CONSUMER EROSKI PARALLEL CORPUS[1]

Asier Alcázar
University of Missouri-Columbia

Abstract

This paper introduces the Consumer Eroski Parallel Corpus, a collection of articles originally written in Spanish and later translated to three languages also spoken in Spain: Basque, Catalan and Galician. The articles have been correlated in the four languages at the sentence level automatically using Moore's bilingual sentence alignment tool (2002). The Spanish section is also annotated morphosyntactically for parts of speech using SVMtool (Giménez and Márquez 2004). The Basque, Catalan and Galician sections may be annotated in a future release with the collaboration of Computational Linguistics Groups in Spain. To my knowledge, the Consumer Eroski Parallel Corpus is the first resource to exist that encompasses a substantial body of parallel text from these four languages spoken in Spain. I would like to thank the Eroski Foundation for granting permission to share the corpus in the public domain. Making this resource public will provide additional opportunities to test, train and develop natural language processing tools in the computational linguistics community. It may also help translators as a reference. With the addition of an advanced search interface, currently under development, the corpus may be consulted by Basque and Romance linguists interested in cross-linguistic research.

1. Introduction

The Consumer Eroski Parallel Corpus (henceforth CEPAC for short) is a database consisting of all the articles published in the online version of the Consumer Eroski magazine (*revista.consumer.es*).

[1] I would like to thank the BIDE 2005 organizing committee (Irene Barbería, Rebeka Campos, Susana Huidobro and Leticia Pablos), Jon Ortiz de Urbina and, last but not least, Jon Franco, for their invitation to present at the conference, and the audience for comments and feedback. Thanks also to Joseba Abaitua and everyone at the DELi Computational Linguistics group at the University of Deusto for discussion and encouragement to pursue this project. This paper has also benefited from a presentation at and attendance to *Ordenagailuz Lagundutako Itzulpena* (Basque for Computer-Assisted Translation; Summer Courses of the University of the Basque Country, Miramar Palace, May 30-June 2, 2005). I am grateful to Josu Waliño Jn. and the Elhuyar Foundation, the presenters and audience for helping me understand the place of the Consumer Eroski Parallel Corpus among the current existing resources for Basque, Catalan, Galician and Spanish. I am indebted to Itziar Otegi for getting me in touch with Ricardo Oleaga, Iker Merchán, and Ainara Zarraga from the Consumer Eroski magazine, all of whom enthusiastically supported this project.

The Consumer Eroski magazine is a free publication produced by the Eroski Foundation (*www.fundaciongrupoeroski.es*). It exists online since January 1998, although the Eroski Foundation has published this magazine on paper under different names for the last 30 years. When the magazine initiated its life online, it did so under the name Consumer, which was eventually changed to Consumer Eroski. Consumer was initially published entirely in Spanish (January 1998) and gradually started to be translated to Basque (November 1998), then Catalan (January 2000) and finally Galician (April 2000). Since the current name of the magazine is Consumer Eroski, I find it fitting to name the corpus CEPAC.

The Consumer Eroski magazine has been published online for 8 years now. During this time the magazine has produced a new issue every month, with the exception of the months of July and August, which make up a joint number, for a total of 11 issues a year. Each issue contains 14 different sections of varying length that cover diverse topics (see table 1), the signature of the magazine being its consumer reports.

Table 1

List of Sections in Consumer Eroski Magazine (Spanish)

Tema de portada	Psicología	Medio ambiente	"Lo más práctico"
Análisis de productos	Miscelánea	Alimentación	Consejos
Informe	Economía doméstica	Nuevas tecnologías	Sentencias
Salud	Entrevista		Consultorio legal

Since the year 2000, the Spanish articles have been translated to the three other languages and the magazine has been published simultaneously in the four languages. Thus, the Consumer Eroski magazine has been published effectively as a multilingual edition since April 2000.

According to the Eroski Foundation, *revista.consumer.es* receives an average of 300,000 monthly visits. These numbers, added to the paper version, make the magazine an important publication in Spain.

At the time of submitting this paper, the current collection contained in CEPAC encompasses all the issues from January 1998 to October 2005. The amount of text for each language is substantial (as there are 14 sections times 11 issues times 8 years). All articles exist in Spanish, for this is the language in which the magazine is originally written. Consequently, the Spanish section of the corpus is the largest (1.078 articles). It is closely followed by Basque (991), Catalan (855) and Galician (806). The word count in each section of the corpus exceeds a million. The reader is referred to table 2 at the end of the next section, which details the current size of CEPAC in the different languages.

The reminder of this paper is structured as follows. Section 2 explains the rationale for CEPAC and the steps I took in building it. Section 3 provides an in-

formal overview of the currently available monolingual and bilingual corpora and places CEPAC in the context of these collections. Section 4 introduces a search interface, currently under development, that permits consulting the corpus beyond simple queries and explores how best to integrate advanced linguistic searches that are research oriented. The last section concludes with the release of CEPAC and future updates.

2. Building CEPAC

This section introduces anecdotically my first contact with the online magazine, the driving force behind a project like CEPAC, and the steps that I took to build the corpus.

2.1. First contact and reason d'être for CEPAC

The first time that I came across the Consumer Eroski Magazine I was browsing the World Wide Web in search for web pages that could be used as a reference for health care terminology in Spanish speaking countries. This endeavor was part of a collaborative effort with Hablamos Juntos (*www.hablamosjuntos.org*), a non-profit organization in the United States. Our goal was to gather a set of websites that would guide the development of a Spanish Glossary for health care terms, one that would pay attention to existing cultural differences in the use of terminology. This glossary would assist translators in the difficult task of rendering health care terms to a diverse Spanish-speaking population that are often particular to the United States health care system. The Consumer Eroski magazine stood out for its consistency in the use of health related terminology among the Spanish websites surveyed in Spain.

The Consumer Eroski magazine also presented a rare opportunity to a linguist and computational linguist like me. The magazine is published in three Romance languages (Catalan, Galician, and Spanish) and a language isolate (Basque). The parallel nature of the text, consisting of full translations of the Spanish original to three languages, called for processing this wealth of materials into a database for research use.

Although it would have been wisest to share the programming load of turning these web pages into a database with other computational linguists, it has not always been the case that parallel corpora, rather than monolingual corpora, has been the primary objective in the field. More recently, however, with the success of statistical approaches to machine translation (Knight & Marcu 2005), the search for and processing of parallel text into parallel corpora has received greater attention. For example, the European Constitution has been turned into a multilingual parallel corpus that exists for the most widely spoken languages among the member countries (*http://logos.uio.no/opus/*). It is unfortunate that this project does not yet include Basque, Catalan and Galician, among many other languages spoken in the European Union. In computational linguistics there is less of an interest in minority languages as compared to widely spoken languages that could result into profitable applications. In contrast, in linguistics and related disciplines, linguistic diversity enriches

our perspectives on language as a cognitive phenomenon. It is in gaps of this kind that CEPAC finds its niche.

2.2. From *revista.consumer.es* to CEPAC

What follows is a brief account of the steps and software involved in the processing of the Consumer Eroski magazine into CEPAC.

The first step consisted in devising a strategy for the automatic download and grouping of the original articles with their translations. To this end, I wrote a spider program in Python that would crawl *revista.consumer.es* and harvest the articles. This was an inappropriate way to access the contents of the magazine and I apologize to the Eroski Foundation for having procured the entire magazine without their permission. I initiated this project as a research activity and did not realize that in time the project could be of use to the research community. My gratitude here goes to the Eroski Foundation for their understanding and their generosity in allowing me to share this resource freely in the public domain.

The second step was to strip the articles from anything but plain text. At this stage I also developed my own programs in Python that would achieve this task. It was important to carefully study the formatting of the web pages to exploit some of these tags as sentence and paragraph dividers for use in the later stages of text processing. Some of these tags were critical in the later processing of the text. For example, the tags that specifically divide sequences of headlines or numbered/bulleted lists might be the only way to effectively divide chunks of text that have no terminating punctuation. Furthermore, later stages of processing attain more satisfactory results if they are fed the appropriate chunks of text. Understandably, a part of speech tagger will better resolve the morphosyntactic labels for a headline (an element that may be a phrase rather than a sentence) if this headline has not been put together with the following or preceding main/secondary/tertiary… headline, possibly creating a long sentence without a verb.

A third step involved breaking the plain text files into sentences with the added complexity of the inconsistency in punctuation that results from manipulating vast amounts of texts. The difficulty in this task was increased by the need to process four different languages, with overlapping yet not identical punctuation conventions. I needed to devote much time to this stage to maximize the results of the two third party applications (the aligner and the tagger) that would complete the current stage of the database. During this process I was able to treat a high number of exceptions in a systematic way, yet I also made certain case specific provisos to rescue a few hundred exceptional cases. The resulting corpus is by no means perfect but it has benefited from extensive human supervision.

To align the different language pairs, I used Robert Moore´s alignment tool (2002), written in Perl, and refer the reader to his paper for the implementation details. It is of interest to note that the success of the alignment tool in the Spanish-to-Catalan and Spanish-to-Galician pairs has been higher than the Spanish-to-Basque pair. In the former two, the alignment exceeds 92% of the text, while in the latter the alignment falls to 84%.

There are at least two potential reasons to account for this difference in the results. One is to assume that the Basque translation has fewer 1-to-1 correspondences with respect to the original text. However, it is difficult to establish this fact with utmost certitude without incurring into a manual verification. All things being equal, the quality of the translation to the three languages, and the agreement observed for Catalan and Galician, suggests that over 90% of correspondences should be expected for Basque as well. This leads us to consider the other possible reason, namely that the agglutinative nature of the Basque language makes the automatic building of a dictionary of word correspondences sparser (this is an intermediate step in the alignment process), as multiple correspondences may be established between Spanish and Basque words due to the different inflectional endings. Indeed, Basque has a rich declensional system with 17 different cases, most of which have four different number forms, and there exist six alternative case forms for animates (see Zubiri and Zubiri 2000). Without a lemmatizing tool for Basque that separated the ending from the stem, it is not possible to overcome this morphological obstacle. This second reason seems a better candidate to explain the lower percentage in the absence of manual verification. It is plausible that a new alignment with Basque lemmatized text would yield more pairs of aligned sentences.

That said, it is perhaps more interesting to bring attention to the fact that the alignment from Spanish to Basque was highly successful regardless. The alignment program had to face that Basque is agglutinative (meaning fewer words: etxe-ra 'to the house (house-to.the)'; see table 2 in the next subsection) and a data sparsity problem arising from the rich morphology of Basque (one to many words in the Spanish-to-Basque dictionary). The usability of the alignment program attests to the robust technology employed in alignment systems and Moore's design in particular. Incidentally, the Spanish-to-Basque character ratio approximates 1.0, as seen for English, French and German (Church & Gayle 1993), so character-length based approaches could also fare well in principle.

Finally, to annotate the Spanish section with part of speech information, I used the C++ version of the part of speech tagger by Giménez and Márquez (2004), based on Support vector Machines, with the models for Spanish based on the LEXESP corpus (5.5 million annotated words). A technical description of this tool can be found at *www.lsi.upc.edu/~nlp/SVMTool*. I have not corrected manually any of the errors that the tagger may have done. As stated above, repeated revisions of the success of sentence dividers visibly improved the performance of the tagger, particularly in headlines and cases where sentence integrity had been compromised.

2.3. CEPAC in numbers

This subsection offers a numeric view of the corpus.

The Consumer Corpus consists of 1078 articles written in Spanish (January 1998 to October 2005). Of these, 804 have been translated to Basque, Catalan and Galician. The table below (2) shows the number of sentences, words and characters for the fully multilingual version of the corpus. As noted earlier, because Basque is an agglutinative language, it has fewer words (0.87 million) than its non-agglutinative

neighbors (between 1.09~1.19 million). Note that the number of characters is similar (6.9 million) compared to Catalan (6.9 too) or Galician (6.7).

The numbers in the table are contingent. The Consumer Eroski magazine continues to be published and is well worth periodical updates that will cause the numbers to increase. Future revisions in sentence division may modestly alter the numbers too.

Table 2

804 articles in four languages in CEPAC

	Spanish	Basque	Catalan	Galician
Sentences	59,111	56,531	56,765	57,027
Words	1.19 m	0.87 m	1.14 m	1.09 m
Characters	7.29 m	6.90 m	6.90 m	6.70 m

The actual number of sentences in each section of the corpus is higher (Spanish: 79,982; Basque: 70,496; Catalan: 60,697; Galician: 57,147). As noted in the introduction, translation started at different times. For this reason, there are pockets of articles that have not been translated to Galician, or to Galician and to Catalan. The first 10 issues of the online version of the magazine were not translated either.

3. CEPAC in the context of other existing monolingual and parallel corpora

The goal of this section is to orient the reader as to the niche that CEPAC occupies among the existing monolingual and parallel corpora. It is by no means intended to be an exhaustive list of corpora available in Spain or elsewhere. For further references, see the Linguistic Data Consortium (*www.ldc.upenn.edu*).

The characterizing property of CEPAC as a parallel corpus is that it contains the same text in four languages, presenting unusual pairs like Basque-to-Galician, Galician-to-Catalan, Catalan-to-Basque, etc.

With the exception of the European Constitution, if it is eventually processed for all the languages spoken in the European Union, there is little chance to find parallel text of this nature. Some companies aim to gain market share in Spain and dedicate part of their budget to better market their products via linguistic localization. However, the bulk of these materials are likely to be advertising texts and product manuals (e.g. telecommunications). While literary resources may be found that are translated in the four languages, literary classics, to name one such source, it is unlikely that these texts will be localized in one place.

As we narrow our focus to particular languages or specific language pairs, we find that substantive and sometimes vast collections of text are available for Basque, Catalan, Galician and Spanish. Many of these can be accessed online, although full access to these resources is generally not possible. The following is a random and informal walk through some of these resources.

The *Real Academia de la Lengua Española* (Spanish for Royal Academy of the Spanish Language) or RAE has two large collections of literary, journalistic and oral texts online divided into two monolingual corpora: CREA for contemporary Spanish (1976 to present date) and CORDE, a historical corpus that extends back to the 19th century (*www.rae.es*). These texts contain information relating to their author, topic, publisher, publication year, and country of origin (Spain, Portugal, Latin America, United States, Philippines). This information provides new opportunities for studies in language variation and sociolinguistics. For example, Mayoral Hernández surveyed the position of frequency adverbials in Spanish in a recent study that made use of data from CREA (2004). The downside of CREA and CORDE is their limited access. Results are constrained to 1000 per query, these matches being either paragraphs and paragraph chunks or units smaller than a sentence.

Given that CREA does not contain the Spanish section of CEPAC, it may one day be added to its press section. While CEPAC alone cannot be used for sociolinguistic purposes (author information is not included), it provides a different opportunity to study linguistic variation in translation or across particular language pairs. To facilitate this task, it is necessary that the search interface provided for users allows more flexibility in its advanced searches (support for part of speech search, for example), and the convenience to return sentences. In any case, the distribution of CEPAC as a free resource will grant access to circa 250,000 sentences.

The monolingual corpus for Basque provided by Euskal Herriko Unibertsitatea (Basque for University of the Basque Country) contains approximately 18 million words (*http://www.ehu.es/euskara-orria/euskara/ereduzkoa/araka.html*), 8.7 million in the press section. CEPAC would be worth including to bring this section closer to the 10 million barrier. The Basque reference corpus is fairly limited in its temporal scope 2000-2005 compared to CREA. Its user interface is more convenient though in that it returns full sentences and allows searches complemented by part of speech information.

The largest parallel corpus for Basque and Spanish is LegeBi or 'Official bilingual gazettes from the Basque Administration (1994-2004)' collected by the DELi Computational linguistics Group (*www.deli.deusto.es/AboutUs/Resources/LegeBi*). Similarly to the European Constitution or the UN proceedings, LegeBi is another example that the administrative domain is a frequent and abundant source of parallel text. Like LegeBi, CEPAC also provides the opportunity to compare the same text in the two languages with the convenience of doing so at the sentence level. The Basque reference corpus and CEPAC more closely resemble the everyday written standard Basque.

Regarding Catalan, CucWeb (*http://ramsesii.upf.es/cucweb/about.en_US.htm*) is an attractive example that serves to illustrate a different type of corpus, and one that will be increasingly available. It consists of over 200 million words collected from web pages written in Catalan and, like the Basque reference corpus, can be consulted with the aid of part of speech information. To its favor, CucWeb and similar corpora possibly outweigh all other collections in its size. On the other side of the coin, CucWeb is an all-encompassing collection of texts, not an organized and annotated collection like CREA or CORDE. For this reason, this type of corpus is not a reference for the written standard.

Finally, for matters relating to Galician monolingual and parallel corpora, it is best to refer the reader to CLUVI (*sli.uvigo.es/CLUVI/*) and references therein.

4. CEPAC as a reference tool

This section illustrates the value of the corpus as a reference tool and introduces an advanced search interface project that aims to bring queries to a higher level of abstraction.

It is little wonder that one of the advantages of parallel corpora is to present the same sentence in more than one language to bilingual speakers (translators, journalists, second language learners). By way of example, I may have come across the Spanish acronym ONG for non-profit organization and may wonder how to express such concept in Basque. Querying the corpus for ONG in Spanish, I find out that GKE is the corresponding acronym and that it stands for 'Gobernuz Kanpoko Erakundea'.

Table 3

Sample sentences containing 'ONG' and their Basque counterparts

Spanish	Son cada vez más frecuentes las noticias sobre malas gestiones de ONGs y los actos reprochables de las específicamente que trabajan en proyectos de desarrollo.
Basque	Gero eta sarriago agertzen dira GKEen gestio txarrei buruzko albisteak eta garapen proiektuetan lan egiten dutenen jokaera gaitzesgarriak.
Spanish	Específicamente, la labor de las ONG es la práctica de ayuda humanitaria, es decir, la asistencia de las sociedades vulnerables y vulneradas de la que nos ocupamos sobre el terreno.
Basque	Gobernuz kanpoko erakundeon eginkizuna giza laguntza ematea da, hau da, erasogarri eta eraso jasanak diren gizarteetan asistentzia ematea.

In time, advances in computational linguistics may provide sophisticated ways to query databases that enable linguistic research beyond its most widespread trends today (authorship, concordances, language variation, sociolinguistics, etc). However, for that matter texts should be linguistically searchable. By linguistically searchable, I mean that the search should ideally be formulated in linguistic terms and at different levels of abstraction. I continue to develop a search interface in Python that explores these new avenues (see table 4).

At this point the search for Spanish can be abstracted to parts of speech because this section of the corpus is annotated. It is important to provide the capability to search beyond particular words and we have seen several resources in the earlier section that offer this feature. For my interface, I have defined an extensive list of linguistically relevant tags that abstract the particular part of speech tags used by the tagger (reduced Parole tagset). For example, the verbs are annotated with 18 different tags: auxiliary, semi-auxiliary and main verb have different tags for the infinitive, gerund, participle, indicative, subjunctive and imperative forms. I group these 18 tags into VERB, and lower levels of abstraction like AUX for all auxiliary forms (also S-AUX, MV), IND for all indicative forms, NFVB for all non-finite verb forms, etc. This allows for searches like 'siempre' followed by VERB followed by INF (any infinitive form). Similar levels of abstraction are defined for the remain-

ing part of speech tags. All lower levels tags, for example VMI (main verb indicative form), are directly accessible as well.

For example: power search all questions, find those with interrogative pronouns with POS tag search, then filter results to those with, say, auxiliaries or indicative verbal forms with Verb form search.

Table 4

Overview of search levels, available categories and sample results

Level	Categories	For ex.: intended result
Morpheme	Prefix, infix, suffix and part of speech	A collection of Spanish nouns ending in -ción
Word	Hundred+ linguistic categories of various levels of abstraction (main verb participle > main verb non-finite > non-finite > verb)	A collection of Spanish sentences where *siempre* precedes a non-finite form
Phenomenon	All of the above plus specific search modes — Verb search — Part of speech search — Morpheme Search — Regular expression search — Power search (a predefined set) — Operator search — Chain search (all of the above)	A collection of sentences with clitic doubling
Construct		A collection of absolute participial clauses
Sentence		A set of questions with interrogative pronouns
Paragraph		A set of paragraphs with overt subject pronouns

On the linguistic side, the alignment provides the possibility to compare search results across three Romance languages and a language isolate. In effect, the search results for a query in a language may be accompanied by its translations to the other three. For example, a regular expression search for clitic doubling in Spanish may find interesting companions in the Catalan and Galician translations. A power search for absolute constructions in Basque, which are ambiguous between absolute participial and gerundival clauses, may find revealing correlations in the Spanish original, which uses distinct non-finite forms. The translations to Catalan and Galician, in turn, offer the possibility to do a case study on non-finite non-complement clauses in Romance. These data have helped my joint research with Mario Saltarelli on participial clauses.

5. The Publication of CEPAC

Thanks to the permission granted by the Eroski Foundation, CEPAC will be placed in the public domain for research or reference. Future versions of CEPAC may have the Basque, Catalan and Galician sections of the corpus annotated for parts of speech and a slightly better Spanish-to-Basque alignment.

References

Gale, W. and Church, K., 1993, «A Program for Aligning Sentences in Bilingual Corpora», *Computational Linguistics* 19:1, 75-102.

Giménez, J. and Márquez, L., 2004, *SVMTool: A general POS tagger generator based on Support Vector Machines***.** Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal.

Knight, K., D. Marcu, 2005, *Machine Translation in the Year 2004*, Proceedings of ICASSP.

Mayoral Hernández, R., 2004, «Importance of Weight and Argumenthood on the Ordering of Adverbial Expressions», *Proceedings of the 23rd West Coast Conference on Formal Linguistics (WCCFL),* edited by V. Chand, A. Kelleher, A. Rodríguez and B. Schmeiser. Somerville, MA: Cascadilla Press, 569-582.

Moore, R. C., 2002, «Fast and Accurate Sentence Alignment of Bilingual Corpora». In *Machine Translation: From Research to Real Users* (Proceedings, 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California), Springer-Verlag, Heidelberg, Germany, 135-244.