From EAGLES to CT tagging: a case for re-usability of resources

Manuel Barbera
SSLMIT Trieste
b.manuel@inrete.it

1. Abstract

Re-usability has been recently identified as one of the main requisites a corpus annotation project must accomplish (cf. Garside, Leech and McEnery 1997: 5; Leech - Wilson 1999, etc.). This subject, developed mainly to favour morphosyntactic tagging of corpora, naturally holds more general implications for corpus linguistics as a whole, and the need for resources to «be reusable, interchangeable, shareable» (Monachini and Calzolari 1999: 149) is now strongly agreed upon even at an institutional level. It is not by chance that international initiatives in this sense have multiplied in the last few years (cf. Monachini and Calzolari 1999: 149-150). In fact, beside the obvious economic and practical reason, there is also a more theoretic attitude toward a more "ecological" and "democratic" conception of linguistic computer sciences, where resources can be shared by fields of diverse nature.

Corpus Taurinense (CT) annotation in connection with EAGLES standard guidelines is, I believe, an excellent example of this approach from many points of view. Not only, in fact, was the CT-tagset conceived according to standards that would allow the CT to be (re)used for extremely different purposes (cf. § 4), but its very conception is an example of how experiences previously accumulated in rather diverse fields may be re-used (cf. § 3). The CT, in fact, is a POS tagged corpus of old Italian formulated for prevalently linguistic-philological aims. EAGLES standards, instead, have been introduced for eminently practical (commercial, economic, etc.) purposes. The fact that technologies designed for society can become valuable to humanistic research, in a sort of cycle and recycle of intercommunication between the two sectors, is a rather new, fortunate situation.

The present paper will present a short documentation of an example of this lucky match.

2. The Corpus Taurinense (CT).

Before fully entering into the subject, it is worthwhile dedicating a few introductory notes to the Corpus Taurinense; while it is certainly not necessary in this paper to spend time on the EAGLES guidelines, so familiar to all.

The Corpus Taurinense[1], as I have already mentioned, is a tagged corpus of old Italian texts (more specifically, old Florentine dated between 1251 and 1300) of 258,310 tokens (for 19,235 forms) which has been developed by Carla Marello and me[2]. The choice of texts was not our responsibility, however, since the CT is truly the annotated reincarnation, improved in the tokenization, of the Padua Corpus, which is a subset of the collection of texts of the TLIO (Tesoro della lingua italiana delle origini) under costruction at the OVI (Opera del vocabolario Italiano)[3]. This collection was made available by Pietro Beltrami and chosen by Lorenzo Renzi and Giampaolo Salvi (cf. Renzi 1998: 29) as the base for the compilation of "ItalAnt – Grammatica dell'italiano antico", a syntax of Old Italian which is considered an ideal prosecution of the "Grande grammatica italiana di consultazione" (Renzi - Salvi 1988, 1991, 1995).

The linguistic annotation which we have implemented is a morphosyntactic tagging, additionally enriched by lemmatic annotation. At present, we are working on the disambiguation of the transcategorizations and on the treatment of multiword entries (Barbera and Marello 2000). In the future we hope to be able to add at least a third level of annotations of textual nature.

The CT is available for UNIX/LINUX on the Corpus Work Bench (CWB) system , built by IMS Stuttgart (cf. Christ and Schulze 1996). A demo of an Internet query interface is already online and

---

1 Its name, analogously to the Padua Corpus, which will be discussed shortly, is taken from the place where the co-financed group is located, namely Torino (Turin, Italy, in Latin Augusta Taurinorum). We could not have simply called it "Corpus di Torino" because, aside from our love for Latin, there already exists a corpus of texts in English put together by students from the University of Turin which is internationally known in studies of applied linguistics as, in fact, the "Turin Corpus".

2 In the field of research co-financed by the CNR "Per una grammatica testuale dell'italiano antico", directed by Bice Morara Garavelli, and coordinated with "Ricerche linguistiche sull'italiano antico", directed by Lorenzo Renzi.

3 Cf. OVI homepage at the URLs http://www.ovisun199.csovi.fi.cnr.it/italnet/OVI or http://www.lib.uchicago.edu/efts/ARTFL/projects/OVI.

under testing at Stuttgart[4]. The CT, thanks to the versatility of the CQP (Corpus Query Processor) of the IMS Corpus Work Bench, allows for the simultaneous display and query of both linguistic (lemma, POS, morphosyntax) and philological annotations (e.g. corrections, text structure, etc.)

3. From EAGLES to the CT

In the first place, EAGLES guidelines have provided us in the planning phase of the works with many useful insights which have helped in building an efficient tagset. Monachini and Calzolari (1996) collects a lot of previous experience in this sector, which has been even more useful since corpus annotation is a branch of computational linguistics that, until now, has rarely worked on "old" corpora, so that we had few specific previous experiences[5] to base our work on.

In particular, in order to create a tagset suitable for old Italian it was necessary to keep the peculiarities of the language and of the documentation in mind as they related to the computational automatisms and the needs of linguistic analyses. Thus, various operations have become necessary. Often they were complex and would have had to be completely "invented" had the EAGLES guidelines not offered us appropriate solutions. This is not the place to discuss the details of these matters. I would, however, like to offer at least a few practical examples of this, restricting the discussion to one specific fragment of the tagset, the "pronominal" area[6].

A more general problem is the division in the tagset into Hierarchy Defining Features (HDF) and Morphosyntactic Features (MSF). Since only HDF are constructed in a typed hierarchy in this kind of architecture, while MSF are freely applied to typed tags, the general criteria hold that each class of alternative features which is POS-specific be arranged in a hierarchy and that each class of alternative features which is applied to various POS be classified as MSF. A simple comparison of EAGLES tagsets for German (ELM-DE) and for modern Italian (ELM-IT), however, reveals how these criteria can be freely modified and shaped in every language according to their requirements for lemmatization and the problems of disambiguation. The ELM-DE scheme, in fact, may seem complex, yet it was the best choice in this respect, as demonstrated by the way the IMS Stuttgart has used it.

ELM-DE

PRON	**personal**	**refl**	**poss**	**demo**	**idf**		**rel**	**interrog**	
	sg;pl		sg;pl	sg;pl	**inflect**	**non-inflect**	sg;pl	sg;pl	
	1;2;3	1;2;3			sg;pl				
	du	*mich*	*seines*	*dieser*	*mancher*	*man*	*die*	*wen*	
+MSF	+gend	+case	+gend	+gend	+gend	–	+gend	+gend	
	+case		+case	+case	+case		+case	+case	
DET			**poss**	**demo**	**idf**		**rel**	**interrog**	
			sg;pl	sg;pl	**inflect**	**non-inflect**		**inflect**	**non-inflec**
					sg;pl			sg;pl	
			seine	*dieser*	*manche*	*manch*	*dessen*	*welchen*	*wessen*
+MSF			+gend	+gend	+gend	–	–	+gend	–
			+case	+case	+case			+case	

Table 1.

For Italian, however, as demonstrated by the ELM-IT solutions, a simpler division was preferred.

4 Cf. the "CT-WWW-Demos des Corpus Query Processor CQP" homepage at the URL http://www.ims.uni-stuttgart.de/projekte/CQPDemos/italant/. Access is reserved, but it can be freely granted by asking me (b.manuel@inrete.it) or Carla Marello (marello@cisi.unito.it).

5 Moreover the Penn-Helsinki parsed corpus of Middle English (http://www.ling.upenn.edu/mideng/) and the Tycho Brahe parsed corpus of Historical Portuguese (http://www.ime.usp.br/~tycho/corpus/index.html) which are perhaps the most relevant experiences in this sector, are both treebanks, and present, therefore, problems which are often different from ours. We do, however, know of some experiments on morphological tagging of Old Italian texts at the CiBIt (Centro interuniversitario Biblioteca Italiana Telematica) in Pisa: http://cibit.humnet.unipi.it/index_ra.htm.

6 This area of the tagset was dealt with specifically in Parallela IX Congress (Barbera 2000); for a general standard description of the CT-Tagset cf. Barbera 2000/2001.

ELM-IT

PRON	**pers**			**poss**	**dem**	**indf**	**int**	**rel**	**excl**
	strg		**weak**						
	nom	**obl**	**obl**						
	io	*mi*	*me*	*mio*	*quello*	*alcuni*	*che?*	*che*	*che!*
+MSF	+pers +gend +numb	+pers +gend +numb	+pers +gend +numb	+pers +gend +numb	+gend +numb	+gend +numb	+gend +numb	+gend +numb	+gend +numb

DET	**poss**	**dem**	**indf**	**int**	**rel**	**excl**
	1,2,3					
	mio	*quello*	*alcuni*	*che?*	*che*	*che!*
+MSF	+gend +numb	+gend +numb	+gend +numb	+gend +numb	+gend +numb	+gend +numb

Table 2

Having seen the EAGLES proposals, we began to consider the idea that the scheme of the tagset, in its typed and non-typed components, could be correlated to the scheme of the lemmary, so the system could be optimized, while considering the specific difficulties Old Italian presented. These are not only due to the variations introduced by diverse philological practices used in the editions initially used, but also to the fluid and not yet prescribed nature of the original texts. The result is a staggering number of graphic and linguistic variations of all forms and the creation of multiple problems in identifying tokens, especially in relation to the presence of pronominal clitics (and less frequently adverbial clitics), particularly abundant in this state of language.

The aim was to render the annotation as distinct and suitable to old Italian as possible, while using the least number of tags necessary[7]:

CT-Tagset

PD	**pers**			**poss**		**dem**		**indf**	**int**	**rel**	**excl**
	strg		**weak**	**strg**	**weak**	**strg**	**weak**				
	nom	**obl**	**obl**								
	io	*mi*	*me*	*mio*	*÷ma*	*quello*	*ne*	*alcuni*	*che?*	*che*	*che!*
+MSF	+pers +gend +numb	+pers +gend +numb	+pers +gend +numb	+pers +gend +numb	+pers +gend +numb	+pers +gend +numb	–	+pers +gend	+pers +gend	+pers +gend	+pers +gend

Table 3

In doing so, as can be seen from the results, we have also managed to remain closer than the ELM-IT to the native "naive"[8] grammatical tradition. The modern perspective of re-usability of computer data, in fact, has underlined more than once that «it is a good idea for adnotation schemes to be based as far as possible on consensual or theory-neutral analyses of the data» (Leech 1997: 7).

From this point of view the principal novelty of the CT-Tagset, compared to ELM-IT, has been to re-organize the pronominal area (Pronouns and Determiners) in one single POS, which we have called "PD", prevalently morphologically based (leaving the syntactic level to a later and different phase), and with an internal organization studied expressly from the point of view of the tagset's structure and its suitability to Old Italian.

It was a pleasant surprise for us, as well as a confirmation that we had worked in the right direction, to find that Geoffrey Leech and Andrew Wilson, in the most recent *Guidelines*, published after the formulation of ELM-IT, had reached conclusions similar to ours:

---

[7] In fact, the computational advantages offered by a limited tagset are well-known. For example, if the tagset contains no more than 70 hierarchical tags, the annotated *corpus* will be most effective as training corpus for a stochastic annotator (cf. Heid 1998).

[8] Cf. also the linguistical notion of "concetto ingenuo" worked out by Graffi 1991.

> The parts-of-speech Pronoun, Determiner and Article heavily overlap in their formal and functional characteristics and different analyses for different languages entail separating them out in different ways. For the present purpose, we have proposed placing Pronouns and Determiners in one 'super-category', recognizing that for some descriptions it may be thought best to treat them as totally different part-of-speech. There is also an argument for subsuming Articles under Determiners. The present guidelines do not prevent such a realignment of categories, but do propose that articles (assuming they exist in a language) should always be recognized as a separate class, whether or not included within determiners. (Leech - Wilson 1999: 63-64).

They then concluded that "the requirement is that the descriptive scheme adopted should be automatically mappable into the present one" (ibidem): and our scheme certainly and easily is.

4. From CT towards the future

This brings us back to the second point of view of re-usability which we bore in mind as we planned the CT. From this other perspective, as information retrieval is concerned, since the CT tagset is fully EAGLES-conformant, information provided by the CT is fully comparable with the main modern language tagged corpora, without even exiting from the same working system environment.

That the interlinguistic comparison is so greatly facilitated is obvious. But the typological perspective is certainly not the only one to benefit from this approach.

For example, if historical linguists and historians of the Italian language have annotated corpora of Old and Modern Italian, which can be easily compared, they could verify the empirical basis of their theories more easily, or even make new discoveries. A small example of what new observations are made possible by this corpus is found in the study of multiword entries that I am pursuing. I have discovered that collocations of the structure "dalla parte mia/tua/…" with variable ending are apparently unknown to thirteenth-century Florentine: in our corpus only the type "dalla mia/tua/… parte" with variable internal element.

Last but not least, corpora annotated in the same way for diverse chronological phases of the Italian language, could be used for lexical acquisition by historical dictionaries. An elementary example could be the evidence of the change in the subcategorization frame of many verbs from the thirteenth century to the modern language. This can be only roughly studied when simple concordance programmes are used. Moreover, since the CT, through the mediation of the Padua Corpus, already constitutes in itself a reuse of the lexicographic resources from the OVI, its ability to return this favour by making a new source of information retrieval available, is another demonstration of that non vicious circle of resources from which our small consideration on re-usability began.

5. Conclusions

At the beginning of this contribution we underlined how technologies designed for society could become valuable for humanistic research. One of the great merits of computational studies, in fact, has been just that: to have built a bridge between these two worlds, which previously had relatively little contact. The example of the CT, in this way, is, I hope, even more noteworthy in that it involves historical linguistics and philology, subjects that until now have benefited from this profitable circulation of resources less then other doctrines, such as logic, applied linguistics and lexicography.

6. References

Barbera M 2000 Pronomi e determinanti nell'annotazione dell'italiano antico. La POS "PD" del Corpus Taurinense. Paper presented at *Neuntes österreisch-italienisches Linguistentreffen / Nono incontro italo-austriaco dei linguisti PARALLELA IX. "Text - Variation - Informatik / Testo - variazione - informatica", Salzburg, 1.-4. November 2000.*

Barbera M 2000/2001 Italiano antico e linguistica dei corpora: un Tagset per ItalAnt, in *VI Convegno Internazionale SILFI "Tradizione & Innovazione": La linguistica e filologia italiana alle soglie di un nuovo millennio. Gerhard-Mercator-Universität Duisburg 28.06.-02.07.2000. Atti.* (Forthcoming).

Barbera M, Marello C 1999/2001 L'annotazione morfosintattica del *Padua Corpus*: strategie adottate e problemi di acquisizione. Paper presented at *Italiano antico e* corpora *elettronici. Padova, 19-20 febbraio 1999. Incontro seminariale*. Forthcoming in *Revue romane* 36(1).

Barbera M, Marello C 2000 (Forthcoming in Revue française de linguistique appliquée.) Entrées de multimots et étiquetage de parties du discours dans le Corpus Taurinense. Paper presented at *AFLA 2000, Paris, 6-8 juillet 2000*.

Christ O, Schulze BM 1996 CWB. Corpus Work Bench, Ein flexibles und modulares Anfragesystem für Textcorpora. In Feldweg H, Hinrichs E (eds) *Lexikon und Text.* Tübingen, Niemeyer.

Garside R, Leech G, McEnery A (eds) 1997 *Corpus annotation. Linguistic information from computer*

text corpora, London - New York, Longman.
Graffi G 1991 Concetti 'ingenui' e concetti 'teorici' in sintassi. *Lingua e stile* 26: 347-363.
Heid U. 1998 Annotazione morfosintattica di corpora ed estrazione di informazioni linguistiche. Paper presented at *Annotazione morfosintattica di corpora e costruzione di banche di dati linguistici. Torino, 26-XI-1998*.
Leech G, Wilson A 1999 Standards for tagsets. In van Halteren 1999, pp. 55-80.
Monachini M, Calzolari N 1996 *Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. A common proposal and application to European languages*. Pisa, EAGLES Document EAG-CLWG-MORPHSYN/R, May 1996.
Monachini M 1996 *ELM-IT: EAGLES Specifications for Italian morphosyntax. Lexicon specifications and classification guidelines*. Pisa, EAGLES Document EAG-CLWG-ELM-IT/F, May.
Renzi L 1998 Perché una grammatica dell'italiano antico: una presentazione. In Renzi L (ed), *ITALANT: per una Grammatica dell'Italiano Antico*. Padova, Centro Stampa di Palazzo Maldura, pp 21-32.
Renzi L, Salvi G (eds) 1988, 1991, 1995 *Grande grammatica italiana di consultazione* I.-III. Bologna, Il Mulino.
Teufel S 1996 *ELM-EN. EAGLES Specifications for English morphosyntax. Draft version*. Stuttgart, EAGLES Document, July.
Teufel S, Stöckert Ch 1996 *ELM-DE. EAGLES Specification for German morphosyntax. Lexicon specification and classification guidelines*. Stuttgart, EAGLES Document EAG-CLWG-ELM-DE/F, März.
van Halteren H (ed) 1999 *Syntactic wordclass tagging*. Dordrecht - Boston - London, Kluver Academic Publishers, *Text, speech and language technology* 9.