Burr, Elisabeth (2001):"Romance Linguistics and Corpora of French, Italian and Spanish Newspaper language", in: Fiormonte, Domenico / Usher, Jonathan (eds.): *New Media and the Humanities: Research and Applications*. Proceedings of the first seminar Computers, Literature and Philology, Edinburgh 7-9 September 1998. Oxford: Humanities Computing Unit, University of Oxford 85-104.

# ROMANCE LINGUISTICS AND CORPORA OF FRENCH, ITALIAN AND SPANISH NEWSPAPER LANGUAGE

> Was allein geeignet ist, als Leitstern, durch das ganze Labyrinth der Sprachkunde hindurchzuführen, findet auch hier Anwendung. Die Sprache liegt nur in der verbundenen Rede, Grammatik und Wörterbuch sind kaum ihrem todten Gerippe vergleichbar. Die bloße Vergleichung selbst dürftiger und nicht durchaus zweckmäßig gewählter Sprachproben lehrt daher viel besser den Totaleindruck des Charakters einer Sprache auffassen, als das gewöhnliche Studium der grammatischen Hülfsmittel. [...] Freilich führt dies in eine mühevolle, oft ins Kleinliche gehende Elementaruntersuchung, es sind aber lauter in sich kleinliche Einzelheiten, auf welchen der Totaleindruck der Sprachen beruht, und nichts ist mit dem Studium derselben so unverträglich, als bloss in ihnen das Grosse, Geistige, Vorherrschende aufsuchen zu wollen. (Humboldt 1827-1829/1963: 186 and 200).

## 1. Introduction

Since 1990 a computer corpus-based course for students of Romance linguistics has been gradually established at Duisburg University. When the course was started, there was only an Italian newspaper corpus, but now there are newspaper corpora for each of the three languages: French, Italian and Spanish. All of these have been enriched with tags which represent the structural features of newspapers and text. To date, the tagging is in the COCOA-Format. These corpora can be used to study grammatical and stylistic features in general or the variation present inside each corpus. Furthermore, as the newspapers in the different corpora were published at the same time, synchronic comparative studies of the three languages can be undertaken. The Italian corpus, being composed of two subcorpora commenced at different times, can also be studied under a diacronic perspective. But why should corpora be used to teach Romance linguistics?

## 2. Romance linguistics

### 2.1. Traditional synchronic Romance linguistics

Linguistics is the scientific study of language. Thus, Romance linguistics, if we just concentrate on the synchronic perspective, studies Romance languages in a systematic way. Up to now, Romance linguistics has been, however, very heavily theory-driven. Language is, in fact, looked upon from an exclusively theoretical point of view, so to speak from above, and is forced into models constructed *a priori*. These models themselves are based on a very small number of isolated phrases, which in the majority of cases have been invented. Indeed, invention is very often considered to be the only effective way to reach a comprehensive description of language.

Such is, for example, the concept behind of one of the most thorough studies of the Italian verbal system. Bertinetto, in fact, invented himself most of the examples which appear in his work. In his view, it would have been an act of pure optimism to rely on naturally occuring speech in order to find the most adequate example for every variation of meaning, given the marginality of some uses:

> La maggior parte degli enunciati illustrativi contenuti in questo lavoro sono frutto di libera invenzione; [...] non nego affatto che l'uso concreto offra formulazioni più ricche e più duttili di quelle che in genere sono partorite dalla mente di chi se le costruisce da sé secondo le esigenze del proprio argomentare; ma è indubbio che la speranza di trovare già bell'e confezionato l'esempio più adatto per ogni minima sfumatura di senso [...], appare un tantino ottimistica, data la marginalità di certi impieghi. (Bertinetto 1986: 13).[1]

He justifies such a procedure with Chomsky's concept of the native speaker, who as Bertinetto says is capable of producing an unlimited number of phrases in their own language by drawing on the registers and styles which a certain type of education has put at their disposal:

> Nessuno mette in dubbio che il parlante nativo sia in grado di produrre infiniti enunciati nella propria lingua, sfruttando tutti quei registri e quegli stili discorsivi che il suo livello d'istruzione gli consente di padroneggiare. (Bertinetto 1986: 13).

Thus using naturally occuring language data would mean restricting one's own native speaker competence to the stylistic registers present in a corpus. Such a restriction, can, however, according to Bertinetto, only be justified in cases where languages are studied by linguists who are not native speakers:

> [...] questa autolimitazione mi pare assennata solo nel caso di studiosi che si propongano di descrivere una lingua straniera, non certo nel caso di chi possieda in proprio la 'competenza' del parlante nativo. (Bertinetto 1986: 14).

A similar principle governs teaching. Students are, in fact, trained to look at models developed on the basis of de-contextualised and invented language 'data' and then to apply

[1] Geoffrey Sampson, however, is of quite the opposite opinion: "If the linguist relies on data invented by himself [sic] in his role of native speaker of the language [...], then it is near-inevitable that the linguist will focus on a limited range of phenomena which the research community has picked out as posing interesting problems, while overlooking many other phenomena that happen never to have struck anyone as noteworthy." (Sampson 1993: 268).

these models to equally invented isolated phrases. Hence, Michael Stubbs' evaluation of modern synchronic linguistics applies, likewise, to Romance linguistics in more specific terms:

> [...] it is so easy to be blind to the very small amount of data on which contemporary linguistics is based, and this is fundamental to the whole intellectual organisation of the discipline: the theories which linguists develop, the types of corroboration they claim, the methods they use, and the ways in which students are trained. (Stubbs 1993: 10).

Furthermore, if the language in question itself is studied at all, analysis is carried out on the basis of very small amounts of spoken or written text or, and this is much more often the case, on the basis of dictionaries and grammars. Dictionaries and grammars are themselves, however, static abstractions of the language in question and present linguistic elements in an atomistic way. Speech, i.e. the only reality of language at our disposal is represented there either via quotations, mostly from literary works, or via artificial examples which have been constructed in such a way that they illustrate precisely the rules in question, or as Bertinetto admits in the quote given above, "secondo le esigenze del proprio argomentare" (Bertinetto 1986: 13).[2]

## 2.2. Data-driven synchronic Romance linguistics

According to Wilhelm von Humboldt, though, whom I have quoted extensively above, language as such can only be studied "in der verbundenen Rede", i.e. in naturally occurring speech, compared to which dictionaries and grammars are, at best, a dead skeleton. Even a comparative study, he says, of unsubstantial and not necessarily well chosen samples of speech allow for more insight into the overall character of a language than the traditional study of tools like dictionaries and grammars. Charles J. Fillmore puts it in similar terms: "every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way." (Fillmore 1992: 35).

It is precisely this insight into the characteristics of languages like French, Italian and Spanish which students are supposed to gain in courses devoted to synchronic Romance linguistics. Gaining insight does not mean, however, that the study of theory, models, dictionaries and grammars has to be excluded from courses devoted to it. On the contrary,

[2] Wallace Chafe, occupying himself with the methods used when studying language, comes to a similar conclusion: „The techniques that have most dominated the field of linguistics [...] have been techniques focused on artificial rather than naturally occurring data.“ (Chafe 1992: 85).

they can and should be integrated and critically explored by comparing their assertions to actual findings.

Studying samples of naturally occurring speech leads, as Humboldt points out, to a "mühevolle, oft ins Kleinliche gehende Elementaruntersuchung", to a tedious and cumbersome study of the single elements of language. But it is precisely these possibly mediocre details which Humboldt claims make up the overall character of the individual languages and, thus, nothing is more of an obstacle to the study of individual languages than trying to find in them only the spiritual, the dominant or the grand. It is obvious, I think, that the type of language study Humboldt is advocating here, would, today be called *data-driven* linguistics, using a term coined by John Sinclair and his school (see f. ex. Sinclair 1992).

At a first glance, then, the type of linguistics which constitutes the framework of part of my research and which I am trying to teach in some of my classes, could be described as data-driven Romance linguistics aimed at allowing students to gain insight into the complexity of naturally occuring contemporary French, Italian and Spanish and to examine critically their more traditional and theoretical descriptions, on the basis of their own findings.

## 3. The conception of language

Being data-driven is, however, not enough to define the type of linguistics in question. Instead, it has to be differentiated from the dominant type of linguistics, which in this century and up to now has not only been very much theory-driven, but has also been characterised by an essentially dualistic conception of language.

### 3.1. The dualistic conception of language

This dualism does not just concern the functional oppositions which are seen as binary oppositions, i.e. *homme / uomo / hombre / man* [+ masculine] - *femme / donna / mujer / woman* [- masculine] (cf. for example Renzi 1989: 318), but language itself. In fact, Saussure's dichotomy *langue* and *parole* is still today the starting point for most approaches. Its endurance is, in a way, sustained by Chomsky's *competence* - *performance* differentiation, which likewise remains central to modern linguistics. The two entities *langue* and *parole* or *competence* and *performance,* are, however, not of equal importance for modern main-stream linguistics. On the contrary, whereas *competence* or *langue* is really the centre of attraction around which gravitate the linguistic models which have been created and the discussions carried out about them, *performance* or *parole* is looked upon as a secondary manifestation of language. For this reason elaboration of a theory of speech is not even seriously envisaged. This means that the situation in linguistics has not really changed since the time of Dell Hymes and that his characterization of modern linguistics is still valid today: „A major characteristic of modern linguistics has been that it takes structure as primary end in itself, and tends to depreciate use“ (Hymes 1972: 272).

It is true that since the beginning of the 'eighties, above all in Britain and in northern Europe a new type of linguistics has evolved and gradually established itself. This type of linguistics, now generally called corpus linguistics or even computer corpus linguistics, tries to turn *performance* into its starting point and to describe, in a systematic way, languages on the basis of natural speech. This does not necessarily mean, however, as the following remark from Jan Svartvik shows, that it has overcome the *competence* - *performance* dualism: „Linguistic competence and performance are too complex to be adequately described by introspection and elicitation alone.“ (Svartvik 1992: 8). Instead, it aims, more than anything else, at the mere substitution of introspection and elicitation by data or their integration with data. The reason for this might be that although speech is considered to be the only reality at our disposal, this type of linguistics has not yet, as far as I can see, developed its own coherent theory of speaking. Because of this, it seems to be constrained either to defend itself against the assertions of the theory-driven type of linguistics and its fundamental opposition to the analysis of speech, or to fill the binary distinction with new content but without ever questioning the dualistic way of looking at language. Halliday's concept of *system* and *instance* (cf. Halliday 1992: 66) seems to me a good example of this tendency.

### 3.2. Social implications of the dualistic conception of language

The dualistic conception constitutes in my eyes a serious problem, however. First of all, an activity like speaking is much too complex to be studied in binary terms, i.e. as the realisation of systematic possibilities or relations.[3] Secondly, this view has far reaching implications with respect to the type of speech which is studied at all and thus for corpus linguistics as well. In fact, as the primary sources of corpus linguistics are not just any portions of speech but corpora constructed on the basis of a particular conception of language, conceiving language

[3] See also John Sinclair's opinion on this: "Language is an abstract system; it is realized in text, which is a collection of instances. This is clearly an inadequate point of view, because we do not end up with anything like text by 'generating' word strings from grammars." (Sinclair 1991: 102).

in dualistic terms and thus basing corpus construction on the *core* and *periphery* distinction[4] will lead to the exclusion of vital parts of natural speech and, what is more important, of whole categories of speakers.[5]

Obviously, such an exclusion has always been the case. Over the centuries the dominant view of what is a language or of what it means to know a language, have in fact worked against a whole range of expressions, language varieties and strata of the population. But there have been historical events which more than others sanctioned such an exclusion. Just think of what happened to the Romance languages and others when printing established itself. Heavy pressure towards normalisation was brought to bear on the language in favour of the economic success of the new technology, which, although fostering progress in cultural and economic terms, lead to a considerable reduction of possibilities of expression. The purist dictionaries which were edited in the 17th and 18th century by the famous academies *Accademia della Crusca*, *Académie Française* and *Real Academia Española* (cf. Nencioni 1990: 345-346) are just one outcome of this process. In our century, the background against which in Italy the *Dieci tesi di una educazione linguistica democratica* were elaborated (cf. De Mauro 1975/$^{3}$1981: 138-151) is a good example, instead, of what happens when on the official level of education only one type of language is allowed to exist and the people who do not know it are pushed aside.

If these sanctions could take place without the technology which we have at our disposal, we have to ask ourselves what will happen in times when the power of deciding what is possible or correct is no longer in the hands of lexicographers or printers who in many cases were at the same time literary scholars, but is handed over to a computer by the speakers themselves, or to be more precise, by those speakers who have access to the technology and its control. That the computer is really made into the arbiter of language use is shown by the following type of arguments you might come across nearly every day, above all in

[4] The concept of *core* was originally developed by Charles F. Hockett in order to account for the fact that people with different idiolects could nevertheless communicate together (cf. Hockett 1958: 332). Later on, however, it was extended to all the different varieties of a language: „however esoteric or remote a variety may be, it has running through it a set of grammatical and other characteristics that are common to all.“ (Quirk/ Greenbaum $^{4}$1975: 1).

[5] For a discussion of the *core* on which is built the BNC and the exclusions it entails see for example Atkins/Clear/Ostler (1992) and Clear (1992). According to Gerhard Leitner the concept of *core* is also at the basis of the *International Corpus of English* (cf. Leitner 1992: 37 & 44-46).

institutional and administrative environments: "the computer can't accept this, this is too long for the computer, it only accepts *il vient* and not *elle vient*, *professor* and not *professora*".

Another important factor, today, are the so-called language industries. Leaving aside, at this point, the problems involved for languages which will either not be able to take part in the development in this field or are not yet taking part efficiently enough,[6] I should point out, that there could be considerable danger involved for those languages, too, which are already at the centre of this process. In order to be more easily processable, these languages might, in fact, be reduced, in the long run, to a presumed *core*.

If we as linguists, given this background and the ongoing collaboration between branches of linguistics and the language industries, continue to maintain a view of language and of linguistic knowledge which does not correspond to its complexity or just takes the traditional hierarchies for granted, we, too, will bear responsibility, should an even more incisive ideological normalisation of language come about in the near future and if in teaching and wide areas of communication, language varieties differing from a supposed *core* are assigned to the *periphery* together with the people who use them.[7]

### 3.3. A non-dualistic theory of speech

What we need for our studies, teaching and corpus building, instead, is a theory of speech or, more precisely, a theory of what the speakers know when they speak, which takes into account precisely the complexity of language. Obviously, by speaking I do not just mean oral expression, but written expression as well. Such a theory is not at all new. On the contrary, it has already been proposed by Dell Hymes (1972) and by Eugenio Coseriu (1988b).[8] Both linguists, in fact, base their considerations on language as it manifests itself in reality and not on an idealisation, and develop their theory in opposition to Chomsky's dualistic conception of language and in opposition to the idea of a *common core*. In both theories, speaking is considered to be a very complex activity based on a whole range of different spheres of knowledge. Chomsky's ideal speaker-listener who acts in a homogeneous speech community

[6] See Francisco Marcos Marín (1994: esp. 49-62) for a discussion of this topic.

[7] According to Michael Barlow the concept of common core has, in fact, already had negative consequences not only for grammatical description but for the teaching of foreign languages, as well (cf. Barlow 1996: 4).

[8] Coseriu has proposed elements of such a theory of speech in many of his works. The book I am referring to, here, has the advantage, however, that these elements are drawn together and elaborated into a coherent theory of linguistic knowledge.

is confronted with the heterogeneous character of real speech communities and with the socio-culturally determined knowledge of languages and language varieties.

The starting point for both theories is, thus, the observation that in most speech communities people are not mono- but plurilingual, i.e. that they speak more than one historical language: "[...] in much of our world, the ideally fluent speaker-listener is multilingual." (Hymes 1972: 274), or, at least, more than one variety of such a language: "Even an ideally fluent monolingual of course is master of functional varieties within the one language." (Hymes 1972: 274). In similar terms Coseriu says that

> [...] verschiedene Normen und Regeln von verschiedenen Sprachsystemen im tatsächlichen Sprechen nicht nur der Gemeinschaft, sondern auch der Individuen angewandt werden oder - kürzer gesagt - daß auch das Individuum in seiner Sprachgemeinschaft und innerhalb seiner historischen Sprache mehrere Sprachen spricht. (Coseriu 1988b: 263).[9]

If we want to take this reality into account, we have to, as Hymes says, "break with the tradition of thought which simply equates one language, one culture, and takes a set of functions for granted." (Hymes 1972: 289).

According to Hymes, we have to realise, furthermore, that not even the different language varieties of a certain language can be related to a common grammar, let alone the different languages spoken in a speech community (cf. Hymes 1972: 275). This fact is perhaps best explained by using two concepts which play an important role in Coseriu's theory and which take languages into account in two different dimensions, i.e. in their heterogeneity and their homogeneity. It will be useful, furthermore, to call to mind the linguistic situation in countries like Italy, Germany or Spain, where dialects which are by no means mutually understandable are still alive.

### *3.3.1. The historical language*

The concept which tries to cover the heterogeneous dimension of a language is the concept of the *historical language* and its external structure or architecture (cf. Coseriu 1988b: 24-25 & 139-148). A historical language constitutes, in fact, an autonomous combination of linguistic traditions. As such it is recognised not only by its own speakers but also by speakers of other historical languages, who identify it with adjectives like English, French, Italian, or Spanish.

[9] different norms and rules of different language systems are not only applied in natural speech of a speech community but also of individual persons or - said in a more concise way - that also individuals speak several languages within their speech community and within one historical language.

Regarding the linguistic traditions of which such a language is composed, we can distinguish three different types on the basis of diatopic, diastratic and diaphasic differentiation.

With respect to the diatopic differentiation a historical language like Italian is composed of primary dialects like Sicilian, Venetian or Tuscan and a historical language like Spanish of Asturian, Castilian, or Aragonese etc. They are primary in the sense that they existed as independent languages long before one of them, - in the case of Italian the Florentine variety of Tuscan, in the case of Spanish, Castilian, - developed into the communal language of the respective country. When the use of such a communal language is spreading all over the country and is thus becoming a sort of *lingua franca* for the speakers or a *roofing language* for its former co-languages, diatopical differences will normally lead to its own internal differentiation and bring about secondary dialectal traditions. In Italian, these secondary dialects are the so-called *italiani regionali* in Italy or Italian in Switzerland, the Spanish secondary dialects are Andalusian, Canarian or the different forms of Spanish in America etc., all of which are distinguished inside the communal language.[10] Diatopical differences might, however, also be present inside the socio-cultural norm, which in a linguistic community is recognised as an exemplary model or standard. These tertiary types of dialects can be discerned, for example, when listening to the speeches Italian or Spanish politicians make in parliament, i.e. in a formal speech situation where the standard language is normally used.[11] The formal type of Italian in Switzerland or the Mexican, Peruvian, Chilean, Rioplatese etc. form of the Spanish standard are to be considered tertiary dialects, too.[12]

By diastratically characterised traditions we refer, instead, to the languages spoken by the various socio-cultural classes or strata. These traditions can be called sociolects or *niveaus*. With respect to Italian we normally distinguish, in fact, between *italiano colto*, *italiano dell'uso medio* and *italiano popolare*, and with respect to Spanish between *español culto*/*habla culta* - *español medio*/*habla mediana* - *español vulgar*/*popular*. Such traditions

[10] Obviously, diatopical differentiation exists inside communal English, as well: In fact, British, American or Indian English are all to be considered secondary dialects of the historical language English.

[11] This does not mean, however, that tertiary dialects are characterised only by differences in pronounciation, instead there are usually also some differences in vocabulary and grammar.

[12] Another good example for a diatopically differentiated standard language is German, where there are not just tertiary dialects in the standard spoken in Germany itself, but different standard languages in the German speaking countries, as well, with their own dictionaries and grammars. The same goes for English where different standard forms have at least developed in the different countries where English is spoken if not inside the individual countries themselves.

can, by the way, also be observed inside the various types of dialects themselves. A sociolect or *niveau* is, thus, a variety which characterises a specific socio-cultural class. As such it can also be differentiated further in diatopical terms.

The last type of differentiation derives from the different communicative situations. Diaphasic varieties or styles can be distinguished within the various types of dialects and within the sociolects, as well. As the various social groups like football fans, women, men, professional people and youths etc., are delimited by the respective point of observation the type of language they speak as a group falls into the category of style, too. When considered as autonomous entities, these varieties show, in their own right, diatopic and diastratic differenciation. With respect to Italy or Spain, this means that the language spoken, for example, inside the family is neither the same in all regions nor in all socio-cultural strata.

Hymes' point that the different language varieties cannot be related to a common grammar, becomes even clearer when we consider that dialects can themselves function as sociolects or styles. This is the case when a certain dialect is normally spoken by a specific socio-cultural class whereas the other socio-cultural classes speak the communal language. The Milanese upper classes speaking Milanese in order to distinguish themselves from the Italian-speaking lower classes may serve here as an example. Dialects function, instead, as styles, when in a certain communicative situation a certain dialect is normally used, whereas in another situation a different dialect or the communal or standard language is normally spoken. An example could be, that inside a family which originally comes from Aragón but lives now in Madrid, Aragonese is normally spoken whereas at work the Spanish communal language is used. It should be obvious, thus, that the varieties of the historical languages Italian or Spanish can by no means all be related to the same system.[13]

The functional substitution of certain varieties by other varieties, which is in question here, is synchronically, however, possible only in one direction. That is why Koch and Oesterreicher call it a *Varietätenkette*, that is a variety chain (Koch/Oesterreicher 1990: 14), which can be presented in the following way:

[13] There are also speech communities where a different historical language functions as sociolect or style. In Italy and Spain the situation in Alto Adige or in the Basque country can be cited here.

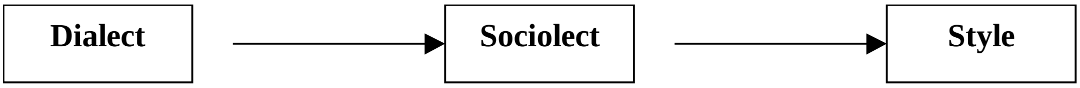

fig. 1 Variety chain (cf. for example Coseriu 1988b: 146)

This combination of three types of dialects, of niveaus or sociolects and of styles together with the possibilities offered by the variety chain take into account the complex heterogeneity of a historical language and constitute, in fact, its architecture or external structure, which can be visualised as in the following figure:

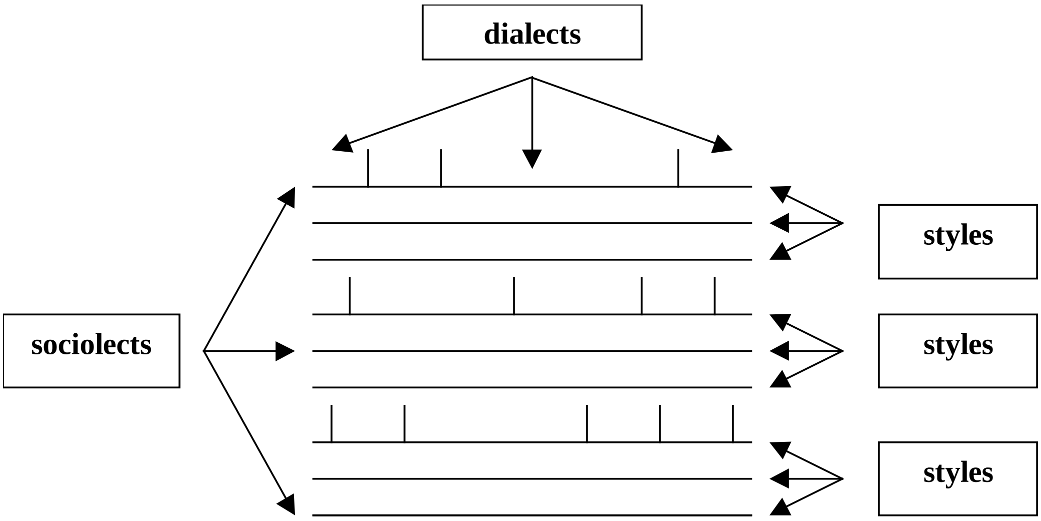

fig. 2 The architecture of a historical language (cf. Coseriu 1988a: 283)

### *3.3.2. The functional language*

The concept which tries to cover the homogeneous dimension of a language, instead, is the concept of *functional language* with its internal structure (cf. Coseriu 1988b: 25-27 & 266-278). In this dimension we find those varieties which are to be considered synchronic, syntopic, synstratic and synphasic entities. In more concrete terms, what is envisaged is a certain dialect or the communal language spoken by a certain socio-cultural class in a specific communicative situation.

With respect to such an entity alone, we can talk about *system* and *norm*, i.e. the two structural levels distinguished by Coseriu within the 'virtual technique'.[14] The *system* in this theory is understood as a very abstract entity which contains only the functional oppositions, the elements and procedures. The *norm*, instead, is a social entity which comprises the already

[14] As can be seen in the figure below, Coseriu distinguishes a third level, i.e. the language type. In our context, this level is, however, not of immediate importance.

traditional realisations. Whether these realisations are functional or not is in this case of no importance. Whereas the *system* is understood, above all, as a system of possibilities, which includes also the virtual side of a variety, the social norm is seen as mainly a system of traditional constraints which have developed through history, not least with the help of influential people or institutions. Through these constraints, the possibilities offered by the system are reduced to those traditionally already realised. It is this social norm, and not the abstract system, which is then realised in a concrete discourse:

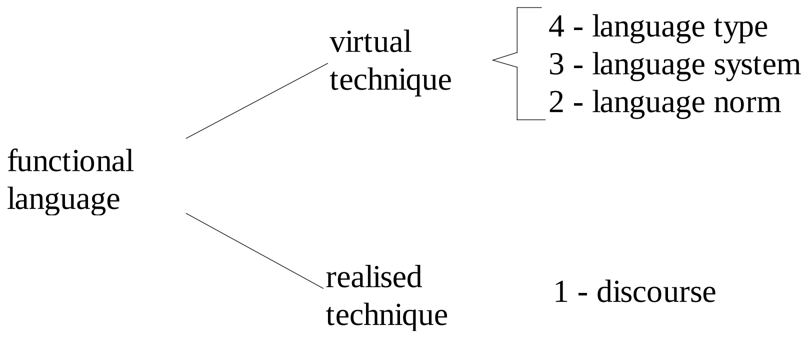

fig. 3 The structure of a functional language - virtual and realised technique (cf. Coseriu 1988a: 293)

### *3.3.3. The knowledge of the speakers*

The speakers' linguistic knowledge of such a variety, however, does not correspond to the variety's synchrony. Instead, the speakers' linguistic knowledge is made up at any point in time of linguistic facts which like *thou* in English do not function in the variety's present synchrony but are functional entities only in a former state of the language. Other linguistic facts, on the contrary, will be valued by the speakers in a diachronic perspective, i.e. some might say, for example that *udire* is not used any more in Italian today, others might still consider *sentire* as an innovation. This diachronic consciousness determines the speakers' own attitude towards the linguistic phenomena and the application of their own linguistic knowledge (cf. Coseriu 1988b: 134-135). In this context it is not really important whether the judgments of the speakers correspond to the objective situation. Instead, the speakers are always right, because it is precisely their own opinion which influences their speaking (cf. Coseriu 1988b: 178). Such a diachronic consciousness exists, according to Coseriu, above all, in linguistic communities with a written literary tradition. In such a community, older forms can even be used deliberately, in order to refer to the respective state of the language or to the respective historical moment. Thus, the individual forms and the norms applicable to them remain present even when these forms no longer belong to the actual synchrony.[15] They can even be revitalised and will then be part of the actual synchrony again (cf. Coseriu 1988b: 135-137).[16]

In addition, the speakers' linguistic knowledge does not comprise just one variety of the respective historical language. Instead, in the normal course of things all speakers know several diatopic, diastratic and diaphasic varieties in various ways and they use them, too. Furthermore, they have a rudimentary (correct or incorrect) knowledge of other varieties of their own language or of other historical languages. This knowledge might very well be confined to the so-called languages of imitation, i.e. the traditions which exist in speech communities with respect to the imitation of varieties of the own historical language or of foreign languages which are not spoken by the members of the speech community itself. Italian communities, for example, which do not speak Tuscan, imitate this dialect via an exagerated use of the *gorgia*, in Spain, *seseo* and *yeísmo* can be used in the same way, and German is generally imitated by putting an -*en* on every single Italian or English word, which then leads to forms like *spaghetten*, *mangiaren* and so on.[17] As these examples show, the individual imitations do not necessarily have to correspond to the realisations which are traditional in the variety or historical language in question. On the contrary, such languages of imitation very often contain elements which do not exist in the variety or languages which are imitated or would even be impossible there (cf. Coseriu 1988b: 148-152). In German, for example, there exist expressions like *uno momento, picco bello, dalli dalli, alles paletti* which are supposed to be Italian, but are not Italian at all.

The speakers' knowledge however, does not correspond to the historical language either. This means that no speaker knows all the varieties a historical language is composed of and not all the speakers know the same varieties or do not know them in the same way. Furthermore, the languages or varieties speakers do know, are not put to use indiscriminately in all the situations of communication. Instead, the speakers possess, according to Hymes, a

[15] Thus, in England the language of St. James's Bible, of the Shakespearean plays, of Dickens' novels or at least some elements of it belong to the actual knowledge of present day speakers.

[16] The example given by Coseriu of the Spanish Suffix -*ì* might suffice here. This suffix, which had been used in older times to form adjectives for persons and facts related to the Orient, had been unproductive for a long time. It was revitalised, however, in our century for the formation of adjectives like *pakistaní* etc. (cf. Coseriu 1988b: 137).

[17] See, for a literary use of such imitation Skakespeare's *All's Well That Ends Well*, Act IV Scenes I & II.

*differential competence*, which allows them to discern the communicative adequacy of each variety:

> fluent members of communities often regard their languages, ot functional varieties, as not identical in communicative adequacy. It is not only that one variety is obligatory or preferred for some uses, another for others (as is often the case, say, as between public occasions and personal relationships). Such intuitions reflect experience and self-evaluation as to what one can in fact do with a given variety. (Hymes 1972: 274).

Apart from that, speakers possess also a *competence for production* and a *competence for reception* which are both socio-culturally determined, but do not coincide with each other in the majority of cases (cf. Hymes 1972: 275).

## 4. Teaching data-driven linguistics of speech with corpora

Parts of the theory I have tried to outline above and which is at the basis of my research, teaching and corpus-building, have in the meantime become an integral part of Romance linguistics. These are above all the concept of the *historical language* and the distinction between *system* and *norm*. The reason for this is that for historical reasons and because of its object, Romance linguistics possesses, in many respects, superior insight into the complexity of historical languages than the linguistics of many other languages and is, thus, less inclined to reduce languages to a *core*. What is generally missing, however, is a systematic investigation into the knowledge of the speakers and an evaluation of the many theories and models which have been elaborated within variational linguistics by confronting them with the reality of speech. Thus, even today, as a recent article by Raffaele Simone (1996) shows, Italian linguistics, despite its very elaborate linguistics of varieties, does not yet really know what is spoken in Italy, when and by whom. Instead, it has to rely on the statistics of institutions like ISTAT or on subjective impressions. To solve this problem it seems necessary to me to combine the above theory of speech with corpus or data-driven linguistics as outlined for example in Sinclair (1991). In order to do this we need, however, corpora which respect the linguistic knowledge of speakers.

### 4.1. The corpus of Romance newspaper languages

Without entering into the discussion about reference or sublanguages corpora, here,[18] I shall now present the corpus I am currently using for teaching. The corpus is composed of whole

[18] Let me just say that neither corpus as such really meets the requirements of the theory of speech, the first type by excess and the second type by deficiency. However, a discussion of this has to be left for a later contribution.

editions of French, Italian and Spanish newspapers which were published around the time of the last European elections in 1994. This period was selected for two different reasons. One reason was that distribution of newspapers is expected to be much larger at the moment of an important event. The other reason was the aim to achieve a high level of correspondence between the newspapers in question with respect to the themes handled, in order to allow for comparative studies of certain styles.[19] With the *European Elections* taking place everywhere in Europe at the same time, it could be expected that every newspaper in every country would report extensively about this event.

### *4.1.1. The composition of the corpus*

At the moment, the corpus consists of the three subcorpora *Le Monde*, *Corriere della Sera* and *La Vanguardia*, each of which is composed of three editions of the same paper.[20] The size of the subcorpora and of their individual components are given in the following table:

| Le Monde | tokens | types |
|---|---|---|
| **total:** | **236.236** | **22.545** |
| 12./13.06.1994: | 89.786 | 13.449 |
| 14.06.1994: | 70.936 | 10.399 |
| 15.06.1994: | 75.514 | 11.328 |

| Corriere della Sera | tokens | types |
|---|---|---|
| **total:** | **303.641** | **31.285** |
| 13.06.1994: | 102.976 | 16.400 |
| 14.06.1994: | 102.441 | 15.652 |
| 15.06.1994: | 98.224 | 16.498 |

| La Vanguardia | tokens | types |
|---|---|---|
| **total:** | **261.133** | **31.330** |
| 13.06.1994: | 86.468 | 14.367 |
| 14.06.1994: | 94.251 | 17.844 |
| 15.06.1994: | 80.414 | 16.234 |

fig. 4 Size of the subcorpora and their components

Although on account of its size and composition this corpus can by no means represent the whole knowledge of the speakers who make up the reading public of the individual papers, its

[19] Theme is after all one of the parameters which delimit speech situations.

[20] Altogether, much more material has been collected which, however, in its present state cannot be called a corpus as yet.

components stand for a "complete experience, that permits distinction but excludes selection" which according to Giovanni Nencioni (1990: 351) has to be taken account of when constructing a corpus. That this is really the case can be justified in the following way. The individual editions of the papers are retained as they presented themselves to the readers the day of their publication. Thus all texts which appeared have been retained in their entirety and phenomena which would disturb the picture if the aim was to represent a *core*[21] have not been excluded, on the contrary they have been assumed to belong to the experience of the readers and are, therefore, part of their knowledge. Such phenomena are the elements of foreign languages and dialects present in probably all the papers and the Catalan elements in the Spanish paper as well as elements which might not function in the present synchrony. Furthermore, the components do not represent a homogeneous language variety but are, in fact, a portrait of the real combination of stylistic and socio-cultural varieties peculiar to the paper.

There are, however, a few limitations (still) with respect to the congruence between the printed editions and their counterparts in the corpus. The reason is that CD-ROM editions and thus the sources of the French and Italian component do not mirror exactly the printed editions[22] and even from the magnetic tape *La Vanguarda* put at my disposal, every kind of publicity had for obvious reasons been excluded. In spite of these exclusions, the newspapers can nevertheless be understood as a quite reliable representation of what the actual readers are presented with and thus of the knowledge expected of them.

As not every reader will read all the paper, some might even just read one special section or type of article, this knowledge as such must be regarded as an abstraction, however. In order to achieve a more realistic notion, the different socio-cultural strata which make up the readership of the individual papers would have to be related using the demographic data at

[21] That to represent a *core of language* is also the aim of the LIP can be understood when considering the following remarks: "Abbiamo deciso di escludere dalla raccolta e dalla trascrizione tutti i documenti di parlato orientati in modo dominante verso il dialetto." (De Mauro 1993: 31) or "Il corpus del LIP non presenta dunque una varietà specifica dell'italiano, ma raccoglie testi di italiano tendenzialmente comune e unitario parlato in tutto il territorio nazionale." (Voghera 1987: 34).

[22] For example, picture captions do not appear in both the CD-ROM editions, the weather report is missing on the CD of *Le Monde* and TV- and radio programs are not retained on the CD of *Il Corriere della Sera*.

our disposal, to the various parts of the subcorpus.[23] Such a distinction is made possible by the markup of the corpus.

### 4.1.2. *The Markup-System*

Apart from the header where relevant information about the corpus and its elaboration is retained:

```
<<!  Spanish Newspaper Corpus "European Elections 1994" Component 1              >>
<<!  Original text copyright La Vanguardia, Barcelona                            >>
<<!  Text source magnetic tape, ATEX format                                      >>
<<!  Data extraction Gert-jan Burggraaf (NL)                                     >>
<<!  Translation ATEX format to DOS codepage 850 Gert-jan Burggraaf (NL)         >>
<<!  with support of Steward Whitelaw & Roger Tripper, Daily Telegraph, London   >>
<<!  Extraction & Translation Period 1995                                        >>
<<!  by extraction of this edition first letter of every line was lost           >>
<<!  has been corrected collating electronic with printed version                >>
<<!  Format MS-DOS Text                                                          >>
<<!  not lemmatised                                                              >>
<<!  preposition-article contraction like "del" to "de+el", "al" to a+el         >>
<<!  verb-enclitics contraction like "veder+la" to veder+la"                     >>
<<!  composed words with dash are retained as one form                           >>
<<!  Apostrophe is followed by blank, i.e. "del' autovia"                        >>
<<!  Hyphenation, line breaks correspond to printed version                      >>
<<!  Disambiguation: punctuation and digits, i.e. # = decimal point, \ = decimal comma >>
<<!  Disambiguation: Quotes and minutes / seconds, i.e. £ = minutes, $ = second  >>
<<!  Mistakes not corrected, but annotated {sic}                                 >>
<<!  Markup textual features                                                     >>
<<!  Markup format COCOA                                                         >>
<<!  Values of variables given in Italian                                        >>
<<!  Occhiello = CINTILLO, Titolo = TITULO, Sottotitolo = SUBTITOLO              >>
<<!  Sommario = ENTRADILLA, Catenaccio = DESTACADO                               >>
<<!  Markup of comments { } and << >>                                            >>
<<!  Correction, Disambiguation, Markup Elisabeth Burr                           >>
<<!  Elaboration Period 1995-1997                                                >>
<<!  Copyright Elisabeth Burr                                                    >>
<<!  The June 13, 1994 issue of the newspaper La Vanguardia, Barcelona, Spain    >>
```

fig. 5 The *header* of the Spanish subcorpus

every component of the corpus has been enriched with the following series of tags in the COCOA-format, which allow for a whole range of differenciations in terms of readership or language varieties.[24]

[23] When doing this, the BNC and the demographic parameters used when creating it could serve as a model. See for example Crowdy (1993).

[24] At the moment a research project is prepared which should allow the corpus to be enriched with parts of speech tags.

| **Reference** | **Code** | **Example** |
|---|---|---|
| paper | <Z> | <Z La Vanguardia> |
| editon | <E> | <E 130694> |
| section | <S> | <S Politica> |
| origin of the text | <A> | |
| signed | | <A firmato> |
| anonymous | | <A non firmato> |
| name of author | <N> | <N Tapia Juan> |
| page | <C> | <C 01> |
| language | <L> | <L Inglese> |
| texttype | <T> | |
| head-line | | <T Occhiello> |
| slugline | | <T Titolo> |
| sub-title | | <T Sottotitolo> |
| abstract | | <T Sommario> |
| in between title | | <T Catenaccio> |
| announcement | | <T Civetta> |
| article | | <T Articolo> |
| front-page strory | | <T Spalla> |
| TV-, cinema program | | <T Programma> |
| film content | | <T Film> |
| commentary | | <T Corsivo> |
| interview | | <T Intervista> |
| column | | <T Rubrica> |
| criticism | | <T Critica> |
| stop press | | <T Flash> |
| news in brief | | <T Breve> |
| leading article | | <T Fondo> |
| letters to the editor | | <T Lettera> |
| listings | | <T Elenco> |
| news | | <T Notizia> |
| weather report | | <T Tempo> |
| title of book, film, song etc. | | <T Nome> |
| picture caption | | <T Foto> |
| type of speech | <P> | |
| running text | | <P Prosa> |
| quote from written source | | <P Citazione> |
| quote from oral source | | <P Discorso> |
| interview question | | <P Domanda> |
| interview response | | <P Risposta> |

fig. 6 The Markup-System

### *4.1.3. Corpora and TACTWeb*

When corpora had been used in the years before quite a few problems arose with the installation of TACT, the textual analysis program to be used, either on the central university server or when students wanted to install TACT at home. The most critical point was, however, that the time needed by inexperienced students to learn to use the program did not really allow for the teaching of a course of Romance linguistics where students would gain insight into the characteristics of languages, learn how to do empirical research and to critically compare their findings to theoretical assertions. Former courses were, thus, devoted mainly to textual analysis as such. The availability of TACTWeb has altered the situation radically.

Part of the corpus, i.e. the edition of the 15.06.1994 of *Le Monde*, *Il Corriere della Sera* and of *La Vanguardia* is now on-line in the form of a database[25] into which it has been converted by means of *TACT 2.1*.[26] This database is installed on a relatively old PC 486 with 33 MHZ and 16 MB RAM running under Windows95 in my office. With the aid of a trial version of *O'Riley Software's WebSite* this PC functions as a server for *TACTWeb*,[27] which itself supports the querying of TACT-databases over the Web by operating with the server software. This has the advantage that students can work with it whenever they want or have time, the setup can be organised in a user-friendly way, it can even be adapted when needed and can be enriched with information which is directly relevant to the course. This corpus can be accessed via the following page http://www.uni-duisburg.de/FB3/ROMANISTIK\PERSONAL\Burr\humcomp\tacthome.htm.

## 4.2. An example

At the moment of writing, the corpus of Romance newspaper languages is used in a course on the Romance verbal system. As the corpus has not been tagged for parts of speech yet, students, first of all, have to tag the wordforms which are relevant for their research, before they can actually retrieve the data, thus gaining some insight into the tagging of a corpus, too. The Spanish component of the corpus, furthermore, is part of a course on variation, where linguistic phenomena which have been put forward in the literature as being characteristic for newspaper language or other written varieties are studied in the sections which are delimitable through the markup system in order to find out whether their usage is, in fact, particular to certain parts of the corpus and thus to certain varieties present inside the corpus. As the

[25] In the meantime, part of my Italian newspaper corpus 1989, where finite verb forms have been tagged according to the categories of tense, mode and aspect, is equally accessible over the web and is used for teaching a course on the verbal system of Italian.

[26] TACT was developed originally by John Bradley and Lidio Presutti at the University of Toronto from 1984 onwards.

[27] TACTWeb is a project developed by John Bradley & Geoffrey Rockwell.

teaching of these two courses is still going on, it is, however, too early for an evaluation of the results.

The example I am going to present, instead, is taken from the first course where TACTWeb was actually used by the students to retrieve the necessary data for their own research projects. As this course was taught at Duisburg University during the summer semester 1998 its results can be evaluated. I will be using for this evaluation a concrete research project which has been carried through by one of my students. Although the course itself was quite experimental in nature this project together with the results achieved should be able to show some of the advantages of courses where an integration of linguistic theory or descriptions of languages and empirical research is sought by using the new technology, text analysis software and electronic corpora now at our disposal.

### *4.2.1. The course*

The course was part of the program for undergraduate students and was devoted to the study of lexical phenomena in the three Romance languages French, Italian and Spanish. Variation was, where possible, to be taken into account, too.

The topics which were proposed for study during the course included:

*a)* foreign words like English *leader,* French *main propres,* German *hinterland,* Italian *a contrario,* Latin *mea culpa*;

b) word formation either with the help of prefixes like anti-, archi-, ex-, dis- or suffixes like -ista/-iste, -ione/-ión/-ion, via composition as in *molino de viento/macchina da scrivere/moulin à vent* or juxtaposition as for example *hombre-rana, café-teatro*, or in more specific terms, terminologies like *direc-trice/-teur, vinci-trice/-tore, ministr-a/-o*;

c) phraseology, for example *más loco que una cabra, bueno como el pan, lancia in resta, povero in canna, son dos caras de la misma moneda, un chien vivant vaut mieux qu'un chien mort*;

d) collocations as *temps mort, natures mortes, salto mortale, trappola mortale, errore mortale, abbraccio mortale, peccato mortale, muerte política, muerte social*;

e) word classes like prepositions, conjunctions or adverbs.

Students were, however, free to propose different topics, as well. In order to be able to judge whether the study was feasible with the means at our disposal, students had to hand in first of all a project proposal, giving a short description of why they wanted to study a certain topic and what they where interested in finding out. Then, at some later point, a draft of the structure of the paper they were to write and the steps they planned for their empirical research was requested. All the way through the course there was discussion of the project with the individual students.

### *4.2.2. A study*

In the framework of this course a particularly interesting study was done by a part-time student, a secondary teacher, who had only just started to use a PC (cf. Steinert-Schmitz 1998). By looking at the denominations of and the terminology for persons she had, in fact, become interested in the question if and how women are present in the *Corriere della Sera*. This concern lead her to formulate the following questions:

- How many women?
- Which are the linguistic means employed to make them visible?
- In which sections do they appear and how?
- What is their position?

#### *4.2.2.1. HOW MANY WOMEN*

In order to find an answer to the first question of how many women appear in the newspaper she retrieved the occurrences of the following pairs and of a few (gender) specific denominations:

> lei/lui, ella/egli, donna/uomo, madre/padre, moglie/marito, mamma/papà, figlia/figlio, femmina/maschio, ragazza/ragazzo, nonna/nonno, cugina/cugino, suocera/suocero, zia/zio, cognata/cognato, amico/amica, bambina/bambino, signora/signore, padrona/padrone, compagna/compagno, impiegata/impiegato, femminile/maschile, prostituta, capo, boss, leader, girl

Whenever it was sensible the wildcard .* was used, in order to retrieve plural forms like *ragazze/ragazzi* and diminutives like *figliola/figliolo*, *ragazzina/ragazzino*, as well. This step led to 440 occurrences of denominations referring to men, against 170 which refer to women.

#### *4.2.2.2. WHICH ARE THE LINGUISTIC MEANS EMPLOYED TO MAKE THEM VISIBLE?*

The second question concerning the linguistic means with which women are made visible was studied first of all on the basis of the suffixes which according to grammars and studies like my own (cf. Burr 1995) are offered by the system for the creation of terms which denominate persons from a perspective of their activity or profession:

> -aio, -aro, -ario, -ino, -tore, -sore, -one, -iere, -aiulo, -aiolo, -ano, -ato, -vendolo, -alaio, -ante, -ista, -ente, -fice, -trice, -aia, -iera, -aiola, -ana, -ara, -aria, -ina, -ona, -ora, -essa, -ata

The explicitly feminine terms which resulted from this query are given in the table below. Obviously, also denominations like *bambina* and *ragazzina* turned up because of the suffix used for their formation but then had to be discarded because they do not belong to the terminology in question:

attrice (3)
danzatrice (1)
direttrice (1)
nuotatrice (2)
organizzatrice (1)
scrittrice (1)
senatrice (1)
sostenitrice (1)
baronessa (1)
dottoressa (1)
presidentessa (1)
poetessa (1)
studentessa (2)
crocerossina (1)
madrina (1)
regina (5)
sovrana (2)
suora (1)
operaia (1)
astrologa (2)
casalinga (1)
collega (1)
candidata (7)

A few more denominations which refer to women like *socia* (1) and *diplomatica* (1) were found by going through the material, others by looking at the co-text of the individual occurrences of formations with *–ista* which by themselves do not indicate the sex of the person. It was found that the terms in the following table actually refer to a woman:

regista (1)
brigatista (1)
inserzionista (1)
cronista (1)
artista (1)
leghista (1)
giornalista (1)

In this way, a total of 51 instances refering to women was put together which compared with the ca. 850 occurrences retrieved for men by using the above given list of suffixes shows that women are clearly underrepresented in the newspaper. Men, on the contrary, are not only present in large numbers but also as performing a very large range of activities. As these would have made up a huge table, for the table below only those terms are retained which appear relatively frequently:

autore (24)
direttore (42)
fondatore (6)
imprenditore (9)
procuratore (17)
scrittore (12)
assessore (15)
presidente (71)
protagonista (11)
regista (18)
ministro (44)
segretario (47)
bersagliere (13)
carabiniere (17)

The next step the student took in order to answer the present question was a co-occurrence search for the combination of the feminin article *la* with non sex specific terms like *presidente* or *leader*. This resulted in the following two examples:

- la presidente della Camera, Irene Pivetti
- la leader dei Verdi

which led her on to the question whether there might be some women hidden also behind denominations which are supposed to be masculine because of their ending or because of the article used together with them. In order to find out about this she carried out another query using the following list of denominations:

> presidente, ministro, deputato, sindaco, magistrato, giudice, procuratore, ambasciatore, funzionario, senatore, consigliere, parlamentare, portavoce, ingegnere, avvocato, pubblicista, interprete, architetto, docente

The result was that *consigliere*, *sindaco*, *presidente*, and *avvocato* each referred on one occasion to a woman:

- il presidente della Camera Irene Pivetti
- anche la moglie è un avvocato di successo
- il sindaco Ariana Cavicchioli
- il consigliere comunale Luisa Camelli

As by then three different ways of denominating a woman presiding had been found, it was only logical that she should ask herself whether there was a difference in meaning between *il/la presidente Pivetti* and *presidentessa* in the following example:

> Questa mostra – spiega Carla Marino, da un anno presidentessa del Comitato provinciale della Croce rossa -, è un'occasione importante per rendere visibile il nostro operato.

The conclusion she came to was that, as some linguists had argued, the different terms were used in accord with the higher or lower prestige of the respective activity.[28]

As *donna* in conjunction with a masculine term in descriptions of Italian is treated as another way to make women discernible and as it was argued in a study that foreign terms were used in most cases in a sex specific way and thus, when they referred to women were either combined with the feminine article *la* or with *donna* (cf. Burr 1995: 153), she then analysed the occurrences of foreign terms like *designer, leader* and *star* and of *donna*. She could, in fact, confirm that foreign terms were mostly used with the feminine article, whereas *donna* appears in those cases where, like in the following example, the exceptionality of the presence of a woman is stressed:

> E di un leader donna che ne dice?

[28] I personally would be tempted to see even a further differenciation in the following terms: *il presidente → la presidente → presidentessa*.

The last step to be undertaken was an investigation into the usage of the different types of names which appear together with the women in the samples retrieved by means of all the afore-mentioned queries. The results obtained confirmed, according to the author, another study carried out on the basis of a larger corpus of Italian newspapers (cf. Burr 1997), where it was argued that in most cases where women appear either their name or their name and surname is used. If the surname is used on its own, the feminine article appears. Her findings were, infact:

| | | |
|---|---|---|
| • name + surname | 30 | |
| • name | 7 | |
| • la + surname | 4 | *la Thatcher*, *la Pivetti* (2), *la Serafino* |

### *4.2.2.3. IN WHICH SECTIONS DO WOMEN APPEAR AND HOW?*

In order to find an answer to the question concerning the sections of the paper women were present in above all, she had to make use of the differenciations made possible by the markup-system with which the corpus had been enriched and which, therefore, can be made visible next to each retrieved occurrence. This lead to the following table:

| | |
|---|---|
| • Cultura/Spettacoli | 40% |
| • Politica | 23% |
| • Cronaca | 31% |
| • Sport | 6% |

Thus women are above all present in the section *Cultura* of the *Corriere della Sera* and much less in sections like *Politica* and *Sport*.

By then analysing more thoroughly the data, she found that female politicians in the section *Politica* are mostly talked about in quite neutral terms. In the section *Cultura*, instead, the appearance and character of women are of great importance and the ideal woman seems, in fact, to possess professionality, intelligence and beauty as in the following example:

> l'inserzionista "laureata, 50enne, manager e donna attraente, intelligente e colta"

Due to the film- and theatre program rwitten about in the section *Spettacoli*, the image of women there is predominantly stereotyped and full of clichés. Women are presented as victims of violence and of the sexual appetites of others or as victims of their own irrationality and unaccountable emotions. A similar picture of women can be found in the *Cronaca*. In the section *Sport*, however, where very few women appear, there are a few cases in which a new conciousness of women is allowed to transcend the reports:

> Le ragazze sono in prima fila; mezzo milione di loro partecipano ai campionati regionali della Federazione Giovanile e si prepara il varo della Lega nel '95.

#### *4.2.2.4. What is their position?*

A query of *moglie* / *mogli* will result in altogether 27 occurrences, whereas *marito* will turn up only 9 times. As Steinert-Schmitz recognised, more than twenty times in this context alone women were portrayed as ornaments of their husbands, indicated by the expression "insieme alla moglie" or variations thereof. The following examples may suffice:

```
(403) razza e cultura, in | compagnia del+la moglie, che per anni ha
(2950)   e le fidanzate de+i giocatori | (la moglie di Bodo Illgner ha
(6787)|  | L' ultimo sfogo con a+l fianco la moglie Alessandra:  "Vogliono
(6800)   l' ultima difesa. a+l suo fianco la moglie | Alessandra, fiera
(9387)         miliardario che insieme al+la moglie Barbara ha aperto uno |
(9421)   partito Philip Gould, insieme al+la moglie Gail | Rebuck.  "Ora
(11815)  Luigi Vecchini, in compagnia del+la moglie, stava aprendo il
(12968)     di Beverly Hills, era con lui la moglie Ginny, con | cui era
(2949)  un | peperone. Mancavano soltanto le mogli e le fidanzate de+i
(5989)   | eta' che tengono per mano le loro mogli. Veronica prende la
```

Apart from this there is also a tendency to define the authority of a woman by means of the position of her husband:

> "La guerra è in città. Non solo in Bosnia o in Ruanda" - continua Carla Martino, che è sorella del ministro degli Esteri -.

## 5. Conclusion

Language can only be studied "in der verbundenen Rede" and even a relatively small sample of speech will, as this study of women in the *Corriere della Sera* shows, allow for more insight than dictionaries and grammars. With her "mühevolle[n], oft ins Kleinliche gehende[n] Elementaruntersuchung" the student was, in fact, able to find out quite a lot about the overall character of the language. As her study shows, the possibilities of the system are, in fact, not just realised in concrete discourse, but they are restricted by the social norm which is itself bound up with a certain world-view, which plays are more or less dominant role according to the domain where language is used.

I am convinced that a study like this could not have been done in the framework of a traditional undergraduate course of Romance linguistics. Instead, it was made possible by the new means at our disposal, in this case an electronic corpus which can be systematically queried in a satisfactorily user-friendly way. This allows students to become curious in the first place and to enter whatever forms come to mind. The first results of a query where a wildcard is used will show them, however, that next to the occurrences they had in mind, forms they had never thought of will be retrieved, as well, and that they, therefore, have to redefine their original definition if they intend to control the complexity of speech. Even if in a traditional course of synchronic linguistics the interest of the students in a certain form, category or usage can be raised, the sheer amount of work implied in a manual study of real speech on the basis of a printed newspaper, for example, would prevent them from following up their curiousity and from defining and redefining certain queries.

The same goes for questions which only come up when working on a research project. As the study presented here shows, quite a few questions were only asked because of a certain result. In order to answer these questions the student actually went through the whole corpus several times. She could only do this, however, because she did not have to reread the whole newspaper. Had she worked on a printed newspaper, instead, it would have been only natural if she had given up asking questions and nobody could have blamed her for that, the research project being part of an undergraduate course and not of a doctoral thesis. In the present case, on the contrary, she was quite prepared to retrieve the data she needed to follow through her own ideas and in the process she learnt how to interpret them.

Apart from making it possible to carry out research projects like the one used here as an example, courses which integrate theory with practice allow for much more participation of the individual student. In fact, every student can look at what the others are doing, ask for help or exchange experiences with others. Whereas in traditional courses all the work for a paper is done in the library or at home and the other members of a course are only presented with the results, here the process of defining and retrieving data is open to everybody and problems and findings can be directly discussed. Thus such a project involves the collective elements of a 'workshop' with individual reflection.

Last but not least, teaching also profits very much from such an organisation of the course. The teacher has, in fact, more time for the individual student. This can be used not only for problem-solving but also for finding out about their reading of the literature which is relevant for their topic. Furthermore, discussing the project or the methodology with them while they are actually working on their research allows for much more insight into the problems involved or the progress they are making, than the traditional counselling during reception hours normally permits.

## Bibliography

Atkins, Sue / Clear, Jeremy, Ostler, Nicholas (1992): „Corpus Design Criteria“, in: *Literary & Linguistic Computing* 7: 1-16.

Barlow, Michael (1996): „Corpora for Theory and Practice“, in: *International Journal of Corpus Linguistics* 1: 1-37.

Bertinetto, Pier Marco (1986): *Tempo, aspetto e azione nel verbo italiano*. Il sistema dell'indicativo. Firenze: Accademia della Crusca.

Burr, Elisabeth (1995): “Agentivi e sessi in un corpus di giornali italiani”, in: Marcato, Gianna (ed.): *Donna & Linguaggio*. Convegno Internazionale di Studi, Sappada / Plodn (Belluno), 26-30 giugno 1995. Padova: CLEUP 141-157.

Burr, Elisabeth (1997): “Neutral oder stereotyp. Referenz auf Frauen und Männer in der italienischen Tagespresse”, in: Dahmen, Wolfgang / Holtus, Günter / Kramer, Johannes / Metzeltin, Michael / Schweickard, Wolfgang / Winkelmann, Otto (eds.): *Sprache und Geschlecht in der Romania*. Romanistisches Kolloquium X (= TBL 417). Tübingen: Narr 133-179.

Chafe, Wallace (1992): „The importance of corpus linguistics to understanding the nature of language“, in: Svartvik, Jan (ed.): *Directions in Corpus Linguistics*. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991 (= Trends in Linguistics. Studies and Monographs 65). Berlin / New York: Mouton de Gruyter 79-97.

Clear, Jeremy (1992): „Corpus sampling“, in: Leitner, Gerhard (ed.): *New Directions in English Language Corpora*. Methodology, Results, Software Developments (= Topics in English Linguistics 9). Berlin / New York: Mouton de Gruyter 21-31.

Coseriu, Eugenio (1988a): *Einführung in die allgemeine Sprachwissenschaft* (= UTB 1372). Tübingen: Francke.

Coseriu, Eugenio (1988b): *Sprachkompetenz: Grundzüge der Theorie des Sprechens* (= UTB 1481). Tübingen: Francke.

Crowdy, Steve (1993): „Spoken Corpus Design“, in: *Literary & Linguistic Computing* 8: 259-265.

De Mauro, Tullio (1975 / $^{3}$1981): *Scuola e linguaggio*. Questioni di educazione linguistica. Roma: Editori Riuniti.

De Mauro, Tullio (1993): „Le scelte per la costituzione del corpus“, in: De Mauro, Tullio/Mancini, Federico/Vedovelli, Massimo/Voghera, Miriam: *Lessico di frequenza dell’italiano parlato*. Ricerca a cura dell’Osservatorio linguistico e culturale italiano OLCI dell’Università di Roma „La Sapienza“ (= Collana Fondazione IBM). Milano: ETASLIBRI 29-32.

Fillmore, Charles J. (1992): „’Corpus linguistics’ or ‘Computer-aided armchair linguistics’“, in: Svartvik, Jan (ed.): *Directions in Corpus Linguistics*. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991 (= Trends in Linguistics. Studies and Monographs 65). Berlin / New York: Mouton de Gruyter 35-60.

Halliday, Michael A. K. (1992): „Language as system and language as instance: The corpus as a theoretical construct“, in: Svartvik, Jan (ed.): *Directions in Corpus Linguistics*. Proceedings of the Nobel Symposium 82, Stockholm, 4-8 August 1991 (= Trends in Linguistics. Studies and Monographs 65): Berlin / New York: Mouton de Gruyter 61-77.

Hockett, Charles F. (1958): *A Course in Modern Linguistics*. New York: Macmillan.

Humboldt, Wilhelm von (1827-1829 / 1963): „Über die Verschiedenheiten des menschlichen Sprachbaus“, in: Humboldt, Wilhem von: *Schriften zur Sprachphilosophie* (= Werke in fünf Bänden III). Darmstadt: Wissenschaftliche Buchgesellschaft 144-367.

Hymes, Dell H. (1972): „On Communicative Competence“, in: Pride, J. B. / Holmes, Janet (eds.): *Sociolinguistics* (= Penguin Modern Linguistics Readings). Harmondsworth: Penguin 269-293.

Koch, Peter / Oesterreicher, Wulf (1990): *Gesprochene Sprache in der Romania: Französisch, Italienisch, Spanis*ch (= Romanistische Arbeitshefte 31). Tübingen: Niemeyer.

Leitner, Gerhard (1992): „International Corpus of English: Corpus design - problems and suggested solutions“, in: Leitner, Gerhard (ed.): *New Directions in English Language Corpora*. Methodology, Results, Software Developments (= Topics in English Linguistics 9). Berlin / New York: Mouton de Gruyter 33-64.

Marcos-Marín, Francisco A. (1994): *Informática y Humanid*ades. Madrid: Gredos.

Nencioni, Giovanni (1990): „The Accademia della Crusca: New Perspectives in Lexicography“, in: *Computers and the Humanities* 24: 345-352.

Quirk, Randolph / Greenbaum, Sidney ($^4$1975*): A University Grammar of English*. London: Longman.

Renzi, Lorenzo (ed.) (1989): *Grande grammatica italiana di consultazione I: La frase. I sintagmi nominale e preposizionale*. Bologna: Il Mulino.

Sampson, Geoffrey (1993): „The Need for Grammatical Stocktaking“, in: *Literary & Linguistic Computing* 8: 267-273.

Simone, Raffaele (1996): „Italiano sì ma quale?“, in: *Italiano & Oltre* XI: 260-261.

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

Sinclair, John M. (1992): „The automatic analysis of corpora“, in: Svartvik, Jan (ed.): *Directions in Corpus Linguistics*. Proceedings of the Nobel Symposium 82, Stockholm, 4-8 August 1991 (= Trends in Linguistics. Studies and Monographs 65): Berlin / New York: Mouton de Gruyter 379-397.

Steinert-Schmitz, Sabina (1998): "Analyse sprachlicher und inhaltlicher Besonderheiten bei der Referenz auf Frauen im Corriere della Sera", Seminararbeit. Gerhard-Mercator-Universität GH Duisburg.

Stubbs, Michael (1993): „British Traditions in Text Analysis. From Firth to Sinclair“, in: Baker, Mona / Francis, Gill / Tognini-Bonelli, Elena (eds.): *Text and Technology*. In Honour of John Sinclair. Philadelphia / Amsterdam: John Benjamins 1-33.

Svartvik, Jan (1992): „Corpus linguistics comes of age“, in: Svartvik, Jan (ed.): *Directions in Corpus Linguistics*. Proceedings of the Nobel Symposium 82, Stockholm, 4-8 August 1991 (= Trends in Linguistics. Studies and Monographs 65): Berlin / New York: Mouton de Gruyter 7-13.

Voghera, Miriam (1993): „Le variabili testuali e pragmatiche“, in: De Mauro, Tullio / Mancini, Federico / Vedovelli, Massimo / Voghera, Miriam: *Lessico di frequenza dell’italiano parlato*. Ricerca a cura dell’Osservatorio linguistico e culturale italiano OLCI dell’Università di Roma „La Sapienza“ (= Collana Fondazione IBM). Milano: ETAS LIBRI 32-38.