Tagging a Corpus of Interpreted Speeches: the European Parliament Interpreting Corpus (EPIC)

Annalisa Sandrelli, Claudio Bendazzoli

Department of Interdisciplinary Studies in Translation, Languages and Cultures (SITLeC), University of Bologna at Forlì
Corso Diaz 64, 47100 Forlì (FC), Italy
Annalisa.Sandrelli@unibo.it cbendazzoli@sslmit.unibo.it

Abstract

The performance of three different taggers (*Treetagger*, *Freeling* and *GRAMPAL*) is evaluated on three different languages, i.e. English, Italian and Spanish. The materials are transcripts from the European Parliament Interpreting Corpus (EPIC), a corpus of original (source) and simultaneously interpreted (target) speeches. Owing to the oral nature of our materials and to the specific characteristics of spoken language produced in simultaneous interpreting, the chosen taggers have to deal with non-standard word order, disfluencies and other features not to be found in written language. Parts of the tagged sub-corpora were automatically extracted in order to assess the success rate achieved in tagging and lemmatisation. Errors and problems are discussed for each tagger, and conclusions are drawn regarding future developments.

1. Introduction: the EPIC corpus[1]

In a previous paper of ours (Bendazzoli et al., 2004), we outlined how we intended to create a trilingual corpus (English, Italian and Spanish) of European Parliament speeches and corresponding simultaneously interpreted versions. The basic aim of the project was to collect a significant amount of machine-readable interpreting data in the three above-mentioned languages, so as to study simultaneous interpreting "... as an activity influenced by the language pairs involved and the language direction in which interpreting is carried out" (Bendazzoli et al. 2004: 33).

Two years on, the European Parliament Interpreting Corpus (EPIC) has been created and the transcription process is continuing to further expand it.[2] The transcribed material has been POS-tagged and lemmatised by using different taggers (§2), and the tagged transcripts are encoded by using the *IMS Corpus Work Bench–CWB* (Christ, 1994). A web interface has been developed to enable users to interrogate the corpus by formulating their queries in the CQP language of *CWB* (see Web references).

The taggers used for EPIC were designed for written texts, with one notable exception, GRAMPAL, the tagger developed for the Spanish materials in the C-ORAL ROM project (see §2, and 3.5). Therefore, the aim of the present paper is to present an assessment of the accuracy rate achieved by the taggers used for the three languages and to try and identify recurring error typologies.

Before the main issue of the present paper can be tackled, some background information about the EPIC corpus is in order. Firstly, it is both a parallel and a comparable corpus,[3] since it is made up of nine sub-corpora, three of source speeches (in English, Italian and Spanish, indicated as "org" followed by the language), and six of target speeches covering all the possible combinations of the three languages (indicated as "int" followed by the source and target language). Table 1 below presents EPIC's current size and composition.

sub-corpus	n. of speeches	total word count	% of EPIC
Org-en	81	42705	25
Org-it	17	6765	4
Org-es	21	14406	8
Int-it-en	17	6708	4
Int-es-en	21	12995	7
Int-en-it	81	35765	20
Int-es-it	21	12833	7
Int-en-es	81	38066	21
Int-it-es	17	7052	4
TOTAL	357	177295	100

Table 1: EPIC's current size and composition.

Although EPIC may seem small in comparison with other corpora of spoken language, it must be highlighted that it is far larger than "traditional" (i.e. non machine-readable) collections of interpreting data. Furthermore, the material in EPIC is very homogeneous, in that owing to the institutional setting in which the debates take place (the European Parliament), a large number of variables are kept under control, including speaker characteristics, topics under discussion, interpreters' preparation and working conditions, etc. Section §1.1 briefly describes the main features of the speeches in EPIC.

1.1. Types of texts in EPIC

As was mentioned above, there are two types of speeches in EPIC, namely source speeches originally

[1] Although the present paper is the result of a joint effort, Annalisa Sandrelli wrote §1, 1.1, 3.1, 3.2 and 3.4, whereas Claudio Bendazzoli is the author of §1.2, 2, 3.3 and 3.5. The Conclusions (§4). were jointly drafted.

[2] For details of the various stages involved in the development of EPIC, see Monti et al. (2005) and Bendazzoli & Sandrelli (forthcoming).

[3] See Sandrelli & Bendazzoli forthcoming for an example of parallel and comparable analyses of EPIC, specifically in relation to lexical density.

delivered by speakers in the European Parliament, and simultaneously interpreted speeches produced by the English, Italian and Spanish interpreters working in their respective booths. The original (source) speeches in the three languages are typical examples of communication in a conference setting, in that they are monological. This type of communication is very unlike ordinary conversation, since speakers address the audience, but do not expect any direct reply, with the exception of specific, ritualized times, such as the final debate (Goffman, 1981). In the European Parliament there are strict rules for the allocation of speaking time, and as a consequence speeches tend to be very fast. Another important feature is their delivery mode, that is, their position along the orality versus literacy scale. We have decided to use three labels to classify our speeches, i.e. impromptu, read, and mixed.[4] However, it must be borne in mind that in actual fact all EP speeches are partly prepared: in this context "impromptu" does not mean wholly improvised, but simply delivered without the aid of a visible written script. Lastly, the range of topics in EPIC source speeches is wide and their degree of technicality is fairly high, even in seemingly less specialised fields such as politics.

As regards the interpreted speeches in EPIC, they cannot be considered as truly spontaneous either, since the semantic content is determined by the original speakers, whereas the formal aspects (syntactic and stylistic features) are influenced by the production conditions, i.e. simultaneous listening and speaking, with the target language (TL) speech being assembled on-line on the basis of in-coming chunks of source language (SL) speech. As Gile (1995) explains, like all multiple-task activities, successful interpreting implies efficient allocation of cognitive resources to the various components (listening, analysis, memory, and production). Increased processing capacity requirements caused by high information density, fast delivery, low sound quality, unknown names and terms, and other factors, may result in reformulation problems or even lead to saturation, with consequent omissions and errors. Typical consequences of fast delivery and high information density may include sentence planning problems (e.g. false starts), hesitations, and articulation problems (including the production of truncated words and mispronounced words). Furthermore, as a consequence of simultaneity, the duration of the interpreted speeches must resemble that of the source speeches as closely as possible.[5]

1.2. Transcription

The EPIC texts are orthographically transcribed on the basis of two main principles, namely user/annotator-friendliness and machine-readability (Leech et al., 1995). Annotation was limited to a basic set of features (Shlesinger, 1998; Armstrong, 1997), related to three different levels: extra-linguistic, linguistic and paralinguistic.

The extra-linguistic level, or metadata, comes in the form of a header at the beginning of each transcript and provides information about the speaker and the speech. Some of these extra-linguistic parameters are used as structural attributes, i.e. search filters available in the EPIC web interface (see Monti et al. 2005).

As regards the linguistic level, all the words uttered by speakers and interpreters are transcribed following the EU orthographic standards indicated in the Interinstitutional style guide. No punctuation was used, but sentence boundaries were identified by combining syntactic information and paralinguistic cues. Units were thus indicated by the double slash (//).

The paralinguistic level is limited to some disfluencies, such as mispronounced and truncated words. Mispronounced words are transcribed exactly as they are perceived after their normalised version (which can normally be inferred from the context) in order to make automatic POS-tagging possible. Systematic pause annotation with the aid of IT tools to measure the exact duration and location of pauses is beyond the present scope of our project. However, empty and filled pauses were annotated on the basis of the transcribers' perception, in an attempt to comply with the user/annotator-friendliness principle. Given the nature of our texts (see §1.1)., no turn-taking symbols are needed.

speech feature	**example**	**transcription conventions**
word truncations	propo pro posal	propo- proposal </pro_posal/>
mispronounced words	chorela	cholera </chorela/>
pauses	filled empty	ehm ...
numbers figures dates	532 4% 1997	five hundred and thirty-two four per cent nineteen ninety-nine
unintelligible		#
units		// (based on syntax and speaker's intonation)

Table 2: EPIC transcription conventions.

2. Tagging EPIC

The EPIC transcripts are tagged and lemmatised by using different taggers.

For the English language, we used the *TreeTagger* (Schmid 1994). The tagset here used is a revised version of the Penn-Treebank tagset, in which the second letter of the POS-tags used for verbs distinguishes between "be" verbs (B), "have" verbs (H) and other verbs (V).

The Italian version of the *TreeTagger* was used on our Italian speeches. The Italian tagset comprises 36 tags expressed as abbreviations in capital letters, with further specifications in small letters following a colon (eg. "VER:geru" = verb in the gerund tense).

For the Spanish language we were able to use two taggers, *Freeling* (Carreras et al. 2004) and *GRAMPAL* (Moreno, 1991; Moreno & Goñi, 1995; Moreno & Guirao, 2003; Moreno & Guirao, forthcoming).

Freeling is a suite of linguistic analysis tools, including Spanish, Catalan and English taggers. Each tag comes in the form of a rich code made up of letters and

[4] Clearly, a more fine-grained classification would have been possible, but we felt that three categories were sufficient.

[5] Interpreters try to wrap up their own speeches as soon as possible after the original speaker has finished.

numbers indicating linguistic and morphological features (e.g. depends = *depende* VMIP3S0 = main verb, indicative, present tense, third person singular).

All the taggers described so far were designed to process written texts. That is not the case of the last tagger taken into account here, namely *GRAMPAL*, developed for the Spanish part of the multilingual C-ORAL ROM project (Cresti & Moneglia, 2005) involving four Romance languages. This is one of the few examples of spontaneous speech tagging with a new tokenization system designed to cope with the specific features of spoken language (see §1.1, 1.2). On the one hand, the C-ORAL-ROM taggers were trained to recognise particular patterns typical of spoken language, such as words that are repeated in case of retracting or sequences with interruptions; on the other hand, problematic cases were treated by exploiting the probabilistic nature of languages and word order. Furthermore, the C-ORAL ROM transcripts include prosodic tagging, i.e. an indication of prosodic breaks corresponding to utterance limits (Cresti & Moneglia 2005). Two distinctive features of *GRAMPAL* are multiwords and the discourse marker tag (DM).[7] Multiwords are word units that express a single meaning and cannot be split by intervening words (e.g. in *fin_de_semana*, *es_decir*, etc.). The discourse marker tag was included in the *GRAMPAL* tagset in an attempt to better reflect the characteristics of spoken language. All the tags are abbreviations in capital letters, in some cases followed by specifications in small letters and numbers indicating morphological features.

3. Methodology and results

3.1. Methodology

In order to assess the accuracy rate of our taggers, we decided to manually check the tagging and lemmatisation of 10% of all of our units (sentences, see §1.2). Thanks to a dedicated Perl script, random units were automatically extracted to create 12 files, each accounting for 10% of corresponding sub-corpus.

As regards lemmatisation, three error categories were identified, namely wrong lemma (but correct tag), wrong lemma and wrong tag, and no lemma assigned to the token. Similarly, several categories were identified for tagging errors, namely wrong tag and wrong lemma (see above), wrong tag (but correct lemma), partially wrong tag,[8] incomplete tag and no tag assigned to the token. All the occurrences were counted and the lemmatisation and tagging success rates were thus calculated for each text collection. The separation between the sub-corpora in the same language was kept in order to verify whether there were any differences in tagger performance depending on whether the speeches had been originally delivered in that language or whether they were interpreted speeches. Finally, a number of cases were recorded under the "doubts" category, i.e. cases in which the tagger in question had correctly applied its own rules, whose formal appropriateness was debatable. For example, the English version of the *TreeTagger* (§3.2) tends to assign the proper noun (NP) tag to words beginning with a capital letter, even when that word is, in fact, an adjective, as in *United States*. Moreover, the EPIC texts often mention institutions (European Commission, Food Safety Authority, etc.), documents (Financial Services and Markets Act, Amsterdam Treaty, etc.) and job titles (President, Secretary-General, etc.) from the European institutions system or from international politics. Whether these words are to be considered as proper nouns or not is subjective, and a consistent criterion across all three languages will have to be adopted during our future training of the taggers. Table 3 below gives details of the examined materials and results.

sub-corpus & tagger	units	tokens	lem success %	tagging success %	doubts
Org-en *TTE*	171	4302	98.61	96.75	2.65
Int-it-en *TTE*	28	584	98.29	98.29	2.56
Int-es-en *TTE*	58	1412	99.3	97.81	1.91
Org-it *TTI*	26	679	97.05	93.51	2.5
Int-en-it *TTI*	161	3529	96.34	91.92	2.49
Int-es-it *TTI*	55	1459	96.5	92.04	2.12
Org-es *FRL*	56	1371	98.18	94.68	2.4
Int-it-es *FRL*	25	632	99.9	97.16	2.53
Int-en-es *FRL*	157	3589	97.83	94.21	2.67
Org-es *GRA*	56	1285	98.44	95.09	N/A
Int-it-es *GRA*	25	478	98.11	94.56	N/A
Int-en-es *GRA*	157	3309	97.67	94.49	N/A

Table 3: Results.
Legend: TreeTagger English (TTE); TreeTagger Italian (TTI); Freeling Spanish (FRL); Grampal (GRA)

3.2. TreeTagger (English)

As was mentioned in §2, the English tagset is very basic. Since it was developed to tag written texts, it is not surprising that there should not be any specific tags for pauses and truncated words. We looked at the attempts made by the tagger on these features. In most cases filled pauses (transcribed as *ehm*) were tagged as nouns. This happened whenever a pause occurred at the beginning of a unit (sentence), between an adjective and a noun, between two nouns, between a noun and a relative pronoun, between a preposition and a noun, or between two phrases. If a pause occurred between a noun and a verb, it was tagged as a verb. In all of these cases, lemmatization was attempted unsuccessfully, with no lemma being assigned to the token *ehm*. Truncated words, on the other hand, were always tagged as nouns, and no lemma was assigned (indicated as UNKNOWN). Since these problems were due to the absence of specific tags for pauses and truncated words in our tagset, these occurrences were not counted as errors. However, they certainly had an impact on the probability calculations made by the tagger in processing the following items, resulting, in some cases, in tagging errors, as in the example below in which an adjective (*dynamic*) is tagged as a noun:

[7] The latter tag is common to the Portuguese tagger as well.
[8] Partially wrong tags are those cases in which word classes are correctly assigned but other details (e.g. gender, number, verb tense, etc.) are wrong.

the	the	DT
people	people	NNS
are	be	VBP
much	much	RB
more	more	JJR
dy-	UNKNOWN	NN
dynamic	dynamic	NN

The tagger was unable to assign a lemma when it encountered certain acronyms, proper nouns of Members of Parliament and other politicians, neologisms and EU jargon words (e.g. *Europol*, *Eurobarometer*, *codecision*), foreign words (*vis-á-vis*), and rare words (*Ossetian*, *ascribed*). Strangely enough, there were also problems with certain numbers written out in full (twenty-six, ninety-seven, etc.) but not with others (sixty-four). Lemmatization errors, on the other hand, were in most cases related to tagging errors, i.e. there were very few cases in which the wrong lemma was assigned but the tagging was correct.

As regards the tagging, there were no instances of unassigned or incomplete tags in any of the English sub-corpora. The interjection tag was assigned very sparingly, with expressions such as *thank you*, *thanks*, *please*, *sorry*, *well* and *Ok* being tagged as verbs or adverbs. There were a number of cases in which the tagger was not able to disambiguate between a noun and a verb form, such as in the following example (*iron*):

to	to	TO
make	make	VV
absolutely	absolutely	RB
clear	clear	JJ
iron	iron	NN
out	out	IN
any	any	DT
interference	interference	NN

A lexical item that posed problems was *that*, used as a relative pronoun or as a conjunction. Other frequent errors involved the distinction between adjective and past participle forms (e.g. *embarrassed* vs. *trained*) and noun and gerund forms (e.g. *labelling* vs. *doing*). Finally, as was mentioned in §3.1, there seems to be some inconsistency in the application of the proper noun tag (NP): for example, all months and weekdays were (incorrectly) tagged as proper nouns.

Despite all of the above, the overall success rate of the *TreeTagger* on our English materials is fairly high (about 97%), especially in the light of the specific features of our texts. Repetitions, reformulations and even self-corrections did not greatly affect the tagger's performance in the selected texts.

3.3. TreeTagger (Italian)

The Italian version of the *TreeTagger* is very basic, as it was developed along the lines of the English version. As a consequence, a lot of potentially useful morphological information (including, for example, gender, number, verb person, etc.) is not included in the tags. This aspect could be improved to better process a morphologically rich language such as Italian.

As regards lemmatization, when the tagger could not identify a lemma (indicated as UNKNOWN), a tag was assigned anyway, almost always resulting in a tagging error. That is the case of several proper nouns, acronyms and certain shortened forms of adjectives, determiners and verbs, e.g. *gran* vs. *grande*, *tal* vs. *tale* (= such), *vien* vs. *vieni* (= come).

Filled pauses (ehm) were consistently tagged as interjections (INT), whereas empty pauses (...) were considered punctuation (PON). Truncated words were not lemmatized, but were assigned a tag which is not specific for truncated words. Therefore, the wrong tag assigned to the truncated word may, in turn, lead to incorrect tagging of the following items (see §3.2 for an example of the same problem with the English version of *TreeTagger*).

The tagging of the relative pronoun and conjunction *che* proved problematic, as indeed was the case with the tagging of the English *that* (§3.2). The Italian tagger almost always tagged it as a relative pronoun, despite the fact that *che* is actually used as a conjunction much more frequently (the same problem was found in the Spanish sub-corpora processed with the GRAMPAL tagger: see §3.4). Numbers were also problematic. They were generally treated as either adjectives or nouns, but in some cases they were not lemmatized, and some numbers were not even recognised as such. The specific tag for numerals (NUM) was actually never assigned to numbers, but it was (wrongly) assigned to two acronyms (ISAC and FAO) in the examined section of the Italian sub-corpora.

Once again, the tagging of capitalised words proved difficult. In some cases (about 2-2.5% in the material analysed), capitalised words are tagged as common nouns, but whether this is acceptable or not will depend on our methodological choice (*Unione europea*, *Patto di stabilità*). Moreover, several proper nouns are not recognised as such and are left unprocessed by the tagger (e.g. *Lisbona*, *Dayton*, etc.).

Verb forms with enclitic pronouns are tagged as verbs only, resulting in loss of information on such pronominal particles (that is not the case of the GRAMPAL tagger: see §3.5):

possa	potere	VER:cpre
evitare	evitare	VER:infi
di	di	PRE
farlo	fare	VER:infi

All possessives and demonstratives are tagged as pronouns even when they are actually used as determiners. Although this approach is questionable, such occurrences cannot be considered errors in terms of tagger evaluation, in that the *Treetagger* is consistently applying its internal rules.

3.4. Freeling (Spanish)

As was explained in §2, the tagset developed for *Freeling* looks fairly elaborate. However, the resulting tagging does provide a wealth of information in comparison with the other taggers examined in the present paper.

Lemmatization was generally very good, with an average success rate across all three Spanish sub-corpora around 98%. The rare lemmatization errors almost always involved plural forms: for unclear reasons, in some cases *Freeling* was not able to find the lemma of certain plural nouns and adjectives and left them in the plural. On the

other hand, it always attempted to lemmatize each token, so there were no instances of unprocessed lemmatization.

Turning to the tagging, once again there were no specific tags for pauses or truncated words. Filled pauses were tagged as adverbs, and lemmatized with the same word form as the token (*ehm*). Truncated words were always tagged as masculine, uncountable nouns, and the identified "lemma" was the same as the token.

The tag for interjections (UH) was assigned very rarely and inconsistently, resulting, for example, in *gracias* (= thank you) being tagged and lemmatized as a plural noun, or in *bueno* (= well) at the beginning of a sentence being tagged as an adjective. Furthermore, the tagger had problems with the lists of nouns indicating the addressees of a speech, to be found at the beginning of the latter. This structure is possibly not present in the training texts used for the tagger. In the example below, the tagger wrongly identifies *señores* as an adjective rather than a noun:

señor	señor	NCMS000
Presidente	presidente	NCMS000
Señores	señor	AQ0MP0
Miembros	miembro	NCMP000
De	de	SPS00
La	el	DA0FS0
Conferencia	conferencia	NCFS000
De	de	SPS00
Presidentes	presidente	NCMP000

The tagger was also unsuccessful when encountering certain words not present in its vocabulary, such as unknown place names (*Afganistán*, tagged as a verb, *Azerbayán*, adverb, *Tampere*, verb, etc.), people's names (*Karzai*, tagged as an adverb, *Perry*, adjective), acronyms (*ECHO*, tagged as a verb), and so on. As regards proper nouns, an important mismatch was found between the user manual and the actual tagset used, which did not include the tag for proper nouns at all. The absence of this tag resulted in all the proper nouns being tagged as common nouns. Clearly, this will have to be rectified in the training of the tagger (§3.2 and 3.3).

Numbers were often incorrectly tagged as nouns, and the relative pronoun/conjunction *que* caused a few problems, as was the case with both the English *that* and the Italian *che* (§3.2 and 3.3, respectively).

Like the Italian version of the *TreeTagger* (§3.3), *Freeling* does not recognize enclitics: in *quiero referirme a* (= I want to refer to) the second verb is simply tagged as an infinitive, and the information on the pronoun is lost.

Despite the above problems, *Freeling*'s reported accuracy rate of 95% (Carreras et al.,2004) is virtually confirmed on our sub-corpora of interpreting data.

3.5. Grampal (Spanish)

As was mentioned in §2, the Discourse Marker (MD) tag and the multiwords are specific features of the *GRAMPAL* tagger. The introduction of a discourse marker tag is certainly an interesting step towards adapting the tagger to the spoken nature of the materials to be tagged. However, in some cases it is not easy to determine whether a specific item is a discourse marker or simply a conjunction, an adverb, and so on. This ambiguity is due to the fact that, in a sense, discourse markers are on a higher level of analysis than morphology. As regards multiwords, the way in which these are tokenized is questionable in some cases, such as proper names of projects (*Espacio_de_Seguridad_y_Justicia*). In some cases, the multiword system produced an incorrect output, i.e. words pertaining to separate units were joined and labelled under the same tag (*Presidencia_de_los_Países_Bajos*) or were not joined at all. In our case, 25% of all instances of multi-words found in the sample were processed incorrectly.

The tagging success rate may vary depending on whether the percentage of incomplete tags is included in the overall error count or not (2.1% in org-es, 0.6% in int-it-es, and 2% in int-en-es, respectively). Incomplete tagging was mostly to be found for verb forms, i.e. only the V tag was assigned without any further information.

The relative pronoun vs. conjunction ambiguity on the word *que* (= that) was not tackled successfully by the tagger: indeed, *que* was tagged as relative pronoun in all cases. Moreover, some tokens were not processed at all, although the reason is unclear, as the problem affected ordinary nouns, such as *mañana*, *cara* (= tomorrow, face), and verbs, such as *tratar*, *tomar* (= to deal with, to take), and so on.

Lemmatization seemed to be less problematic. The most frequent problem was that in some cases *GRAMPAL* did not manage to identify either lemma or tag. Another common error was found with the word *gracias* (= thank you), that was never recognised as an interjection except for one case (the same problem encountered by *Freeling*, see §3.3). Pauses (both filled and empty ones) and truncated words are neither tagged nor lemmatized.

In the "partially wrong tag" category there are several cases of auxiliary verbs that are assigned the infinitive mood tag and not their own conjugation-related information. It looks like that in these cases the tagger processed the lemma and not the token.

An interesting feature of this tagger is that it is the only one among the ones considered in the present article to separate verbs and enclitics, thus tagging the two items individually, as in the following example:

tendríamos	tener	AUXcond1p
que	que	C
preguntar	preguntar	Vinf
nos	nos	PPER1p

Capitalised words are generally tagged as proper nouns, which makes it possible to correctly tag even foreign names. However, this rule attributes a proper noun status to several common nouns which were spelt with an initial capital letter following the guidelines contained in the EU Interinstitutional style guide.

4. Conclusions

As was seen above, none of the taggers we analyzed includes specific tags for filled pauses and truncated words, which are typical features of the EPIC speeches. Although the taggers react differently to such tokens, their presence has an impact on tagging results. Therefore, specific tags will have to be added to the taggers' current tagsets. Another aspect which seems to characterize our materials is the high number of capitalized nouns (referring to people, places, institutions, legislation, etc.). A decision will have to be taken regarding their status as

either proper or common nouns, and specific rules will have to be devised to enable the taggers to distinguish them. The performance of all taggers was also affected by the presence of certain features typical of spoken language, such as repetitions and ungrammatical structures (e.g. gender mismatching in adjective + noun phrases), which often resulted in partly incorrect tagging. Lists of items (nouns, verbs, or adjectives) and interjections (please, thank you, hello, etc.) also misled the taggers which could not draw on any punctuation-related information (see §3.4). A form of prosodic annotation, such as the one devised for the C-ORAL-ROM project (prosodic breaks), may be a suitable solution to the problem.

On the basis of the present study, we are going to modify the tagsets to better adapt them to our materials; then, we are going to train the taggers in an attempt to further improve their performance. The EPIC project is one of the few experiments in the creation of a POS-tagged interpreting corpus. It is hoped that our experience will serve as a basis for discussion and further research.

5. References

Amstrong, S. (1997). Corpus Based Methods for NLP and Translation Studies. *Interpreting* 2-1/2, pp. 141-162.

Bendazzoli, C. and Sandrelli, A. (forthcoming). An Approach to Corpus-Based Interpreting Studies: Developing EPIC (European Parliament Interpreting Corpus). In L., Jiang, S., Buhl, S., Bazzanella, and K., Mysak (Eds.), *Challenges of Multidimensional Translation.* Manchester: St Jerome Publishing.

Bendazzoli, C., Monti, C., Sandrelli, A., Russo, M., Baroni, M., Bernardini, S., Mack, G., Ballardini, E. and Mead, P. (2004). Towards the Creation of an Electronic Corpus to Study Directionality in Simultaneous Interpreting. In N. Oostdijk, G. Kristoffersen, and G. Sampson (Eds.), *Compiling and Processing Spoken Language Corpora*, LREC 2004 Satellite Workshop, Fourth International Conference on Language Resources and Evaluation, 24 May 2004, pp. 33-39.

Carreras, X., Chao I., Padró, L. & Padró M. (2004) Freeling: an Open-source Suite of Language Analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*. Lisbon: ELRA (1), pp. 239-242.

Christ, O. (1994). A Modular and Flexible Architecture for an Integrated Corpus Query System. *COMPLEX '94*, Budapest.

Cresti, E. & Moneglia, M. (Eds.) (2005). *C-ORAL-ROM Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam/Philadelphia: John Benjamins.

Gile, D. (1995). *Basic Concepts and Models for Interpreter and Translator Training*. Amsterdam/Philadelphia: John Benjamins.

Goffman, E. (1981). *Forms of talk*. Oxford: Basil Blackwell.

Leech, G., Myers, G. & Thomas J. (Eds.) (1995). *Spoken English on Computer: Transcription, Mark-up and Application*. New York: Longman.

Monti, C., Bendazzoli, C., Sandrelli, A. & Russo, M. (2005). Studying Directionality in Simultaneous Interpreting through an Electronic Corpus: EPIC (European Parliament Interpreting Corpus). *Meta*, 50(4).

Moreno, A. (1991). Un modelo computacional basado en la unificación para el análisis y generación de la morfología del español. Ph.D Dissertation, Universidad Autónoma de Madrid.

Moreno, A. & Goñi, J. M. (1995). A Morphological Model and Processor for Spanish Implemented in Prolog. In M.I. Sessa & M. Alpuente Frasnedo (Eds.), *Gulp-Prode '95: Joint Conference on Declarative Programming, Marina di Vietri sul Mare, Italy, 11-14 september, 1995*. Salerno: Poligraf Press, pp. 321-331.

Moreno, A. & Guirao, J. M. (2004). Tagging a Spontaneous Speech Corpus of Spanish. In N. Nicolov, K. Bontcheva, G. Angelova & R. Mitkov (Eds.), *Recent Advances in Natural Language Processing III, Selected Papers from RANLP 2003, Borovets, Bulgaria*. Amsterdam/Philadelphia: John Benjamins.

Moreno, A. & Guirao, J. M. (forthcoming). Morpho-syntactic Tagging of the Spanish C-ORAL-ROM Corpus: Methodology, Tools and Evaluation. In Y. Kawaguchi, S. Zaima, T. Takagaki & M. Usami (Eds.), *Usage-Based Linguistic Informatics Vol.V*. Amsterdam: John Benjamins.

Sandrelli, A. & Bendazzoli, C. (forthcoming). Lexical Patterns in Simultaneous Interpreting: a Preliminary Investigation of EPIC (European Parliament Interpreting Corpus). In *Proceedings from the Corpus Linguistics Conference Series*, 1(1). On-line: www.corpus.bham.ac.uk/PCLC.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *International Conference on New Methods in Language Processing*. (44-49) Online: http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf

Shlesinger, M. (1998). Corpus-based Interpreting Studies as an Off-Shoot of Corpus-based Translation Studies. *Meta*, 43(4), 486-493.

5.1. Web references

EPIC web interface: http://sslmitdev-online.sslmit.unibo.it/corpora/corpora.php

FreeLing: http://garraf.epsevg.upc.es/freeling/

IMS Corpus Work Bench: http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

TreeTagger: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

C-ORAL-ROM: http://lablita.dit.unifi.it/coralrom/

6. Acknowledgements

We wish to thank Prof. Antonio Moreno Sandoval (University of Madrid) and Prof. José María Guirao Miras (University of Granada) for tagging the Spanish materials with *GRAMPAL* and for providing us with plentiful information about it. We also thank Dr. Marco Baroni (University of Bologna) for devising the *Perl* script we used, and more generally, for his constant and generous support. Finally, our heart-felt thanks also go to Dr. Sara Piccioni (University of Bologna) for her precious help in using the *TreeTagger*.