Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 77-84 recibido 29-11-2016 revisado 10-01-2017 aceptado 07-02-2017

ANALHITZA: a tool to extract linguistic information from large corpora in Humanities research

ANALHITZA: herramienta para extraer información lingüística de corpus extensos para su uso en investigaciones de ciencias humanas

Arantxa Otegi[1], **Oier Imaz**[2], **Arantza Díaz de Ilarraza**[1], **Mikel Iruskieta**[1] y **Larraitz Uria**[1]
[1]IXA Group. University of the Basque Country, UPV/EHU
[2]PRAXIS Research Group. University of the Basque Country, UPV/EHU
arantza.otegi@ehu.eus, oier.imaz@ehu.eus, a.diazdeilarraza@ehu.eus, mikel.iruskieta@ehu.eus, larraitz.uria@ehu.eus

Resumen: El tamaño reducido de los corpus en ciertos campos de investigación se debe a la falta de herramientas para procesar el lenguage de forma masiva y sencilla. En este artículo presentamos ANALHITZA, una herramienta que estamos desarrollando dentro del proyecto Clarin-k que tiene como objetivo principal la creación de tecnologías lingüísticas útiles para la investigación en Ciencias Sociales y Humanidades. ANALHITZA ha sido diseñada para extraer información lingüística online de textos extensos de una forma sencilla. Además, es una herramienta multilingüe que permite analizar textos escritos en tres lenguas: euskera, castellano e inglés. En este artículo, a modo de ejemplo, presentamos tres estudios en los que se ha usado esta herramienta, que puede ser rediseñada para cubrir las necesidades de investigación de muchas de las ramas de Humanidades.
Palabras clave: Herramienta, tecnologías del lenguaje, corpus, análisis de texto, PoS

Abstract: The reduced size of corpora in some areas of research is due to the lack of tools to process massively and easily the language under study. In this article, we present ANALHITZA, a tool which is being developed within the Clarin-k project, whose aim is the creation of linguistic technologies that are useful for research on Social Sciences and Humanities. ANALHITZA has been designed to extract linguistic information online from large corpora in an easy way. Besides, it is a multilingual tool which can process texts written in three languages: Basque, Spanish and English. Moreover, we present three real examples of study where ANALHITZA has been used. The tool can be redesigned or changed, according to the needs of the scientific community in the field of Humanities.
Keywords: Tool, language technologies, corpora, text analysis, PoS

1 Introduction

How can Language Technology (LT) tools be applied in the Humanities research? How can these technologies help in, for example, getting accessible the needed corpora for such researches? Humanities projects are grounded in a dataset that, from a quantitative point of view, is typically used in some kind of statistical analysis to confirm or not a particular hypothesis which will be developed in the process of exploring the dataset. Usually, the size of the dataset used is reduced, because the analysis of bigger amounts of texts is not manually affordable. Having these aspects in mind, some key questions arise: are the researchers in Humanities aware of the possibilities offered by LT tools? Are the researchers in Natural Language Processing (NLP) ready to tackle the problems researchers have in the Humanities field?

There are several reasons why the researcher in Humanities avoid the use of LT tools: *i*) there are not available many tools which can analyze the linguistic phenomena in the language under study; *ii*) in case there is a tool, it may require economic costs or technical expertise to use it; *iii*) the output quality of the tools available cannot be compared to the results obtained by human annotation, or *iv*) the tool is unknown to the community.

Therefore, it is important to make avail-

ISSN 1135-5948

able to researchers in the Humanities and Social Sciences digital multilingual tools that can be easily chained to perform complex operations in order to support them in their work.

In this article, we present three preliminary studies we have been working on in the fields of Humanities and Social Sciences. These studies have been developed on the results produced by ANALHITZA, an application which provides users with linguistic information concerning written texts. Such information is based on an automatic morphosyntactic analysis, which is carried out using NLP tools. The application is still under development.

The article follows the subsequent structure: Section 2 presents some related work. Section 3 describes the system ANALHITZA. In Section 4 we present the results of three studies carried out using the tool. Finally, Section 5 sets out the conclusions and future work.

2 Related work

In the Virtual Language Observatory[1], created in the framework of CLARIN (Common Language Resources and Technology), we can find several tools for the automatic processing of language oriented to eHumanities as well as some interesting resources for different languages. General projects such as MetaShare,[2] developed in the context of MetaNet[3] and ELRA catalog,[4] offer an interesting and useful overview of collections of tools and linguistic resources for general purposes. AntConc[5] and LancsBox[6] are, for example, two interesting tools that provide an easy access to the results, but with the inconvenience that cannot be used online. Another useful tool is CONTAWORDS, an application presented in Villegas et al. (2012), who show that Language Resources and NLP can help in different researches in the Humanities.

In this way, content analysis is accessible with LT, because one can find some related or hidden semantic structures in a text body or check if the semantic structures of the language or knowledge fits with our predictions. Content analysis is aimed at data reduction (Alonso and Volkens, 2012), since texts are very complex and entail high degrees of variability in terms of linguistic expressions (Krippendorff, 2004). Thus, analysis begins with the application of several preprocessing techniques to reduce the complexity of 'texts as data' (Grimmer and Stewart, 2013). Depending on the method and the aimed results, one can use different approaches, to cite some: *a*) Topic Models (TM) erase any information about ordering (bag-of-words) reducing texts to lists of unique words (Blei, 2012) or *b*) Network Text Analysis (NTA), on the contrary, retains ordering to maintain the pattern of textual linkage between concepts in terms of their proximity (Carley, 1997).

The tool we present in this paper, ANALHITZA, aims to be helpful at least in the directions shown here with three different studies: *i*) a specific task of linguistic textual analysis in literature, *ii*) content analysis of transcripts of a deliberative exercise and *iii*) data manipulation to analyze the best indicators of the main topic in a multilingual corpus.

3 System description

ANALHITZA is a tool that, in a nutshell, processes text automatically and extracts some linguistic information concerning the analyzed text.

The in-house version of the system has a simple command-line interface that offers the possibility to pass a single document or a directory which could contain many documents to analyze. The online version, which is publicly available,[7] does not offer the option of analyzing more than one document at once. But using its simple interface (cf. Figure 1), the user can submit a text to be analyzed in one of the following three ways: *i*) uploading a plain text file, *ii*) writing the text in a text box or *iii*) specifying the URL of the website that contains the text. Both versions of the system are able to process texts in three different languages (Basque, English and Spanish). The user has to specify the language, submit the texts to be analyzed and the system will provide the results to the user in a spreadsheet (Excel file).

[1] https://www.clarin.eu/content/virtual-language-observatory.

[2] http://www.meta-share.org/.

[3] http://www.meta-net.eu/.

[4] http://catalog.elra.info/.

[5] http://www.laurenceanthony.net/software.html.

[6] http://corpora.lancs.ac.uk/lancsbox/.

[7] http://ixa2.si.ehu.es/clarink/analhitza.php?lang=en.

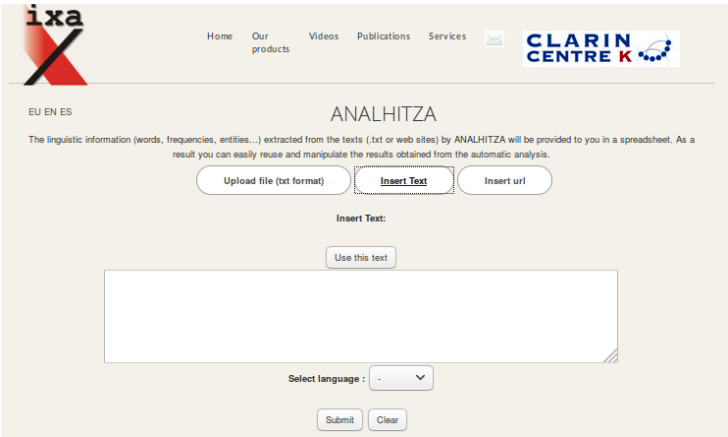

Figure 1: The interface of ANALHITZA

3.1 Automatic text processing

The input of the system is the text itself, which is analyzed and processed with some NLP tools such as tokenizer, lemmatizer, Part of Speech (PoS) tagger and Named Entity Recognizer and Classifier (NERC).

For that purpose, IXA pipes[8] and ixaKat[9] tools are used. IXA pipes is a modular set of NLP tools (or pipes) which provide easy access to NLP technology for several languages (Agerri, Bermudez, and Rigau, 2014). Similarly, ixaKat is a modular chain of NLP tools especifically created for Basque by means of hybridation techniques that combine knowledge and statistical approaches (Otegi et al., 2016). One of the main features of both set of tools is their modularity. That is, the tools in the pipe or in the chain can be picked and changed, as long as they read and write the required data format via the standard streams. All the tools in both sets read and write NAF format,[10] a linguistic annotation format designed for complex NLP pipelines (Fokkens et al., 2014). This way, it is possible the interaction between ixaKat and IXA pipes modules. In that way, Basque texts are analyzed using first a ixaKat tool, and afterwards a IXA pipe tool. Namely, the robust and wide-coverage morphological analyzer and PoS tagger from ixaKat (*ixa-pipe-pos-eu*) is linked to the NERC tool provided by the IXA pipes (*ixa-pipe-nerc*) (Agerri and Rigau, 2016).

Regarding the performance of these tools, all the tools used obtain state-of-the-art results. The Basque morphological analyzer which ixa-pipe-pos-eu is based on obtains 95.17% in accuracy on PoS tagging. The ixa-pipe-pos tool for English and Spanish lemmatization and PoS tagging obtains, respectively, 96.88% and 98.88% in accuracy. The NERC tool obtains a performance of 0.7672 (F1), 0.8621 (F1) and 0.8016 (F1) for Basque, English and Spanish, respectively.

Once the linguistic processing is carried out and based on the information in the output NAF document, some basic maths (such as counting different words, cases or entities, for examples) and filtering are applied.

3.2 Output

All the resulting data are compiled and presented in a spreadsheet to the user. The information is presented in several worksheets or sheet tabs (18 sheet tabs for Basque and 17 for Spanish and English).

In the first sheet tab, some general information is shown, including number of letters, words, lemmas, nouns, adjectives, verbs, adverbs, determiners, conjunctions, prepositions, named entities, sentences as well as the average sentence length and the number of words in the shortest and longest sentences.

In tab sheets from the second to eighth the lemmas of different nouns, adjectives, verbs, adverbs, determiners, conjunctions and prepositions found in the text are listed, with their respective frequency counts. The ninth sheet tab is only available for Basque texts and it shows the different declension cases found in the text. In fact, lemmatization is necessary to recognize the lemmas and the attached determiners (the indefinite singular $-a$ 'one' or the indefinite plural $-ak$

[8] http://ixa2.si.ehu.es/ixa-pipes/.
[9] http://ixa2.si.ehu.es/ixakat/.
[10] http://wordpress.let.vupr.nl/naf/.

'some') and/or the declension cases (ergative *−k*, absolutive *−a/ak*, dative *−(r)i* 'to', ablative *−tik* 'from', destinative *−tzat* 'for', inessive *−n* 'in' and genitive *−(r)en* ''s', among others).[11] In the tenth sheet tab, named entities are listed specifying their frequencies and also their classification-type (person, location or organization). In the following two sheet tabs, different lemmas and word forms (including all PoS tags) with their frequency counts are listed, respectively. Next, different alphabetic letters are listed in the thirteenth sheet tab. Tab sheets from fourteenth to sixteenth show 2-grams, 3-grams and 4-grams extracted from the text. The last two sheet tabs show the lemmatized (with PoS) and unformatted text, respectively.

Based on all that information, users can easily analyze the results and make conclusions in regard to the linguistic aspects of the text.

4 ANALHITZA in Humanities and Social Sciences

ANALHITZA offers users the possibility to extract linguistic information from large corpora in a very easy way, and it can be used to analyze any type of text in most of the disciplines related to Humanities and Social Sciences. For example, it is very useful for analyzing the linguistic characteristics of any text type, for comparing literary texts, news or students' essays written in same or different languages, for studying the language acquisition process of children or second language learners, for creating specialized dictionaries based on real corpora, for analyzing or even reducing the complexity of the texts, etc.

In this section, we briefly explain some experiments carried out using ANALHITZA to show, as example, how it can be exploited in different tasks: *i*) a comparative analysis of two Basque literary books, *ii*) a preprocessing task for content analysis on a bilingual oral corpus and *iii*) an experiment based on n-grams to identify expressions to detect the main topic of each text in a multilingual corpus.

[11] Basque is an ergative and an agglutinative language that constructs phrases by attaching free and bound morphemes (Hualde and de Urbina, 2003).

	Arrieta (2012)		**Alberdi (2013)**	
No of pages	159		139	
No of tales	8		9	
Av. words per tale	3,210		1,974	
No of words in all	25,677		17,765	
No of diff. words	7,793	30.35%	5,739	32.30%
No of diff. lemmas	4,150	16.16%	3,041	17.11%
No of verbs	8,856	34.49%	7,291	41.04%
No of nouns	9,229	35.94%	6,284	35.37%
No of adjectives	1,914	7.45%	1,055	5.93%
No of NE	622	2.42%	382	2.15%
No of decl. words	9,212	35.87%	6,725	
Av. words per sent.	9		7	
Words in longest sent.	97		52	

Table 1: Statistics concerning both books

4.1 Comparative linguistic analysis of two literary books

Because of the linguistic information this tool offers, we consider ANALHITZA a very suitable LT for text analysis. As example, we present a pilot comparative study of two literary books in Basque: *Alter Ero* (Arrieta, 2012) and *Euli Giro* (Alberdi, 2013).[12] Both books are composed of several tales and as they have very similar external characteristics (date of publication, genre, age and place of birth of the authors), we wanted to see whether they are also linguistically similar or not (because books having similar external features can be linguistically very different).

Analyzing the resulting data (cf. Table 1), we have been able to extract some conclusions in quite a fast and easy way (Table 1). For example, Arrieta's stories are a bit longer than Alberdi's (average of 3,210 vs. 1,974 words per tale). In Alberdi's book there are a bit more different lemmas than in Arrieta's, which shows that the lexicon in Alberdi's book is more varied than in Arrieta's.

We have seen that some of the most common lemmas are not content words, which has awaken our curiosity to verify whether in Basque prose there are, in general, more function words than content words (a larger corpora must be analyzed for that). Paying attention to the nouns (which are content words), the most common nouns do not coincide, and are, in general, varied in both books. This shows that each individual tale in the two books relates a different story. However, and although a deeper study is necessary to get more precise conclusions, the most common nouns in Alberdi's book (*ama* 'mum', *esku* 'hand', *andere* 'woman', *etxe*

[12] We want to thank the Susa publishing house for making available many literary works in digital support.

'house', *gizon* 'man') give quite a clear clue about what her stories are related to.

As regards the categories of the words, we have seen that Arrieta uses less verbs and more adjectives than Alberdi, which means that his stories are more descriptive and include less actions than Alberdi's tales, where occur more actions but things are described in less detail. We have found more different NEs in Arrieta's book whereas Alberdi has repeated the same NEs more times. This can be useful to analyze, for example, whether the stories happen to same characters and in same places or not. In this case, the scenarios change from tale to tale.

ANALHITZA also extracts information about letters and declensions cases. Vowels are more frequent than consonants in both books. The average of the declined words and the most common declension cases are very similar in both books (absolutive, inessive and genitive cases are the most common ones and the ergative is a little bit higher in Arrieta's book than in Alberdi's). But are these two facts also some intrinsic characteristics of Basque or just a coincidence?

The main aim of this first analysis was to see what kind of conclusions can be obtained using ANALHITZA. The clearest differences between both books are that one is a bit more descriptive than the other one, and that one contains longer tales than the other. Both conclusions can be useful, for example, for readers or teachers when selecting or recommending a book (the most descriptive one for those who prefer less action and vice versa; and the shortest tales for those who have more difficulties on reading).

In addition, the data obtained with ANALHITZA in this task have raised new questions about the linguistic characteristics, not only of the analyzed literary works but also of Basque literature and even of our language in general, characteristics that can be additionally compared with the main features of Spanish and/or English literary works. However, we have to continue analyzing larger corpora to obtain information and get to such conclusions.

	Basque		**Spanish**	
	No	**Diff.**	**No**	**Diff.**
Sentences	113		94	
Total of words	1126		1423	
Words/lemmas	680	438	580	475
Nouns	480	229	331	185
Adjectives	86	48	129	86
Verbs	312	103	255	111
Entities	34	17	22	11

Table 2: Statistics concerning both languages

4.2 Preprocessing tasks for content analysis in a bilingual corpus composed by political texts

Our aim in this second experiment is to show whether and to what extent ANALHITZA reduces the complexity of the analyzed corpus. Complexity reduction is a necessary step before other techniques for content analysis can be implemented in a text corpus. But, to our knowledge there is not any other tool to preprocess texts of a multilingual corpus including texts written in the Basque language.

The sample for this trial is composed of 40 short argumentative texts (20 in Basque and 20 in Spanish) written by citizens in a deliberative exercise named '*Konpondu*'[13] (CICIR, 2007; CICIR, 2009). The corpus consists of open-ended responses written by participants for oral presentations. We have randomly selected the sample set among those sharing a similar length (more than 300 characters), written in a similar date (April, 2008) and responding to the same question: *In the current situation which difficulties and opportunities do you see for peace and political normalization?*; although, in different towns.

These 40 texts were analyzed with ANALHITZA in two different clusters for each language. Table 2 shows the linguistic characteristics of texts written in two languages: number of elements (No) and different elements (Diff.).

At first glance, ANALHITZA seems to be very effective in terms of data reduction for both languages, but lemmatization is more efficient in Basque than Spanish as

[13]We want to thank Aitziber Blanco and Paul Rios from *Lokarri*, Igor Ahedo and Asier Blas from *Parte-Hartuz* (UPV/EHU) and Gorka Espiau and the *Agirre Lehendakaria Center* (`http://agirrecenter.eus/`) for helping us recollecting the documentation of the '*Konpondu*' initiative.

Difficulties		Opportunities	
Prepr.	**Post.**	**Prepr.**	**Post.**
que	politico	que	vez
de	**partido**	de	oportunidad
la	**violencia**	la	tener
y	ir	a	politico
a	dificultad	para	cada
el	sociedad	las	creer
en	tener	el	querer
no	existir	y	poder
se	poder	en	política
los	parte	una	decir

Table 3: Spanish most frequent word lists

Difficulties		Opportunities	
Prepr.	**Post.**	**Prepr.**	**Post.**
eta	**alderdi**	eta	bake
ez	politiko	da	aukera
da	eta	euskal	herri
alderdi	bake	aukera	gizarte
ere	lortu	behar	euskal
bakea	arazo	dut	eman
politikoen	herritar	bakea	ikusi
dute	jarrera	ez	nahi
behar	**biolentzia**	gure	bide
beste	euskal	bat	herritar

Table 4: Basque most frequent word lists

expected. If we compare data reduction from word/lemma types to tokens, the list of words drops down until 61.11% for Basque and until 66.62% for Spanish. But reduction from word types to lemma types represents a 21.5% in Basque while 7.37% in Spanish.

Moreover, the results facilitate a more informative approximation. This can be seen in Table 3 and in Table 4, where we present the resulting word-frequency lists for Spanish and Basque respectively. Each table is divided in two main sections according to the answers respondents gave about their thoughts in two different papers reflected in the corpus: Difficulties and Opportunities. Each section of the table contains two columns reporting a list of 10 most repeated words before processing (Prepr.) and after processing (Post.) the corpus with ANALHITZA.

Results show that the most frequent word list is much more informative on the post-processing column than before the processing. We see that two words belonging to the Difficulties list are repeated in Spanish and Basque texts: *i*) *partido* (in Spanish) and *alderdi* (in Basque) meaning 'political party', and *ii*) *violencia* (in Spanish) and *biolentzia* (in Basque) meaning 'violence'. In regards the underlined term *sociedad* (meaning 'society' in Spanish), it is in the Difficulties list whereas *gizarte* (meaning 'society' in Basque) is in the Opportunities list. In case we do not preprocess, the columns of Difficulties (first column of Table 3) and Opportunities (third column of Table 3) remain more or less the same (7 words of 10 are the same and none of them are content words).

In addition, NEs provided by ANALHITZA allowed us both *i*) avoiding word ambiguity in several cases: "elkarri" (reciprocal pronoun) and "elkarri_org" (organization), or "eta" (conjunction) and *ETA* (eta_org) (organization) and *ii*) further reduction by identifying several N-grams (Jurafsky, 2009) using PoS lists: *ley_partidos* 'Law on Political Parties' or *Euskal_Herri* 'Basque Country'. Indeed, PoS lists (nouns, verbs and adjectives) could help further reduction due to the fact that other words tend to be discarded as non-informative for content analysis.

Finally, another interesting feature of ANALHITZA from a NTA perspective is the lemmatized text, since the original ordering is retained and this permits network type data extraction from the corpus.

The network maps below (Figure 2), for example, represent two clusters of words to which the term 'violence' belongs in Basque and Spanish responses to the question over Difficulties for peace and normalization.[14] Departing from lemmatized texts provided by ANALHITZA, we have extracted word co-occurrence maps. The size of each word represents the degree of connectivity in the network while links between words show the strength of ties between words in terms of number of co-occurrences. In this example, we can see that the cluster of words to which the term 'violence' belongs differs considerably between both sets. While in the Basque set 'violence' is linked to words like *politika* and *politiko* 'political', *eus_ta_ask* 'eta_org' or *gatazka* 'conflict' and *epaiketa* 'trial', in the Spanish set the cluster is formed by words like *nunca* 'never', *existir* 'exist' or *asesinato* 'killing'.

[14] ConText (`http://context.ischool.illinois.edu/`) was used to extract the co-ocurrence network and Gephi (`https://gephi.org/`) to identify clusters (implementing the modularity algorithm) and network visualization.

Figure 2: A network map of the term "violence" in Basque and Spanish datasets

4.3 Indicators of the main discourse topic in a multilingual parallel corpus

ANALHITZA has been also used to analyze the Multilingual RST Treebank (Iruskieta, Da Cunha, and Taboada, 2015) which contains 15 abstracts for each language (Basque, English and Spanish). These abstracts were published in the proceedings of the International Conference about Terminology celebrated in 1997. The corpus consists of 16,830 words and is available at `http://ixa2.si.ehu.es/rst/`.

As the corpus is annotated with rhetorical structure trees (RS-trees), we extracted the central unit (CU) of each language subcorpus and built new corpora: *a*) a corpus for each language containing only CUs and *b*) a corpus for each language, containing all the text that is not a CU. The aim of this experiment is to know how a linguist can study some word combinations (or n-grams) which indicate the CU of a text in a parallel corpus. Indeed, the detection of the CU of a text can be very useful for different NLP tasks such as question answering, summarization and sentiment analysis. To do so, we analyze in Table 5 whether the combination of the pronoun 'this' ('este' in Spanish and 'hau' in Basque) with a noun is a good indicator of CU and whether it could be used in a CU detector (Iruskieta, Labaka, and Antonio, 2016), filtering the information of n-grams in all the three languages.

Moreover, we see that in the corpus built with CUs, a noun 'N' after the pronoun 'this' is significant because we find nouns that indicate the CU in the three languages (there are other nouns with the pronoun 'this' that are not indicative of CU 17.2%), such as paper

This/este	**Lemma**_Noun	***Hau***	**Freq.**
this	paper_N		6
	artikulu_N	hau	1
this	Study_N		1
este	trabajo_N		1
	lan(txo)_N	hau	2
este	ponencia_N, comunicación_N, presentación_N		6
	hitzaldi_N, komunikazio_N	hau	7

Table 5: Pronoun and noun combinations

– *artikulu*, study – *trabajo* – *lan(txo)*, *ponencia* – *hitzaldi* ('talk' in English). We can see also that Spanish and Basque use similar words *ponencia* and *hitzaldi* 'talk', *comunicación* and *komunikazio* 'communication', *presentación* 'presentation', while in English the most used term is 'paper' in this small corpus.

Additional comparisons based on n-grams could be done in this multilingual corpus to find some collocations or to describe how definitions or examples are indicated.

5 Conclusions

In this paper we have presented ANALHITZA, an application for language processing to extract linguistic information from large corpora in Basque, English and Spanish.

The tool is being developed under the Clarin-k project, whose aim is to offer useful LT tools for Humanities and Social Sciences. As starting point in the creation of NLP based LT tools, we have carried out three experiments which have shown us that ANALHITZA is indeed a very useful and interesting tool to be applied in Humanities as well as in Social Sciences. In fact, it offers

many possibilities for research, such as the comparison of texts written by same or different authors, the comparison of texts written in different periods, genres or languages, the analysis of language acquisition process, the creation of lexicons, the reduction of text complexity, the detection of CUs, etc.

Apart from the three studies presented here, and based on all those opportunities ANALHITZA offers for text analysis, our aim is to continue working on the improvement of the tools as well as on its application and dissemination in different real scenarios such as in class assignments at the university, secondary schools and Basque language academies.

Meanwhile, we continue working in the following improvements: *i*) to extract more information from the analyzed text, such as multi-words, *ii*) to use another external tool for data visualization, *iii*) to allow analyzing other file formats (PDF files), multiple files or a ZIP file in the online application.

In addition, the tool could be improved or redesigned in the foreseeable future, in case we detect that there is some need in a specific branch of study, or if researchers ask us for that.

References

Agerri, R., J. Bermudez, and G. Rigau. 2014. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In *Proceedings of LREC 2014*, pages 3823–3828.

Agerri, R. and G. Rigau. 2016. Robust multilingual Named Entity Recognition with shallow semi-supervised features. *Artificial Intelligence*, 238:63 – 82.

Alberdi, U. 2013. *Euli giro*. Susa.

Alonso, S. and A. Volkens. 2012. *Content-analyzing political texts. A quantitative approach*, volume 47. CIS.

Arrieta, B. 2012. *Alter ero*. Susa.

Blei, D.M. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Carley, K.M. 1997. Network text analysis: The network position of concepts. In Carl W. Roberts, editor, *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*, Routledge Communication Series. pages 79–100.

CICIR. 2007. *Building Peace: the Challenge of Moving from Desire to Implementation*. Columbia University.

CICIR. 2009. *The Challenge of Moving from Desire to Implementation*. Columbia University.

Fokkens, A., A. Soroa, Z. Beloki, N. Ockeloen, G. Rigau, W.R. van Hage, and P. Vossen. 2014. NAF and GAF: Linking linguistic annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*.

Grimmer, J. and B.M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*.

Hualde, J.I. and J. Ortiz de Urbina. 2003. *A grammar of Basque*, volume 26. Walter de Gruyter.

Iruskieta, M., I. Da Cunha, and M. Taboada. 2015. A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora. *Language resources and evaluation*, 49(2):263–309.

Iruskieta, M., G. Labaka, and J.D. Antonio. 2016. Detecting the central units in two different genres and languages: a preliminary study of Brazilian Portuguese and Basque texts. *PLN*, 55(4):77–84.

Jurafsky, D. 2009. *Speech & language processing*. Pearson Education. India.

Krippendorff, K. 2004. *Content analysis: An introduction to its methodology*. Sage.

Otegi, A., N. Ezeiza, I. Goenaga, and G. Labaka. 2016. A Modular Chain of NLP Tools for Basque. In *Proceedings of the 19th International Conference on Text, Speech and Dialogue*, pages 93–100.

Villegas, M., N. Bel, C. Gonzalo, A. Moreno, and N. Simelio. 2012. Using Language Resources in Humanities research. In *LREC 2012*, pages 3284–3288.