural and Metadata attributes:

Those attributes over word level can be introduced though the green boxes (structural) and light brown boxes (metadata). As the positional attributes, all structural and metadata attributes must be described by the ID, Name, and the optional Info field.

We can also create a text box or a drag and drop menu.

Once the interface is defined according to corpus attributes, the corpus file must be uploaded at the server and a click on the Corpus generation button will compile the corpus and create the interface with 3 search modes (simple, advanced and statistics).

6. Conclusions

In recent years considerable efforts have been made to develop more user-friendly interfaces for corpus exploitation, that allow much more possibilities than simple concordancers without needing to be acquainted with intricate syntax languages. These efforts are particularly welcome in the context of translation studies, in which most researchers come from the humanities. In order to overcome the multiplicity of interfaces derived form the usual association between corpus and interface, new common platforms have been created, which allow the access to multiple corpora. The platform we have created and presented in this paper, IAC, is a new input in that sense. Besides being a platform to host and to exploit different types of corpora (mono- and bilingual) for which textual and quantitative results can be extracted, the most novel and relevant contribution of IAC is that of being at the same time a tool to upload new corpus without programming knowledge. This possibility should be particularly useful taking into account the increasing amount of corpora for different languages, purposes etc. Furthermore, this functionality offers researchers a greater autonomy in their work facilitating at the same time new possibilities to exchange and collaboration in the field.

Notes

[1] IAC uses the IMS Open Corpus Workbench

CWB available at:: http://cwb.sourceforge.net (accessed: 8 Oktober 2010).

[2] See, for example, the MeLLANGE Learner Translator Corpus

Avialable at: http://corpus.leeds.ac.uk/mellange/ltc.html (accessed: 8 Oktober 2010).

[3] See *The British National Corpus*, version 2 (BNC World). 2001. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Available at: http://www.natcorp.ox.ac.uk/ (accessed: 8 Oktober 2010).

[4] IDS Corpus available at: http://www.ids-mannheim.de/cosmas2/ (accessed: 8 Oktober 2010).

[5] REAL ACADEMIA ESPAÑOLA: Database (CORDE) [on line]. *Corpus diacrónico del español.* Available at: http://www.rae.es (accessed: 8 Oktober 2010).

[6] CORIS/CODIS available at: http://corpora.dslo.unibo.it/coris_ita.html (accessed: 8 Oktober 2010).

[7] Linguee was founded by Gereon Frahling and Leonard Fink. Linguee has a web crawler over bilingual web or other valuable sources include EU documents and patent specifications and that extracts the translated sentences. The texts are then evaluated by a machine-learning algorithm which filters out the high quality translations for display.

References

CQP available at: http://cqpweb.lancs.ac.uk/ (accessed: 8 Oktober 2010).

Granger, S. (1998). *Learner English on computer*. London: Longman.

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* (3), p. 538-546.

Johansson, S. (1998). “On the role of corpora in cross-linguistic research”. In S. Johansson & S. Oksefjell (Eds.), *Corpora ans Cross-linguistic Research* (pp. 1-24). Amsterdam and Atlanta: Rodopi.

Kilgarriff, A. and G. Grefenstette (2003). "Introduction to the Special Issue on the Web as Corpus". *Computational Linguistics*, vol. 29, n. 3, pp. 333-348.

Sharoff, S. (2006). “Open-source corpora: using the net to fish for linguistic data”. In *International Journal of Corpus Linguistics* 11(4), 435-462.

Sketch Engine: http://www.sketchengine.co.uk/ by Adam Kilgarriff, Pavel Rychlý, Jan Pomikálek.

Jaguar: http://melot.upf.edu/cgi-bin/jaguar/jaguar.pl by Rogelio Nazar (accessed: 8 Oktober 2010)

Nazar, R.; Vivaldi, J. & Cabré, MT. ; (2008). "A Suite to Compile and Analyze an LSP Corpus" (PDF), Proceedings LREC 2008 (The 6th edition of the Language Resources and Evaluation Conference) Marrakech (Morocco), 28-29-30 May 2008.