language. The corpus evidence largely supports my position, suggesting at least that the corpus is a good sample of the kind of English that seems natural to me. There are, however, many kinds of English, and an area like the titles of Societies may be more influenced by the globalisation of English than others. Where English is an instrument of communication for millions of speakers of many languages, we cannot expect the niceties of Table 2 to be maintained, and perhaps the *of* structure – already numerically predominant – will become acceptable for all cases.

Acknowledgement

All the instances quoted above are taken from *The Bank of English* at the University of Birmingham. I gratefully acknowledge the important contribution to scholarship made by the University and Cobuild Ltd in maintaining this invaluable resource for research purposes.

REMA ROSSINI FAVRETTI

Corpus Linguistics in Italian Studies

1. Introduction

The English language has long been the focus of enquiry in corpus linguistics. In the last few years interest in the field of corpus construction and analysis has spread to a large number of languages. Today it is possible to refer to an increasing number of large computer-based text corpora, available in different languages, and several projects are emerging at an international level: in Italy a general corpus of written Italian – CORIS – has been under construction at the Centre for Theoretical and Applied Linguistics of Bologna University (CILTA) since 1998 and is available on-line. CORIS contains 100 million running words and will be updated every two years by means of a built-in monitor corpus. It consists of a collection of authentic texts in electronic form chosen by virtue of their representativeness of written Italian.

Besides the defined model, a dynamic model (CODIS) was designed, which allows the selection of subcorpora pertinent to specific research and also the size of every single subcorpus, in order to adapt the corpus structure to different comparative needs.

The description of the construction of CORIS/CODIS will be the aim of the first part of the paper. In the second part, some concordances from CORIS will be considered to exemplify the contribution which can be given to Italian studies from corpus evidence and to highlight new insights which may emerge in linguistic analysis.

2. CORIS/CODIS development

2.1. CORIS design and construction

In order to design and construct CORIS, some preliminary choices were necessary to lay the foundations for successive stages. First of all we[1] defined the aim of the project, and the type of corpus it was intended to create. From the very beginning the aim of the project was to construct a general reference corpus and, at the design stage, to create a collection of texts in electronic format representing, in the widest sense, present-day Italian. The identification of this aim provided a solution to one of the first issues which arose in the planning of the corpus, the choice between synchronic and diachronic dimensions. It was decided to select texts synchronically in order to permit a generalised description of commonly used Italian.

2.1.1. Background

In the choice between written or spoken and written language, priority was given to written texts at this stage of research. The decision was based both on external and internal criteria. First of all, it was influenced by the general panorama of Italian linguistics and the position that the corpus would occupy alongside works such as the *Lessico di frequenza dell'Italiano Parlato* (LIP, 1993), *Lessico di frequenza della lingua italiana contemporanea* (LIF, 1972), *Vocabolario elettronico della lingua italiana* (VELI, 1989) and *Letteratura Italiana Zanichelli in cd-rom* (LIZ, ^{1}1993, ^{2}1995, ^{3}1997) to name just the most significant. Mention should also be made of the corpus of written Italian (IRC)[2] developed at the Institute of Computational Linguistics in Pisa as part of the PAROLE project (1996-1998) as well as of the LABLITA corpora of spoken Italian. Secondly, in the light of transformations in communication technologies, it was preferred not to pose the problem of the relationship between the language traditionally considered as standard spoken Italian and its technological developments via telephone, radio, television and/or computer technology. On the basis of these considerations, it was therefore decided to develop a synchronic corpus of written language, selecting texts dated, with some approximation, from the 1980s and 1990s, with a somewhat longer timescale as far as fiction is concerned.

2.1.2. Size of the corpus

The definition of the size was more problematic. On the one hand, a survey of currently available corpora clearly revealed that it was not possible to make reference to any standard size. On the other, the criteria at the basis of first-generation corpora, such as the *Brown Corpus*, which were mainly influenced by the potentiality of information technology, appeared to be no longer valid. Developments in information technology over the past years, the present speed of the processing of material and the low cost of mass storage units make it possible to create corpora consisting of hundreds of millions of words, such as the *British National Corpus* and the *Bank of English*.

It would seem that, as far as written language is concerned, the standard of one million words has given way to a standard of one hundred million. But any generalisation is debatable, as is any definition of a set limit. The *Brown Corpus* (1967), with one million words, with 500 written text samples of 2000 words each, representing in equal measure the main text types, is still considered a valid model. One of the most recent and appreciated English language corpora, the *Longman Spoken and Written English Corpus* (Biber *et al.*) consists of about 40,000,000 words and contains 37,244 texts. The structure and proportions among subcorpora also vary considerably (Table 1).

Bearing in mind the inevitable discrepancies, the size of the corpus, though 'large', was not predetermined but related to the choice of linguistic varieties thought to be representative and, as such, set as an intermediate research goal following the construction of a pilot corpus. The choice was made easier by the new situation created by the introduction of monitor corpora. Through the implementation of the monitor corpus, the aspects of determinacy and permanence which were defining characteristics of the size of a corpus over the past decades seem to lose their relevance. The corpus takes on a dynamic configuration, which seems all the more relevant and advantageous in

[1] The planning and coordination of the project was carried out by R.Rossini Favretti, the software was designed by F. Tamburini.

[2] The corpus contains some 16,000,000 words and is composed primarily of journalism and fiction.

relation to the development of new technology and memory, which make it possible to manage a corpus the principal components of which are delimited. At the same time, a monitor corpus is open and able to record innovations and modifications in current usage. This combination permits to access a corpus which is available in a finite form – either on-line or on CD-Rom – and can be updated by means of the monitor as well as by the introduction of supplementary subcorpora representing further varieties.

Corpus	Composition			
BNC - 90Mw	Books	52.5 Mw	-	58.6%
English	Press	27.8 Mw	-	31 %
Written section	Miscellanea	7.4 Mw	-	8.3 %
LSWE - 28Mw	Fiction	5 Mw	-	17.8%
English	News	10.6 Mw	-	37.7%
Written section	Academic Prose	5.3 Mw	-	19 %
	General Prose	6.9 Mw	-	24.6%
The Oslo Corpus - 22.3 Mw	Fiction	3.8 Mw	-	17 %
Norwegian	Newspaper/Magazine	10.6 Mw	-	47.5 %
	Factual prose	7.8 Mw	-	35 %
Corpus de Referência do	Newspaper	55 Mw	-	60.8%
Português Contemporâneo	Books	20.5 Mw	-	22.6%
(CRPC) - 92 Mw	Periodical	7 Mw	-	7.7%
Portuguese	Decisions of Supreme			
Written section	Court of Justice	1.8 Mw	-	2 %
	Miscellanea	3.9 Mw	-	4.3%
	Leaflets	0.3 Mw	-	0.3%
	Correspondence	0.1 Mw	-	0.1%

Table 1. The composition of some reference corpora.

2.1.3. Representativeness

A later step was the definition of representative linguistic varieties. CORIS was intended to represent a wide range of varieties in written Italian, mainly press, fiction, academic, legal and administrative prose, ephemera, and miscellanea. It was designed to create a 'balanced' and 'representative' general reference corpus, easily accessible and user-friendly. It is to be noted, though, that representativeness, a crucial aspect in the creation of a corpus, is still one of the most controversial issues in corpus linguistics studies[3] and has been the subject of considerable methodological discussion. There is a certain ambiguity inherent in its use in particular as a consequence of the interplay of quantitative and qualitative connotations. In some studies the extension of corpora to include hundreds of millions of words is seen as making up for a slight differentiation in the varieties represented, in others a wide differentiation in varieties is seen as an essential condition for any act of generalisation.

Even in the first phase of our research the problem of representativeness did not disappear with the possibility of enlarging the corpus. In spite of the increase in size to hundreds of millions of words, each corpus represents a limited sample of language in use. An operation of sampling, however extensive it may be, inevitably turns out to be simplified in the light of the complexity of the phenomenon under examination. Even building random selections into the corpus construction, in the transition from the sample to the generalisation, certain degrees of approximation were considered inevitable, while trying to identify parameters which might counterbalance them

When defining the selection and construction criteria, reference was made to both external and internal parameters in order to reduce the researcher's choice to a minimum. Furthermore, considering the context of CORIS as well as the wide availability of existing and planned corpora, a further criterion was introduced, that of 'comparability'. This led to the identification of the initial construction phase – provided by the subcorpora – in which it was possible to refer to some macro-varieties identified on the basis of external appearance or the material elements of the text, extremely clear in their characterisation and easily comparable. These are considered as a collection of documents identifiable on the basis of both external and internal features[4] in which the peculiarity of the single variety fades away in comparison to the mass of data. Although the corpus included specialist areas, such as legal, scientific and

[3] See in particular the papers by Biber (1993) and Varadi (2001).

[4] As the distinction between 'published' and 'unpublished' texts was considered to be too simple, various kinds of publications from various types of volumes and essays were then selected, and various handwritten, printed and above all electronic texts were grouped together in a section under the heading of 'ephemera'.

administrative language, an attempt was made to bring together not so much a collection of specialist texts as a variety of types which, according to our investigations, can be placed along a continuum, overlapping and integrating one another.

Having defined some basic macro-varieties, it was thought necessary to apply a second level of articulation (based on the sections which could be divided into subsections) which, again using external parameters as a basis, still enabled collected data to be contextualised. It was clear that a sampling of the 'press' population could not be undertaken except on the basis of a second phase connected to the socio-cultural reality of the nation. This was considered to be a fundamental point in order to reach a definition of a population's components, albeit with some degree of approximation. The choice of these parameters led to the following structure.

Subcorpus	Sections	Subsections
PRESS	newspapers, periodic, supplement	national, local, specialist, non-specialist, connotated, non-connotated
FICTION	novels, short stories	Italian, foreign, crime, science fiction, women's literature, for adults, for children, adventure
ACADEMIC PROSE	human sciences, natural sciences, physics, experimental sciences	books, reviews, scientific, popular history, philosophy literary criticism, arts, law, economy, biology, etc.
LEGAL AND ADMINISTRATIVE PROSE	books, reviews	legal, bureaucratic, administrative
MISCELLANEA	books, reviews	books on religion, travel, cookery, hobbies, etc.
EPHEMERA	letters, leaflets, instructions	private, public, printed form, electronic form

Table 2. CORIS corpus structure.

As a following step, texts for the entry of the single subcorpora were prepared and, in order to comply with the basic criterion of representativeness, the documents were randomized within each subcorpus,[5] to minimize any bias or skewing in the data. The macro-varieties in Table 3 were included.

	Size of Subcorpora (Mw)
PRESS	38
FICTION	25
ACADEMIC PROSE	12
LEGAL AND ADMINISTRATIVE PROSE	10
MISCELLANEA	10
EPHEMERA	5

Table 3. Sizes of the main macro-varieties in CORIS.

2.2. CODIS design and construction

Considering the vital role which would be played by the comparability of a general reference corpus, such as CORIS, it seemed important to provide for an alternative corpus structure which would make it adaptable to the needs of different researchers.

To deal with the comparability issue in CORIS a further corpus – CODIS – was designed. Aimed at specialist needs arising in the context of interlinguistic analysis, CODIS presents a dynamic and adaptive structure that allows the selection of the subcorpora which are pertinent to a specific research project and also the size of every single subcorpus. CODIS is designed to be dynamically adapted to different comparative needs.

As shown in Table 4, each CORIS subcorpus was split into four parts of different sizes. The sizes were carefully selected in order to allow, by means of various combinations, the creation of subcorpora of virtually any size. For example the subcorpus *Miscellanea* can be constructed from 0, 1, 2, 3 (2+1), 4, 5 (4+1), 6 (4+2), 7 (4+2+1), 8 (4+2+1+1), 9 (4+3+2), 10 (4+3+2+1) million words. This fine granularity creates an extremely flexible corpus structure that can be adapted to almost any possible comparison with other reference corpora in different languages.

[5] For a more detailed illustration, see Rossini Favretti (2000).

Subcorpus	*User-selectable sizes (Mw)*			
Press	20	10	5	3
Fiction	13	7	3	2
Academic Prose	5	4	2	1
Legal and administrative prose	4	3	2	1
Miscellanea	4	3	2	1
Ephemera	2	1	1	1

Table 4. CODIS user selectable subcorpora and their sizes.

3. Applications and case studies

3.1. Application potential of CORIS

The CORIS/CODIS running corpus constitutes a source data for a large number of research projects and it is aimed at a wide spectrum of potential users, from Italian language scholars to Italian and international students engaged in linguistic analysis driven by or based on authentic data.

The empirical evidence that authentic data on a large scale can provide is an important new resource for research into contemporary written Italian. Particularly in this moment of considerable interest in methodological issues, problems may arise as to the way in which corpus data are used in linguistic analysis and as to the status which is assigned to the data and to the evidence that corpus data provide. Their role in language description may be seen as supplementary or, as I would argue, fundamental. Different approaches may be identified, but any dichotomy may, at the moment, be reconciled in consideration of the number of issues posed by the unprecedented evidence which is provided and of the new kind of language research which is made possible.

In my opinion, the quantity of authentic data which can be processed and examined and the brevity of query time may be considered as the strong points of this approach, which is firmly rooted in the tradition of structural linguistics and makes it possible to reconsider some of the issues in the fresh light of hardware and software developments opening new perspectives for enquiry (Rossini Favretti 2001).

It goes beyond the limits of this brief overview to investigate the possible applications of corpora such as CORIS to theory and practice, but it would seem appropriate to discuss some of the main issues by means of some examples.

3.2. Case studies

The findings shown by the concordances can be illustrated considering, by way of example, some instances of the word *grado* in the singular and the plural form. The word *grado* presents 21,577 occurrences.

```
contenuto potenzialmente il grado superiore ; per il second
 queste proscimmie erano in grado di localizzare le prede e
genti in uno Stato messo in grado di coordinare e catalizza
o che le sue foglie sono in grado di migliorare l ' apporto
 livello di rischio si è in grado di sopportare . I portafo
ieduto da Chicco Testa è in grado di stendere su dodicimila
n può certo considerarsi in grado di giudicare le sottiglie
care che il referendum è in grado di fare irruzione da un m
re il premier sembra più in grado di controllare le consegu
icolo3 , paragrafo 2 , è in grado di provvedere al coordina
o viene nuovamente messa in grado di comperare il prodotto
lities " si è dimostrato in grado di proteggere dal virus .
 la dei paesi africani è in grado di offrire un servizio qu
anismo primordiale non è in grado di evolversi naturalmente
 visto che non siamo più in grado di affrontare questo giga
condensazione o comunque in grado di risparmiare . Acqua ca
organizzato dev ' essere in grado di utilizzare e sprigiona
 si riconosce o non si è in grado di gestirlo . Tabella 3 -
 nelle conclusioni di primo grado ) a " pronunce e leggi "
92 ( v . dal 44 65 fascic I grado ) . Da altresì atto , ess
```

A quick scan through the concordance, selected automatically, will suffice to identify a dominating pattern, in which *grado*, as a node, collocates with *in* in N-1 position and with *di* in N+1 position. *In grado di* occurs in the corpus in 14747 instances and it appears to function as a discrete unit of meaning, of which a relevant colligational feature may be found in the association with the infinitive from immediately to the right and with a form of *be* on the left. Even if the information is not given in most dictionaries, the phrase, or better the 'lexical phrase' (Sinclair 1996, Hunston and Francis 2000: 8) as it is often referred to, is dominant (14,747 instances) and it shows an almost constantly positive prosody.

Other strong collocates of *grado*, in N-1 position, are ordinal numbers with particular reference to legal proceedings. It is interesting to observe that also the word *gradi* strongly collocates with numbers.

```
raneamente , a più di 1.000 gradi : ne risulta un prodotto
l Sole si scalda fino a 120 gradi sopra zero \ C . La vita
saggio da 300 gradi K a 150 gradi K , e , se ci si riusciss
i situati in posizione 19,2 gradi Est , da tempo in grado d
' eclittica è di circa 23,5 gradi . La fattura , decisament
a diurna massima di meno 29 gradi C . Sembrava che non vi f
 oscillava tra i 32 ed i 34 gradi . Niente di eccezionale m
uperficie del pianeta a 360 gradi . Ancora una volta l ' Uo
lusso di aria rovente a 400 gradi fa evaporare i residui di
ratorio viene ruotato di 90 gradi in senso antiorario . L '
di seconda classe , due dei gradi al vertice della carriera
pollenza di alcuni titoli e gradi , esami integrativi , l '
sicologica , di più elevati gradi di affettività e , non da
sizioni terrestri secondo i gradi di latitudine e di longit
ta nuvola rovente di 5 mila gradi . Il nostro pianeta torne
e in un intervallo di pochi gradi . Lo stesso per quanto ri
eratura è scesa a circa tre gradi sopra lo zero assoluto (
orazione a trecentosessanta gradi dell ' universo neorealis
a luce . Segnava ventisette gradi , la stessa temperatura d
zero gradi Celsius . A zero gradi la bolla emette circa 10
```

However, the plural form, unlike the singular, is preceded by cardinal numbers. Going through the concordances, we see that *gradi* usually refers to temperature, latitude, longitude, inclination angles and so on. The singular and plural form appear to differ not only in their collocation, but also in their contextual association. As argued by Sinclair in his seminal work on "extended units of meaning", words "enter into meaningful relations with other words around them" (1996: 76) and do not remain (1996: 82-3)

> perpetually independent in their patterning unless they are either very rare or specially protected [...]. Otherwise, they begin to retain traces of repeated events in their usage, and expectations of events such as collocations arise. This leads to greater regularity of collocation and this in turn offers a platform for specialisation of meaning, for example in compounds. Beyond compounds we can see lexical phrases form, phrases which are to be taken as wholes in their contexts for their distinctive meaning to emerge [...].

Words can be observed in the multiplicity of their combinations both on the syntagmatic and the paradimatic axes. The ways in which, in Firthian terms. they are 'mutually expectant and mutually apprehended' can be highlighted. There are a number of points that may arise in this connection. Some of them go beyond the limited purposes of this analysis; others, like the combinatorial relationships holding between syntagmatically and paradigmatically related units, can be briefly discussed through the observation of some examples. To begin with we will consider a sample of the forms of two words which are presented as synonyms in a number of dictionaries: *sapere* and *conoscere*.

```
s con incredulità ; come  sapevo    , non credeva alle mie c
me questa , quando non    sapete    affrontare una crisi se
 chiamarci , se solo    sapessimo   ascoltarli , se solo lui
zione . Gli scienziati    sanno     bene che lo stato attual
a meravigliarsi . Io non    so      bene come vengano fatte
' unica a farmi pena .    Sapevo    che il marito la tradiva
 per arrivare a rete e    sapeva    che , in questo modo , a
nsiglia al timido ( " Lo    so      che sei un personaggio t
e gli investigatori ,    sapevano   chi era stato ad uccider
ate a scuola insieme -    seppi     che era andata come inse
rava in crisi , ma non    sapeva    come sostituirlo . Capri
? " Nel senso che io non   so       cosa ha in mente D ' Ale
e s ' allontana . Io non   so       dove siamo né dove andia
lta che ne parlo , e non   so       neppure se dovrei , bene
endo da est o da ovest ,   sa       come nasca un pensiero s
mpo di morire ~ . Non si   sa       quanti bambini illegitti
po ' fu tranquilla . Non   so       quanto può essere durata
i luoghi che indicherete saprebbero ricompensarvi per la men
 nella tua mente . Non   sapevo     se prima o poi avresti c
ce come una sferzata ?   Seppi      subito che quella era la
```

The concordance lines above are a limited sample automatically reduced from the total of 16,540 instances present in CORIS. The citations, alphabetically ordered on the right, show that in a large number of cases the forms of *sapere* are followed by a complement clause either explicit or implicit. In a limited number of cases it is followed by an infinitive form.

Let us now consider *conoscere*. A quick analysis of the concordance shows that the forms of *conoscere* usually collocate with noun phrases, in N+1 or N+2 position, or with pronouns in N-1 position. They may be formed by proper names or common nouns, both animate and inanimate, preceded by definite or indefinite articles.

There is no concordance in which a form collocates with a subordinate clause.

```
notizie che ancora non   conosceva    , per cui riteneva che
he ebbero occasione di   conoscere    David , nessuna ebbe a
e errore di non averlo  conosciuto    a fondo , di non esser
va il suo direttore le   conosceva    a memoria : cene meste
 era inevitabile . No ,   conosco     a sufficienza Louise A
berg ( Germania ) * Non   conosco     abbastanza la situazio
go . Ora , anche se non   conosco     ancora bene il dettagl
olta confusione e non   conoscevamo   bene la zona . - Dov '
loda all ' improvviso .   Conosco     donne che dopo aver de
orriso etrusco ? Perché   conosco     due sorrisi : quello d
ritico verso di lei . "   Conosco     e stimo Amato , con il
 il tipo di ragazza che   conosce     i nomi dei fiori di ca
rima dell ' Accademia .   Conobbe     i surrealisti ma li tr
 seminario . Perciò non   conosco     il latino . " Mi parve
zioni telefoniche . Non   conosco     il contenuto della man
e in Lombardia . Non ne   conosco     la formula ma trovo ch
 E anche se non avesse   conosciuto   le Sacre Scritture , c
suno dei marocchini che   conosco     lo fa più . Credo che
ie della riserva che io   conosco     molto bene , essendoci
tto inaudita , non si   conoscessero  regole giuridicamente
```

Considering the collocational patterns and the co-textual relations of *sapere* and *conoscere* we can see that synonymy is limited to some fixed phrases. It is apparent that the combination and co-selection of words are based on lexical as well on grammatical features.

I should like to make another point and focus attention on the insights that, through corpus evidence, can be gained of linguistic data. Access is simultaneously given to a word – the node – and to its collocational profile and colligational patterns. At the same time the possibilities of combination and selection of a word are highlighted. In my opinion the contribution that can come to language description by the simultaneity of analysis has often been underrated. This may be a challenge to our studies and highlight new approaches to language descriptions.

Many examples could be given. To conclude, it may be interesting to focus attention on some Italian words such as, for example, *voler*, *saper*, *poter*, *batter*, *uscir*, *ciel*, *amor*, *affar*, and so on. Apocope has not been the subject of detailed analysis. I do not mean that it has been overlooked, but that its occurrence has hardly ever been associated with the presence of phraseological patterns or

extended units of meaning.[6] The phenomenon has not generally been related to the environment where the word occurs or referred to the syntagmatic relations that the word holds with its co-text.[7]. As an example, the short concordance below illustrates the restricted range of environment in which the apocoped word *affar* occurs. In CORIS *affar* has just 83 occurrences. In almost all of the instances – 80 out of 83 – the node is followed by a possessive form (*mio, tuo, suo, nostro vostro, loro*) and is not preceded by a definite article in N-1. In the majority of the instances – 79 – the node is preceded – in N-1 or N-2 position – by a form of the verb *essere* (*è, era, sono, erano, fosse*). The full form *affare* (1,699 instances in CORIS), far more common than the apocoped one, is associated with a possessive form in a very limited number of cases (15 instances, 0,8%). The analysis of various instances would suggest a correlation between apocope and the conceptual relations holding in the noun phrase or in the verb phrase.

```
 ava Pete e Paul , non era affar loro . " Pete , ritira que
 uell' orologio ? Quello è affar mio , parliamo di quello c
   che dopo , non sarà più affar mio . Si guardi intorno .
glia di ballare ballo ed è affar mio e se vuoi scoprire il
io non dico niente , non è affar mio , e se questo dà loro
a e del giudizio non fosse affar mio . È un fatto però
 quegli scritti , ma non è affar mio . Di fronte ad una val
hio testardo e selvatico . Affar suo . Tanto , anche senza
is si . I protettori erano affar suo . E Farge , lo conosce
ella Foresta Maligna , era affar suo . A pensarci bene , la
he quella morta lì non era affar suo ) . Prisca pensava che
detto che la cultura non è affar suo , signor direttore , r
 diritti televisivi , sono affar suo . Ma chi decide davver
e la rincorsa alla Lazio è affar suo , lui aveva lasciato l
agnoni : il superG non era affar suo e la prima parte della
lgesse Domaris , era anche affar suo . Deoris sperava di cu
agnoni : il superG non era affar suo e la prima parte della
   deciso . " Questo non è affar tuo ! » Daniel sorrise sin
lo . In Islanda ? Questo è affar tuo . E va bene , disse Mu
eravate i cacciatori . Era affar vostro . " Jack di nuovo f
```

[6] In most grammars it has been considered and described particularly in a phonetic or graphematic perspective. See, for instance, the description given by L. Serianni (1991: 29-33)

[7] As E. Tognini Bonelli observes in her analysis of *saper* and *sapere*, "Usually grammars differentiate *saper* from *sapere* because the *-er* form is followed by an infinitive form, but the difference between the two forms is not even mentioned in a dictionary [...] as they are supposed to share the same meaning" (2001: 96).

The observation of a number of cases seems to support the hypothesis. An illustration can be given by a quick scan of an automatically reduced short sample of *vuol*:

```
  pds è d ' accordo perché  vuol dimostrare il massimo risp
osce al paganesimo e di cui vuol dimostrare tutta la gravit
a meglio che in città . " " Vuol dire che c ' è qualcuno di
edecessore e diventare re , vuol dire risolvere il conflitt
ria : ovviamente questo non vuol dire che sono cambiati i g
ecchezza vuol dire questo , vuol dire sobrietà . Più ch
" così complicato dire cosa vuol dire essere di sinistra ..
o " Signora , Marcolino non vuol dire le preghierine , vuol
ttimo ha realizzato : crisi vuol dire rimpasto , rimpasto s
 . Se credo in un ribasso , vuol dire che stimo che il prez
mormora un accademico . Che vuol dire ? Che il prestigio de
a miele vergine integrale . Vuol dire che il miele non ha s
e dalla logica psichiatrica vuol dire anche smettere di pen
 aperto . L ' Estate Romana vuol dire anche mostre . Il Pal
, dice : siete allegre eh ! vuol dire che la vita vi va ben
 a tutelare i cittadini " . Vuol dire che le circolari mini
dirittura in pensione , ciò vuol dire che esisteva un tempo
 ostei ... poi ricordò . «  Vuol dirgli per piacere che Flo
izio . Che so : se Alitalia vuol far sapere , attraverso un
arà . Se la signora Renauld vuol tener nascosto qualcosa ,
```

Not only is *vuol* followed in all its instances by an infinitive form, but a particularly strong association emerges with some verbs such as *dire*.

Once again we notice the co-selection of a number of words and the forming of units of meaning, extended beyond the word level, where lexical and syntactic elements are strictly combined.

The strong relation marked by the apocope is further illustrated in the following sample of *batter*.

```
 Ms Winshaw la smettesse di batter cassa e che non ci si po
ato » disse Batisti , senza batter ciglio . « Non ho altro
eddoloso . Sopportava senza batter ciglio gli scherzi pesan
e ? ) , riprendendosi senza batter ciglio e seguitando a ca
seduto e mi ascoltava senza batter ciglio . Con gli occhi f
orn | ) Lo ascoltammo senza batter ciglio , come se non ci
di secondo grado ( " Può il batter d ' ali di una farfalla
ii , ma anche un molteplice batter d ' ali di sincopata bel
torie , violentandole in un batter d ' occhio in qualunque
este prime ascendenze ed in batter d ' occhio mi trovo fina
upi di niente , torno in un batter d ' occhio . Sam non ave
e cura il raffreddore in un batter d ' occhio . Un ' altra
```

```
rrotolarla nuovamente in un batter d ' occhio . Ma non fiat
e trasferite a Roma , in un batter d ' occhio dalla Corte d
o venire le 7 di sera in un batter d ' occhio . il cellular
omincia la tempesta . In un batter d ' occhio sono bagnati
e . Il frate assentì con un batter di ciglia . Maros aggiun
co e pensare a un rinnovato batter di tacchi . In ogni caso
 di scorgere , sorridente , batter le mani tra i satiri un
o contentissimo e mi misi a batter le mani a mia volta . Il
```

4. Conclusions

Further analysis is needed to validate these results, but it is important to observe how patterns of regularity emerge which should not be overlooked. Data observation may lead to the identification of regularities which may lead to new hypothesis of generalisation and description. We 'still have a long way to go'[8]. In Italian studies, we must admit, we have not yet seized the opportunities given by the approach, in particular with reference to extended units of meaning and lexical phrases. This could be the starting point for a reconsideration of the network of relations existing in the language as well as of the units of meaning formed by the intertwining of lexical and grammatical features.

References

Aijmer, Karin / Altenberg, Bengt (eds.) 1991. *English Corpus Linguistics*. London / New York: Longman.

Baker, Mona / Francis, Gill / Tognini-Bonelli, Elena (eds.) 1993. *Text and Technology. In Honour of John Sinclair*. Amsterdam / Philadelphia: Benjamins.

[8] See Sampson (1992). Even if, in the article, the remark is referred in particular to computational corpus studies, it seems to summarize most of the considerations contained in this paper.

Biber, Douglas 1993, Representativeness in Corpus Design. *Literary and Linguistic Computing,* 8/4, 243-257.

Biber, Douglas / Conrad, Susan / Reppen, Rand 1998. *Corpus Linguistics. Investigating Language Sstructure and Use*. Cambridge: Cambridge University Press.

Biber, Douglas / Johansson, Stig / Leech, Geoffrey / Conrad, Susan / Finegan, Edward 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Cresti, Emanuela 2000. *Corpus di italiano parlato.* Firenze: Accademia della Crusca.

De Mauro, Tullio / Mancini, Federico / Vedovelli, Massimo / Voghera, Miriam 1993. *Lessico di frequenza dell'italiano parlato.* Milano: Etas Libri.

De Mauro, Tullio 1999. *Grande dizionario italiano dell'uso – GRADIT.* Torino: UTET.

Habert, Benoit / Nazarenko Adeline / Salem André 1997. *Les linguistiques de corpus*. Paris: Colin/Masson.

Hunston, Susan / Francis, Gill 2000. *Pattern Grammar. A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam / Philadelphia: Benjamins.

Leech, Geoffrey 1991. The State of the Art in Corpus Linguistics. In Karin Aijmer and Bengt Altenberg (eds.) *English Corpus Linguistics.* London / New York: Longman, 8-29.

McEnery, Tony / Wilson, Andrew 1996. *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Rossini Favretti, Rema 1998a. Using Multilingual Parallel Corpora for the Analysis of Legal Language: The Bononia Legal Corpus. In Wolfgang Teubert, Elena Tognini Bonelli, Norbert Volz. (eds.) *Translation Equivalence. Proceedings of the Third European Seminar.* The TELRI Association e.V., Institut fur Deutsche Sprache, The Tuscan Word Centre, 57-68.

Rossini Favretti, Rema 1998b. Cross-language Analysis and Large Multilingual Corpora. *Studi italiani di linguistica teorica e applicata,* XVII/3, 415-434.

Rossini Favretti, Rema 1999. Scientific Discourse: Intertextual and Intercultural Practices. In Rema Rossini Favretti, Giorgio Sandri, Roberto Scazzieri (eds.) *Incommensurability and Translation: Kuhnian Perspectives on Scientific Communication and Theory Change.* Cheltenham: Edward Elgar, 201-216.

Rossini Favretti, Rema 2000. Progettazione e costruzione di un *corpus* di italiano scritto: CORIS/CODIS. In Rema Rossini Favretti (ed.) *Linguistica e informatica. Corpora, multimedialità e percorsi di apprendimento.* Roma: Bulzoni, 39-56.

Rossini Favretti, Rema 2001. La linguistica dei *corpora* in Europa: prospettive di analisi. *Lingua e Stile,* XXXVI/2, 367-381.

Sampson, Geoffrey 1992. Probabilistic Parsing. In Jan Svartvik (ed.) *Directions in Corpus Linguistics,* Berlin / New York: Mouton de Gruyter, 425-447.

Serianni, Luca 1991. *Grammatica italiana. Italiano comune e lingua letteraria*. Torino: UTET.

Sinclair, John M. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sinclair, John M. 1996. The Search for Units of Meaning. *TEXTUS,* IX/1, 75-106.

Sinclair, John M. 1998. The Lexical Item. In Weigand Edda (ed.) *Contrastive Lexical Semantics*. Amsterdam / Philadelphia: Benjamins, 3-15.

Sinclair, John M. 2000a. The Computer, the Corpus and the Theory of Language. In Gabriele Azzaro and Margherita Ulrych (eds.) *Transiti linguistici e culturali.* Trieste: E.U.T., 1-15.

Sinclair, John M. 2000b. Current Issues in Corpus Linguistics. In Rema Rossini Favretti, (ed.) *Linguistica e informatica. Corpora, multimedialità e percorsi di apprendimento.* Roma: Bulzoni, 29-38.

Sobrero, Alberto (ed.) 1993. *Introduzione all'italiano contemporaneo. I. Le strutture. II. Le variazioni e gli usi.* Bari: Laterza.

Svartvik, Jan (ed.) 1992. *Directions in Corpus Linguistics.* Berlin / New York: Mouton de Gruyter.

Tamburini, Fabio 2000. Annotazione grammaticale e lemmatizzazione di corpora in italiano. In Rema Rossini Favretti, (ed.) *Linguistica e informatica. Corpora, multimedialità e percorsi di apprendimento.* Roma: Bulzoni, 57-74.

Teubert, Wolfgang 1996. Editorial. *International Journal of Corpus Linguistics,* 1, 1-2.

Thomas, Jenny / Short, Mick (eds.) 1996. *Using Corpora for Language Research.* London / New York: Longman.

Tognini-Bonelli, Elena 2001. *Corpus Linguistics at Work.* Amsterdam / Philadelphia: Benjamins.

Váradi, Tomás 2001. The Linguistic Relevance of Corpus Linguistics. In Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie, Shereen Khoja (eds.) *Proceedings of the Corpus Linguistics 2001 Conference*, Lancaster: UCREL, vol. 13, 587-593.