Computational Lexicography and Carlo Emilio Gadda, *Principe dell'Analisi e Duca della Buona Cognizione*

Maria Luigia Ceccotti, Manuela Sassi

ILC-CNR
Via Moruzzi, 1 - 56100 Pisa
[luigia, manuela.sassi]@ilc.cnr.it

Abstract

The aim of this study is to stimulate interest in searching the Electronic Archive of Carlo Emilio Gadda's Works and the web site dedicated to this important Italian author in order to receive and implement suggestions, and to create new lexical resources for students, translators and readers. The article starts with a brief description of the editorial features of this Archive, that runs under the 2004 version of the Textual DataBase (DBT), and continues with the presentation of the web site, to be intended as a cultural laboratory containing texts by Gadda, bibliographical data, links and special lexicographic resources developed in three ways: by applying simple DBT functions (Hapax Legomena List, Index Locorum and Concordances), by submitting results to the semi-automatic lemmatization process (Concordances lemmatized), and by using multiple Archive outputs to be transformed into new DBT archives.

1. Background

The project for the creation of the "Electronic Archive of Carlo Emilio Gadda's Works" was started in 1994, when publisher Garzanti made their published text materials (Gadda, 1988-93), already prepared for photo-composition and for study purposes only, available to the *Istituto di Linguistica Computazionale* (ILC, Institute of Computational Linguistics).

1.1. The first version of the Corpus

The Corpus was created starting from these texts (33), appropriately encoded under DBT[1] format and by applying a mark-up to point out various phenomena (italic characters, capital words of proper nouns or decided by the author, hyphens, dates, numbers, formulas, abbreviations, acronyms, pictures, author's and editor's footnotes, dialect and foreign words, dialogue, poetic language). The first version of the Corpus, running in DBT 3.0, was presented on November 14, 1997 (the 104st anniversary of Gadda's birthday) at the CNR (National Research Centre) in Rome.

1.2. The first results

In 1997 we became the first users of this database, which is the only electronic archive existing to contain all the works of a contemporary Italian author. While the community of Gadda scholars is starting to demand something that cannot be distributed but is available for consultation at ILC – concordances -, we have started to:

- produce lexicographic tool standards by means of an automatic processing system;
- apply innovative methodologies onto sample subsystems.

2. The CEG Project

The CEG (Carlo Emilio Gadda) Project was created in 1999 with two objectives:

- to transfer the Archive into the new DBT version, thus allowing new functions to be used;
- to create a cultural laboratory containing texts by Gadda, bibliographical data, links and special lexicographical resources.

2.1. The Textual DataBase (DBT)

We will now briefly describe the peculiar features of the DBT 2004, the most authoritative heir and witness of ILC's thirty-year experience in the sector of computational lexicography. In addition to the functions existing since the first version of the programme, the most recent version of this textual database also allows for searches by lemma, by synonym, by complex family of forms and/or lemmas, by specialised words, and by descriptors. However, the DBT2004's potential for application to texts written in a given language is based on the "results" provided by two lexicographic instruments, the dictionary and the Thesaurus, for the first three types of search, and on entering specific codes into the texts examined, in the last two types of search. As regards the consultation of Gadda's archive by lemma, the dictionary and Thesaurus that are presently referred to are those of general use. However, we deem it necessary to design a project for the creation of both a Gadda dictionary and a Gadda thesaurus.

Obviously, for an author like Gadda, whose lexicon is rich of variants, archaisms, neologisms, dialect terms, etc., the results provided by the database after a search by lemma or by synonym are not always exhaustive.

2.1.1. Search by lemma

If we perform a search with the word *patire,* the first result we obtain is a list of the forms retrieved in the texts, with an indication of the number of texts of the corpus in which they appear:

patente	(10)	patenti	(5)
patì	(10)	{A}patì	(1)
patii	(1)	patir	(6)
patirà	(1)	patire	(12)
{A}patire	(1)	patirò	(1)
patirono	(1)	patisce	(7)
patisco	(1)	patiscono	(3)
patisse	(2)	patita	(9)
patite	(7)	patiti	(4)
{A}patiti	(1)	patito	(14)
{A}patito	(1)	pativa	(8)
pativano	(3)	pativo	(2)

1 CNR Patent by Eugenio Picchi, chief-researcher of the ILC.

The relevant contexts of all the forms subdivided by text can also be obtained.

2.1.2. Search by synonym

From the general use Thesaurus, searching with the synonym function, will deliver the adjectival forms labelled as synonyms of the adjective *utile*:

efficace	(11)	efficacemente	(4)
efficaci	(5)	efficacissima	(1)
efficacissime	(2)	efficacissimi	(1)
efficacissimo	(1)	valida	(11)
validamente	(4)	valide	(6)
vàlide	(2)	validi	(10)
validissime	(1)	validissimi	(1)
validissimo	(1)	valido	(8)
vantaggiosa	(2)	vantaggiose	(1)
vantaggiosi	(1)	vantaggioso	(3)
opportuna	(12)	opportunamente	(14)
opportune	(8)	opportuni	(7)
opportuno	(19)	proficua	(5)
proficuamente	(2)	proficuo	(2)
salutar	(2)	salutare	(16)
giovevole	(2)	giovevolissimo	(1)
giovevolmente	(2)		

2.1.3 Search by complex family (forms and/or lemmas)

When more complex searches by groups of words / lemmas are required, the system will give the contexts retrieved by the search for *patire* and *fame*:

1) urla, i brillanti .... e che loro hanno **patito** il freddo e la **fame** per le pere, non sa neanche lui cosa dice - RR1-CD.1.II.152.p.0612.5
2) Bestie pazze! per cui ho **patito** la **fame**, da bimbo, la **fame**! Cinquecento pesos! cinquecento: di munificenza pirobutirrica: - RR1-CD.1.III.652.p.0636.37
3) i brillanti ... e che loro hanno **patito** il freddo e la **fame** per le pere, non sa neanche lui cosa dice - RR2-AG.7.110.p.0719.37
4) Bestie pazze! per cui ho **patito** la **fame**, da bimbo, la **fame**! Cinquecento pesos! Cinquecento: di munificenza pirobutirrica: - RR2-AG.7.941.p.0741.12
5) I bisognosi e gli affamati 'a' continuarono a **patire** la loro ingenza e la loro **fame**, i ricattatori si rivolsero ad altri messeri dal passato - RR2-RAI.3.46.p.1110.17
6) un articolo rabberciato da un precedente, dovremo dire che **pativa** la **fame**? Il gioco è stato bello: e gioco sia - SGF1-SD.27.147.p.0829.24
7) pazienza, senza alcuna inspirazione. Gran debolezza fisica: patii molto la **fame**, come il solito: a cena un mestolo di - SGF2-GGP.D17.596.p.0673.19
8) altro di viveri di riserva. - Sebbene prevedessi la **fame** che avremmo **patita**, preferii le vesti ai viveri. Feci male, - SGF2-GGP.M.1615.p.0739.7
9) compie la gamba e cioè con sudare, volere, **patire**, aver **fame**, pagare l'albergo, ecc.). Così l'occhio - S(V)P-MM.1.2.X.259.p.0705.27
10) hary\... Non crederete, volevo dire, che Corvo **patisce** ... una **fame** da sognarsela \m.luisa\ Una fame? - S(V)P-HJ.1.376.p.1055.4

2.2. The web site

In 1999, when we had become more familiar with computer graphics and HTML or XML formats, we began to plan the creation of a web site on the works written by and about Gadda. Although we had not received an *ad hoc* funding, at the beginning of 2000 we started to create the CEG web site exclusively based on our know-how. The web site advertised the history of the project and made our first paper publications available online.

At present, the web site's home page presents a menu that the user can use to perform searches by:

MAP [MAPPA]
: the site map, which is regularly updated

PRESENTATION [PRESENTAZIONE]
: the presentation of the project and its objectives

REFERENCES [BIBLIOGRAFIA]
: a brief secondary bibliography

CHRONICLE [CRONOGRAFIA]
: here we store important dates (the start of a publishing project, data entry, etc.)

LINKS
: links to significant web sites or any other reference to Gadda in the web

REPORT
: our HTML introductions to Gadda-related data collections and some monographs

LOOK UP [CONSULTA]
: collections of Gadda-related data, both in XML and HTML

CONCORDANCES [CONCORDANZE]
: the traditional concordances produced up to today, in PDF

INSTRUMENTS [STRUMENTI]
: lexicographic instruments obtained by processing DBT function outputs

DEMO
: a demo of the Corpus, which can be consulted online with a DBT-Web interface built in compliance with copyright restrictions

SHORTVIEW
: slides used to present the project at several conferences.

Furthermore, a NEWS section can be directly reached from the home page where the last works completed or in progress are shown.

3. Linguistic Resources

The linguistic resources are created based on our thirty-year experience in textual data processing and on the suggestions obtained either from readings or from researchers, teachers and students who have looked at new technologies for classical studies through the intermediation of ILC.

3.1. Linguistic resources available

The first group of the listed resources has been produced in both electronic and hard copy format, while the second only exists in the electronic format.

1) Hapax Legomena Inverse Index (Ceccotti, 1998):

forme elise	(letica' - mmiezz')
-aa / -na	(baccaa - checcanzuna)
-pa / -za	(sénapa - sapienzuzza)
-b / -me	(sub - frustume)
-ne / -dre	(recane - scolopendre)
-ere / -ze	(ère - elegantuzze)
-f / -chi	(ciàaf - arciduchi)
-ghi / -oi	(esofàghi - vvoi)
-pi / -zi	(crepi - complimentuzzi)
-k / -ho	(ciciàk - serrucho)
-io / -lo	(aio - lapislazzulo)
-mo / -no	(aricordamo - novantuno)
-oo / -z	(abeto - capataz)

2) *L'accentazione in Gadda* [Accentuation in Gadda] (Ceccotti, 1999a)
3) *Un primo censimento di termini gaddiani* [A first census of Gaddian terms] (Ceccotti, 1999b)
4) *Il latino in Gadda* [The use of Latin in Gadda] (Ceccotti, 2002)
5) *Annotazioni su composti in -cola* [Annotations on compounds terminating in *-cola*] (Ceccotti, 2003)

~~~

a) Concordances by form of *La cognizione del dolore*
b) Concordances by form of *Pasticciaccio*
c) Complete concordances of the question mark (contexts with right-hand arrangement by ?)
d) Complete concordances of the question mark (contexts with left-hand arrangement by ?)
e) Complete concordances of the exclamation mark (contexts with left-hand arrangement by !)
f) Statistic co-occurrences of *Giornale di guerra e prigionia*
g) (partial) Index Locorum of Latin forms in Gadda
h) Latin forms in Horace and Gadda – Comparison table
i) Comparisons between the two versions of *Pasticciaccio*
j) Gaddian iterations
k) *System* item in Pocket Gadda Encyclopedia [http://www.arts.ed.ac.uk/italian/gadda/Pages/resources/walks/pge/cnrtsistem.html] (Ceccotti, 2002a).

3.2. Linguistic resources in progress

While in the first phase of our database usage activity we have produced lexicographic resources mainly by applying simple system functions, the method we are prevalently using at present consists in processing the results obtained with different instruments.

3.2.1. Contrastive concordances

The experience made with the construction of a comparison between the two versions of *Quer pasticciaccio brutto de via Merulana* - the first (QPL) being issued in instalments in the years 1946-47 in the *Letteratura* magazine and the second (QP) published by Garzanti in 1957 - was very useful to identify a method to perform virtually automatic comparisons between multiple text versions. These two texts stored in the archive have been isolated and used to create a sub-corpus. The DBT-Corpus table frequency function has been used to generate a table containing a first comparison between words. In the subsequent step, the forms found in both texts were ignored, while only those found in one of the two texts were taken into consideration, namely those of the second version. These data have been re-entered into the search system to extract the relevant contexts, which have been matched by highlighting the differences between them, as shown in Table 1.

QP-16	riconosceva l'interessato: «il dottor **Ingravallo** me l'aveva pur detto».
QPL-282	riconosceva l'interessato: «il dottor **Ingràvola** me l'aveva pur detto».
QP-16	fronte e delle palpebre e quel nero **pìceo** della parrucca.
QPL-282	fronte e delle palpebre e quel nero **piceo** della parrucca.
QP-17	può **stà** ssicure ch'è nu **guaio**: quacche **gliuommero**.de **sberretà** ...
QPL-283	può **sta** ssicure ch'è ca' **gguaio** …, quacche **gliommero** ... de **sbrretà** ...
QP-17	soffiate addosso a molinello (come i **sedici** venti della rosa dei venti
QPL-283	soffiate addosso a molinello (come i **16** venti della rosa dei venti
QP-17	«ch'i femmene se retroveno **addó** n'i **vuò truvà**».
QPL-283	che 'e femmene se retroveno **addo'** n'i **vuo' trovà**.
QP-17	qualche prete **più edotto dei molti danni del secolo, alcuni subalterni**, **certi** uscieri, i superiori, sostenevano che
QPL-283	qualche prete **suo conoscente**, **gli** uscieri, i superiori, sostenevano che
QP-17	come non altre ad **acclieccare** gli **sprovveduti**, gli **ignari**.
QPL-283	ma servono a **gettar la polvere** negli occhi alle genti.
QP-17	di fumare la sua mezza sigheretta, **regolarmente** spenta.
QPL-283	di fumare la sua mezza sigheretta spenta.

Table 1: QP and QPL matched contexts

3.2.2. Iterations

A study has been conducted on Gadda's corpus with the purpose of finding out all the places where the author used the technique of word repetition, a significant technique used in literature and, in particular, in twentieth-century literature. This research, which is still ongoing, has been carried out using DBT functions. The data obtained have been further processed and presented in the web site (see 3.1 j). Two distinct results have been obtained: in the first case, we collected the pairs of words repeated in a sequence, while in the second case we took the iterations of words separated by punctuation (in most cases) or by other words. Table 2 shows a graph with the most frequent iterated forms (from 56 to 14 over a total of 2631 iterations in 1,632,597 occurrences).

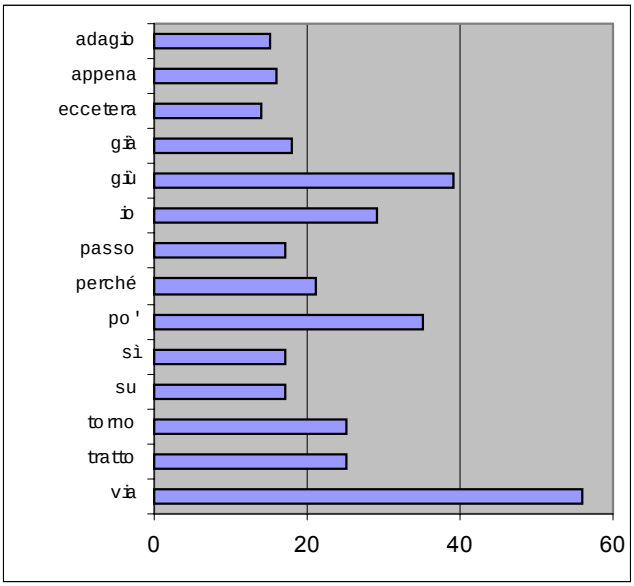

Table 2: Most frequent iterations
~~~

3.3. The future of linguistic resources

Many more data have been collected and are currently available, and we hope we will be soon able to add the resources produced by the application of the tagger to the whole archive in our web site. In particular, we are extracting adjectives, which are numerous and sometimes neologisms created by Gadda himself, and represent one of the most interesting testimonies of the lexicon of this important author of the Italian twentieth-century literature. In addition to the objective of recording the lexical universe of the author, we also aim at creating supports that may be useful, for instance, for translators and teachers. All the co-occurring adjectives for the substantive *prosa* have been extracted:

acre	antonginiana	aulica
avanzata	bellica	bettiana
brillante	buona	buona
burocratica	chiara	classica
commovente	comune	culturale-media
curiale	deformatrice	discorsiva
distratta	dura	elegante
elegantissima	eloquente	evasiva
fluente	formidabile	frantumata
generica	giornalistica	girovagante
icastica	imperfetta	imprecisa
inarrivabile	incollata	insinuante
insuperata	inutile	irreprensibile
italiana	italiana	leggera
magnifica	manzoniana	media
meravigliosa	narrativa	narrative
nervosa	ottima	ottimistica
pacata	pacchianotta	peciona
peggiore	perfetta	pessima
piscatoria	pneumatica	polposa
poltigliosa	potente	povera
preferibile	pronta	robusta
sapida	sàpida	scritta
secca	seria	solida
sorvegliata	squisita	tecnica
tersa	tipo	toscana
vigorosa		

We are presently planning to create conceptual structures and apply them to Gadda's works through the DBT. But some explanation is necessary in this regard. The DBT statistic co-occurrence function allows us both to capture all the pairs of words used in a text at a maximum pre-established distance and to build syntagms, that is to say related semantic aggregations, starting from a given lemma. These results can then be used to build and check thematic trees to be applied to the texts through the conceptual structure search function. More functions available for this purpose are: sequential search, analysis of prepositions and repeated segments. This is a phase where linguistic resources are being created to build the infrastructure required for the TAL (Automatic Language Treatment).

4. Conclusion

We would like to conclude by quoting Gadda himself, in a passage that seems to us to be appropriate to comment upon the presentation of this contribution of ours to a conference where analysis and good cognition are important reference parameters: "Com'è noto sono stato recentemente insignito del titolo nobiliare ereditario di Principe dell'Analisi e di Duca della Buona Cognizione per motivi che tutti conoscono e con un rituale che sarà oggetto d'un particolare e distinto referto. L'impresa del Principe dell'Analisi comporta il motto bruniano (della Cena delle ceneri): «Umbra profunda sumus» ed è questa secondo alcuni la più dolorosa conclusione alla quale possa pervenire un indagatore. Essa però non deve interpretarsi in senso scettico: una constatazione d'ombra non chiude la dolorosa fatica del minatore. Così come i vincoli o legami non impedirono a Michelangiolo di esprimere il suo Schiavo legato." (da *La casa*, Racconti incompiuti, RR2, p.1127; Gadda, 1989)

"As is well-known, I have recently been conferred the hereditary noble title of Prince of Analysis and Duke of Good Cognition for reasons that everybody knows and with a ritual that will be the subject of a special and distinct report. The deeds of the Prince of Analysis involve Bruno's motto (in the *Cena delle ceneri*): «Umbra profunda sumus» and this is, according to some, the most painful conclusion that could be reached by an investigator. But it should not be interpreted as a sceptical statement: the ascertainment of a shadow will not close the painful fatigue of the miner. Similarly, bonds or ties did not prevent Michelangiolo to express his tied Slave."

References

Ceccotti, M.L. & Sassi, M. (1998). Apax in Gadda – Un Indice Inverso, ILC-CNR, S.T.A.R., Pisa.

Ceccotti, M.L. & Sassi, M. (1999a). Forme accentate in Gadda. Un index locorum, ILC-CNR, S.T.A.R., Pisa.

Ceccotti, M.L. & Sassi, M. (1999b). Alla ricerca dei termini gaddiani. Una pre-concordanza, ILC-CNR, S.T.A.R., Pisa.

Ceccotti, M.L. & Sassi, M. (2002a). Sistema. *In A Pocket Gadda Encyclopedia*, Edited by Pedriali F.G., EJGS 2/2002. EJGS Supplement no. 1, ISSN 1476-9859.

Ceccotti, M.L. & Sassi, M. (2002b) La cultura latina in C.E. Gadda, ILC-CNR, S.T.A.R., Pisa.

Ceccotti, M.L. & Sassi, M. (2003) L'Archivio Elettronico delle Opere di Carlo Emilio Gadda. Da redattori a fruitori di un data base testuale, in *Linguistica Computazionale*, Special Issue (Zampolli A., Calzolari N., Cignoni L. Editors), Istituti Editoriali Poligrafici Internazionali, Pisa-Roma, pp.221--250.

Gadda, C.E. (1988). Romanzi e Racconti I, Collana I Libri della Spiga, Garzanti, Milano.

Gadda, C.E. (1989). Romanzi e Racconti II, Collana I Libri della Spiga, Garzanti, Milano, 1989.

Gadda, C.E. (1991). Saggi Giornali Favole I, Collana I Libri della Spiga, Garzanti, Milano, 1991.

Gadda, C.E. (1992). Saggi Giornali Favole II, Collana I Libri della Spiga, Garzanti, Milano, 1992.

Gadda, C.E. (1993). Scritti vari e postumi, Collana I Libri della Spiga, Garzanti, Milano, 1993.

Gadda, C.E. (1993). Bibliografia e Indici, D. Isella, G. Lucchini e L. Orlando (a cura di), Collana I Libri della Spiga, Garzanti, Milano, 1993.

Picchi, E. (1997). DBT 3 - Data Base Testuale: Guida all'uso, versione 3.1. Lexis Ricerche s.r.l. su licenza del C.N.R., Roma.