Cécile Desoutter, Dorothee Heller & Michele Sala (eds)

Corpora in specialized communication
Korpora in der Fachkommunikation
Les corpus dans la communication spécialisée

Indice

I corpora in contesti pedagogici

I corpora in contesti legali

I corpora in contesti professionali

ALESSANDRA MOLINO

1. Compiling a Stratified Corpus for the Cross-cultural Study of Academic Writing: Methodological Challenges and Research Opportunities

1. Introduction

This chapter explores a crucial methodological issue in cross-cultural research: the challenge of comparability. Comparability has always been a highly debated topic in Contrastive Rhetoric (CR), particularly when, in the second half of the 1980s, the focus moved from the analysis of second language writing to comparisons of authentic professional texts in English and in other languages (Leki 1991). As this approach developed, researchers became increasingly aware of the necessity to compare instances of the same text type across languages. Thanks to studies in text and genre analysis, which provided sounder bases to identify equivalent texts, and the contribution of corpus linguistics, which allowed researchers to collect corpora for quantitative descriptive analyses, CR moved significantly beyond anecdotal evidence or speculative conclusions on culturally-determined writing features (Connor 2002).

In order to make CR research more rigorous and consistent, Mauranen (2001) and Yakhontova (2006) have argued for an approach to rhetorical differences aiming at providing empirical evidence of culture-specific patterns rather than seeking explanations for dissimilarities. Such a focus on quantitative descriptive analysis preserves the issue of comparability at the top of the agenda of cross-cultural research. Among the challenges that need to be faced are those of corpus design and the control of significant contextual

variables. As Moreno (2008: 34-37) observes, we need to control contextual factors statistically by carefully designing the corpus for analysis and choosing an effective approach to sampling. Moreno suggests adopting stratified sampling, as this technique enables researchers to control the impact of extra-linguistic factors that may interfere with the variable of writing culture (e.g. topic, text superstructure, gender).

The aim of this chapter is to test the feasibility of stratified sampling in a special situation in cross-cultural research: the comparison of two languages with unequal chances of being used for the same purposes, such as English and Italian for research and publication. Given the predominant role of English as an international language for academic communication, this situation characterises other European contexts too, where local languages risk losing domain to English (see for instance Pérez-Llantada *et al.* (2011) for the Spanish academic context).

One of the main difficulties that may hinder the use of stratified sampling in such contexts is the different availability of data to investigate relevant contextual factors and compile maximally similar corpora. This chapter addresses this issue by trying to answer the following questions:

- Can stratified sampling be applied to the collection of corpora from target populations of different size?
- What are the practical constraints that influence ideal corpus design?
- How can those difficulties be overcome?
- What consequences do possible compromises on corpus design have on the validity of results?
- Is quantitative descriptive research still meaningful when maximal comparability is difficult to achieve?

In order to answer these questions, the technique of stratified sampling will be applied to the collection of a corpus of Linguistics research articles (RAs) in English and Italian. The difficulties encountered during the compilation will be discussed so as to examine the degree to which stratification is possible in this context. The usefulness of

corpus stratification will be illustrated through a case study of 'reader engagement' (Hyland 2005a) as signalled by the use of inclusive authorial references in subject position.

This chapter is organized as follows. By resorting to Connor/ Moreno's (2005) notion of *tertium comparationis*, Section 2 discusses the importance of identifying appropriate criteria of comparison during the choice of primary data for analysis. The technique of stratified sampling will be explained in sub-section 2.1. Section 3 describes the comparable corpus of Linguistics RAs in English and Italian and comments on the practical constraints that affected the design of the sampling frame. Section 4 presents the results of the study of reader engagement. Section 5 concludes commenting briefly on the issues of sampling, balance and comparability.

2. Contrastive research and the notion of *tertium comparationis*

The aim of quantitative descriptive cross-cultural investigations is to assess the impact of the independent variable 'writing culture' on the dependent variable 'form and content of the text' (Moreno 2008). The label 'writing culture' refers to the set of norms, values and common practices of a particular community, which are expressed through a given language code. Cross-cultural investigations need appropriate parameters of comparison, or *tertia comparationis* (Connor/Moreno 2005), a requirement that is particularly important when looking for differences, as these become apparent only if matched against a background of likeness being shared cross-culturally.

Connor/Moreno (2005) argue that *tertia comparationis* should be placed at two major phases: 1) when choosing the primary data for comparison and 2) when establishing comparable textual concepts. The choice of primary data, which is the focus of this chapter, involves the identification of the target population and the parameters of the study. In cross-cultural investigations, the parameters to include

in the sampling frame are those extra-linguistic factors that Moreno (2008) calls 'confounding factors' because, if not controlled during the sampling, they may interfere with the effect of the main independent variable under study, i.e. language as a vehicle of culture.

In descriptive studies of academic discourse as well as in cross-cultural analyses of academic writing, the principal contextual factors that have been identified are genre (Swales 1990), discipline (Hyland 2000, Hyland/Bondi 2006), textual macrostructure (Tarone *et al.* 1998), date of publication (Salager-Meyer *et al.* 2003) and size of the audience (Burgess 2002). Moreno (2008: 34) argues that "if the design is able to maintain constant the values of […] confounding factors and manages to include the same proportion of texts representing those values in each sample, we can then say that the two corpora are equivalent to the maximum degree of similarity".

2.1. Stratified sampling techniques

An important issue characterising the debate over corpus design is the nature of the relationship between the variables in the study (Hunston 2008: 157). In other words, corpus designers have to decide whether each sub-corpus should contain a balanced number of texts, or tokens, or whether the amount of text material in each sub-corpus should be proportionate to its significance in the target population. Generally, it is the purpose of the study that determines the choice.

In a corpus designed to be representative, the function of variables is to increase coverage; therefore, the corpus designer may wish not only to collect a large amount of data, but also to allocate text material in proportion to its significance in the target population. For example, in Italy, in the macro area of the Humanities, about 22% of publications are journal articles, 17% is constituted by book chapters, 60% by books and the remaining 1% is constituted by other

minor research products.[1] In a corpus aiming at being representative of this specific target population, sub-corpora should mirror the above proportions. The procedure best suited to this frame is stratified random sampling with proportional allocation. Using this technique, texts are chosen randomly from each sub-group, or *stratum*, but the amount of text material reflects its real distribution in the target population. The advantage of this technique is that it increases representativeness. However, it may not be the best choice when the aim is to compare data sets, as the sub-corpora may be significantly different in size. Therefore, in a corpus designed for research with a very specific focus, the compiler may decide to privilege the attribute of balance over representativeness for the sake of comparability.

When balance is an important goal, a uniform allocation of the sample is advisable rather than its proportional distribution. With uniform allocation, each stratum is assigned the same number of texts irrespective of the size of strata in the target population. Once having established the parameters of the sampling frame, the texts to include in the corpus may be collected randomly or using other techniques.[2] The advantage of uniform allocation is that comparisons between corpora within individual strata are possible, because uniform allocation ensures that no contextual factor predominates over the others. In addition, comparisons across strata within each corpus can also be carried out to assess whether and to what extent different contextual factors influence language use.

---

[1]   These data were provided by the Comitato Italiano di Valutazione della Ricerca (CIVR, National Committee for the Evaluation of Research) for the period 2001-2003.

[2]   Dörnyei (2007) observes that in most Applied Linguistics research, probability sampling procedures, such as simple random sampling, are rarely fully achieved. In fact, research in this area often employs non-probability techniques, such as 'quota sampling', which is similar to stratified random sampling "without the 'random' element" (Dörnyei 2007: 98). The first step is usually the design of the sampling frame, followed by the definition of the proportions of the various sub-groups. The samples are then collected, but not necessarily adopting strict random sampling procedures. Rather, compilers meet the quotas by selecting participants they can have access to.

The disadvantage of uniform allocation is that it reduces representativeness, as the smaller strata in the population are promoted and the larger strata are demoted. A corpus designed for the purpose of comparison, therefore, does not provide data reflecting the actual distribution of language features in the target population.

3. Corpus design

This section describes the sampling frame of the comparable corpus of Linguistics RAs in English and Italian used in this study. Table 1, which is inspired in its display by the example provided in Connor/Moreno (2005: 159), illustrates the sampling frame; Table 2 provides quantitative information about the corpus.

Tertium comparationis	*Value of prototypical feature perceived as constant across corpora*	*Number of texts in each independent corpus*
Genre	Research article	32
Discipline	Linguistics	32
Text superstructure	Introduction – Method – Results – Discussion	16
	More variable superstructure (logico-argumentative papers)	16
Gender	Women	16
	Men	16

Table 1. Sampling frame for genre, discipline, text superstructure and gender.

Each independent language corpus is composed of 32 texts, meaning that the corpus as a whole contains 64 RAs. The first parameter is that of genre: only original research papers were collected, excluding other related genres such as research reports or review articles. The reason is that different overall discourse goals may affect rhetorical and linguistic configurations.

	Number of texts	*Total number of words*	*Average text length*
Linguistics RAs in English	32	215,461	6,733
Linguistics RAs in Italian	32	223,229	6,976

Table 2. Comparable corpus of Linguistics RAs in English and Italian.

The second constant feature is discipline. The influence of discipline on discourse practices has been documented in numerous studies (e.g. Hyland 2000, Hyland/Bondi 2006). Therefore, only papers that can be described using the umbrella term of 'Linguistics RAs' were collected. However, the sample was not stratified in terms of sub-discipline (e.g. sociolinguistics, applied linguistics, pragmatics) or topic.

The third *tertium comparationis* is text superstructure. This parameter refers to the overall organization of discourse in macro-rhetorical units. Text superstructure may have a considerable impact on language use. In this sampling frame, RAs were grouped according to two sub-groups. One sub-corpus collects 16 papers conforming to the so-called IMRAD structure (Swales 1990). The IMRAD structure is typical of articles featuring an Introduction, a Method section, and Results and Discussion sections (conflated together or in separate units). All the papers in this sub-corpus present a clearly identifiable section reporting on the methods and procedures, and an equally unmistakable section presenting the research findings. However, the methodologies are varied, ranging from corpus analysis to fieldwork. It can be hypothesised that different types of approach entail different rhetorical strategies in reporting procedures. This aspect will have to be considered when analysing results. The second sub-corpus includes papers with a more mixed configuration. They are predominantly logico-argumentative in nature, following macro patterns such as the Problem-Solution(-Evaluation) pattern or the Situation-Analysis pattern.

Finally, I also controlled the individual factor of gender. The relationship between language and gender has received considerable

attention in the study of language variation. Language has been described as an important means by which gender is enacted (see Ehrlich 2004). For this reason, I included this feature in the design of the corpus, collecting 16 RAs written by women and 16 written by men.

Having described the composition of the comparable corpus, I now turn to some remarks on the difficulties encountered during the sampling. The first complication was the comparatively limited availability of data in Italian. There are various reasons for such an imbalance. The first is that the number of specialized journals in Italian is much more restricted. A cursory glance at lists of publications such as the ERIH (European Reference Index for the Humanities) Initial List of Linguistics journals (2007) will prove ample evidence for this phenomenon.[3] Scrolling down the list, I counted 246 journals with an English title, 58 journals with a French title, 47 journals with a German title, 44 publications with a Spanish title, 25 journals with an Italian title and 10 with a Portuguese title. Due to the different population size, there may be some relevant contextual factors, such as sub-discipline or topic, that do not recur frequently enough in the smallest population to be able to collect an adequately large sample of texts.

In the case of the present corpus of Linguistics RAs, the very dependent variable of language was at risk. It has become common practice for Italian journals to accept papers in other languages (mainly in English). As a result, in numerous latest issues, articles in Italian are rare. My original plan was to collect papers reflecting the current use of language by taking texts from 2008 to the present. However, given the constraint of the sampling frame, this was not possible. Therefore I decided to extend the time span of the study and to go back to 2001 to collect the desired amount of texts.

Although going further back in time may be a solution, the disadvantage is that over a long period of time, rhetorical conventions

---

3   The 2007 list and the 2011 list with revised journal categories are available at: <https://www2.esf.org/asp/ERIH/Foreword/search.asp> (last accessed 26 June 2013).

may change, thus turning the time coordinate into another significant extra-linguistic influence. In order to partially avoid this risk, in the present corpus the bulk of articles was taken from the 2003-2008 period (see Appendix). In so doing, the isolated cases of more recent or more dated works can be compared to the general tendencies of the corpus to verify whether they diverge considerably, thus skewing overall counts.

A second complication related to the dependent variable of language is that while in Italian journals the population of scholars writing in Italian is rather homogenous, as almost all of them are Italian native speakers, the same cannot be said for international journals, where scholars from a variety of first language backgrounds publish in English. Ideally, the value of native-speaker status should be kept constant. In studies of academic writing, the name and affiliation of the writer have often been considered as parameters to control this factor (Hewings *et al.* 2010, Molino 2010). The same was done in the present study. However, in order to respect the sampling frame, some exceptions were made. In the Italian corpus, one article is written by an Italian native speaker based in the UK; on the other hand, in the English corpus, while all articles are written by authors based in Anglo-American universities, some of them are signed by authors who may well not be native speakers of English, judging from their name. In those cases, before including the paper in the corpus, the scholars' academic profile was checked on the Internet and when a mostly Anglo-American profile emerged, the article was included. The assumption is that English being these scholars' professional language, their written productions are likely to encode the values and expectations of the Anglo-American academic writing culture. Nevertheless, special attention should be paid to the quantitative data coming from such texts to verify whether they diverge significantly from the rest of the corpus.

4. A cross-cultural study of reader engagement in English and Italian Linguistics RAs

This section presents a cross-cultural study of reader engagement as signalled by the use of inclusive *we* and *noi* [we] in subject position. I will refer to these engagement devices as 'inclusive authorial references' because while in English the subject is always expressed, in Italian, person and number are generally signalled through verb endings.

4.1. Previous research on reader engagement

In recent years, studies on academic discourse have focused on its interactive nature demonstrating that the often-claimed neutrality of academic writing is not borne out of empirical evidence. Writers construct for themselves appropriate academic *personae* and demonstrate awareness of their audience by considering their readers' expectations, anticipating their objections and providing metadiscourse guidance (Hyland 2005a). Hyland (2005a) refers to the devices used to recognize the presence of readers as 'engagement markers'. The most direct engagement strategy is the use of reader pronouns (*you* and *your*). However, as Hyland notes, these forms are less typical in academic writing than in other more informal registers. In academic texts, readers and writers are often bound together through the use of inclusive *we*, whose overall function is to claim solidarity and membership (Hyland 2005b: 364).

Reader engagement has often been studied in terms of student and expert interaction (Freddi 2005, Hyland 2005b, Mei 2007). Other studies include comparisons across disciplines (Fløttum *et al.* 2006) and between popular and professional science (Hyland 2010). Engagement markers have also been investigated cross-culturally in Spanish and English Business Management RAs (Mur Dueñas 2008, 2011) and, more indirectly as part of an analysis of first person

markers, in native English, native Italian and non-native English medical editorials (Giannoni 2008).

With regard to gender-related traits, there has been little research, probably because of the scarcity of gender-related studies in academic discourse in general, as noted by Sala and D'Angelo (2009). Exceptions are, precisely, Sala and D'Angelo's (2009) study of RAs in English in the fields of Applied Linguistics, Economics and Law, and Tse and Hyland's (2006) investigation of academic book reviews in Philosophy, Biology and Sociology. These studies are relevant to the present investigation of engagement as in both of them, men were found to be more reader-oriented than women, using engagement markers more often.

4.2. Method

Inclusive authorial references were retrieved using the Concord Tool of WordSmith Tools 6.0 (Scott 2011). As for English, it was sufficient to search for all the occurrences of the pronoun *we*, whereas in the case of Italian, all the relevant verb endings (i.e., *-amo*, *-emo*, *-imo*, *-mmo*, *-ssimo*) were retrieved. A disambiguation process followed, aiming at distinguishing between inclusive authorial references and forms of editorial *we/noi* and *pluralis majestatis*. Finally, occurrences were classified and quantified according to semantic referent and discourse function.

4.3. Results

4.3.1. Frequency of use of inclusive authorial references and functional analysis

Table 3 shows the incidence of use of inclusive authorial references in subject position in English and Italian Linguistics RAs. Italian writers appear more eager than their Anglo-American colleagues to use this engagement strategy. This result is not directly comparable to other findings in the literature. However, in a study of metadiscourse in Business Management RAs, Mur Dueñas (2011) found that Spanish

writers, too, use inclusive authorial references more often than Anglo-American scholars. Her interpretation is that Spanish scholars are aware of addressing a fairly small and homogeneous audience, which would make them more inclined to appeal to mutual disciplinary understandings. This hypothesis may also hold for the present comparison; however, in order to gain a deeper understanding of differences, it may be useful to investigate how data are distributed across semantic referents and discourse functions.

	Inclusive authorial references	
	Raw	*Normalised*
Linguistics RAs in English	241	11.16
Linguistics RAs in Italian	358	16.04

Table 3. Occurrences of inclusive authorial references (raw and normalised figures per 10,000 words).

Four main semantic referents were identified as being encapsulated in inclusive forms: general academic referents (i.e. the other members of the disciplinary community), readers, language users and general non-academic referents (e.g. 'we as people/humans'). Overall, I found a correspondence between type of referent and discourse function, namely:

- *Appealing to the disciplinary background.* When academic referents are brought into the discourse, the function of inclusive forms is to appeal to the disciplinary background in terms of communal knowledge, values, objectives, interests, problems and expectations, as can be seen in the following examples:

(1)   I have argued that *we* need a more dynamic view of language and of its learning. (AL4)

(2)   To examine language socialization within a community of practice framework, it is necessary to look at the micro-interactional level of ongoing practical activity. This means that *we* have to come to understand what people do. (IJAL6)

- *Elaborating an argument.* When writers resort to inclusive forms to elaborate their argument, they ask for a collaborative construction of meaning by readers, as if the discussion were elaborated simultaneously by both participants, as in examples (3) and (4):

(3) If *we* take the first premise of each case to be a (very) simple context C, in C (22), *we* can say that P is relevant in its own right […]. (L2)

(4) *We* can thus explain why (45), the Finnish counterpart to (1) with a null subject, repeated here, is ill-formed […]. (SL4)

- *Providing metadiscourse guidance*. In this case, too, readers are invoked as semantic referents. Writers guide them through the text thanks to previews and reviews, and by indicating where they can find examples and other information (e.g. tables and appendices), as in (5) and (6):

(5) Although later studies showed some variability across L1/L2, and *we* will consider this in more detail towards the end of this paper […]. (AL2)

(6) *We* have also seen in (12), however, that it can at times serve simultaneously to reduce speaker responsibility […]. (L2)

- *Making generalizations.* Writers refer to typical situations using *we* to include language users or non-academic referents in general.

(7) [Participants] were also told that they would probably not detect every contradiction, as *we* humans have been telling stories for a long time […]. (IJAL7)

(8) How does Derrida inform our thinking in Applied Linguistics? He shows how completely *we*, even we linguists, take language for granted […]. (AL5)

Table 4 presents the frequency of occurrence of inclusive authorial references across discourse functions.

Discourse function	*English*	*Italian*
Appealing to the disciplinary background	4.08	1.97
Elaborating an argument	3.85	6.94
Providing metadiscourse guidance	1.49	5.96
Making generalizations	1.76	1.16

Table 4. Inclusive authorial references across functions (normalised figures per 10,000 words).

The results show quite distinct patterns across the two languages. Only the function of *making generalizations* is similar in frequency in English and Italian. Anglo-American linguists seem to use inclusive references mainly to claim disciplinary solidarity and membership, and to engage readers in the elaboration of the argument. Italian writers, on the other hand, make a more modest use of appeals to solidarity, but engage readers to a greater extent, by involving them in the construction of the line of reasoning and providing more metadiscourse guidance.

4.3.2. The role of text superstructure

Tables 5 and 6 provide the same results as Table 4, but broken down in terms of text superstructure.

Discourse function	*IMRAD English*	*IMRAD Italian*
Appealing to the disciplinary background	1.67	1.25
Elaborating an argument	2.04	3.49
Providing metadiscourse guidance	1.07	3.94
Making generalizations	0.79	0.58

Table 5. Inclusive authorial references across functions (IMRAD text superstructure). (Normalised figures per 10,000 words).

Discourse function	LA English	LA Italian
Appealing to the disciplinary background	2.37	0.72
Elaborating an argument	1.81	3.45
Providing metadiscourse guidance	0.70	2.02
Making generalizations	0.70	0.58

Table 6. Inclusive authorial references across functions (LA text superstructure). (Normalised figures per 10,000 words).

In both Tables 5 and 6, inclusive authorial references in Italian texts tend to correlate with the functions of *elaborating an argument* and *providing metadiscourse guidance*. Therefore, it seems that the association of inclusive authorial references with these two functions is a crosscutting writing habit of the community of Italian linguists, regardless of the macro rhetorical configuration of RAs. Equally, the preference for the association of inclusive authorial references with the functions of *appealing to the disciplinary background* and *elaborating an argument* is found in English irrespective of the sub-type of RA. These results seem to corroborate those obtained in Table 4, indicating that the two writing cultures follow quite distinct patterns of use of inclusive authorial references.

These findings, however, do not imply that text superstructure has no influence on discourse patterns. Tables 5 and 6 indicate that it may affect the relative frequency of functions in texts. This is particularly evident in English IMRAD papers, where the most common use of inclusive authorial references is to elaborate an argument and not to appeal to the disciplinary background, as appeared in Table 4. This difference suggests that if text superstructure is not properly balanced, it may affect overall results.

4.3.3. The role of gender

Tables 7 and 8 show the results of Table 4 broken down in terms of gender distinctions. Contrary to expectations, both Italian and Anglo-American male scholars resort to inclusive authorial references slightly less than women, although the difference in the Italian sub-corpus is more noticeable. As for discourse functions, the most striking data is the marked preference for the function of *appealing to*

the disciplinary background by female Anglo-American writers. Therefore, it would seem that gender affects writing not only in terms of frequency of use of items, but in English, in particular, also in terms of preferred form-function association patterns.

Discourse function	English (women)	Italian (women)
Appealing to the disciplinary background	3.11	0.81
Elaborating an argument	1.53	3.99
Providing metadiscourse guidance	0.32	3.18
Making generalizations	0.97	0.63
TOT	5.93	8.61

Table 7. Inclusive authorial references across functions in RAs by women (normalised figures per 10,000 words).

Discourse function	English (men)	Italian (men)
Appealing to the disciplinary background	0.97	1.16
Elaborating an argument	2.32	2.96
Providing metadiscourse guidance	1.44	2.78
Making generalizations	0.51	0.54
TOT	5.24	7.44

Table 8. Inclusive authorial references across functions in RAs by men (normalised figures per 10,000 words).

In order to assess whether gender alone is responsible for this discrepancy between female and male English writers, I considered results obtained for individual texts. Two texts by female writers, i.e. AL4 and AL5 (see Appendix), feature a considerably high use of inclusive *we*, i.e. 30 and 40 occurrences respectively, while the frequencies of the rest of the articles by women range between 1 and 18 hits. This divergence suggests that other factors than gender may account for the high use of inclusive authorial references in AL4 and AL5.

The question, therefore, is why AL4 and AL5 feature such an extensive use of inclusive *we,* particularly for the function of appealing to the disciplinary background (15 and 24 raw occurrences respectively), thus skewing the overall results of Table 4. In the

attempt to answer this question, I will consider the impact of aspects such as text length, superstructure, native-speaker status and date of publication. Text length may be an influencing factor in the case of AL4, which totals 7,631 words; however, it seems not to be relevant for AL5, which counts 6,813 words, thus not diverging significantly from average text length, i.e. 6,733. Text superstructure may also play a role, at least in part. Despite being shorter, AL5 (LA paper) features a higher number of inclusive *we* than AL4 (IMRAD paper), which might suggest that the logico-argumentative nature of AL5 encourages more appeals to communal knowledge. Nevertheless, the frequencies of *we* in AL4, too, are well above the average of the IMRAD papers in the corpus. Attributing the high use of *we* to the supposedly non-native status of writers (see section 3) is equally disputable. While one writer may actually not be a native speaker of English (AL5), web-based searches confirmed that the other is a native speaker (AL4). Finally, the time coordinate does not seem to be a significant factor either, as AL4 was published in 2006 and AL5 in 2008. Therefore, the high use of *we* to appeal to the disciplinary background seems to be related to either individual idiosyncrasies or to other contextual factors that could not be controlled during the sampling process, such as sub-discipline and topic.

This analysis suggests that individual divergences may have a great impact on overall results, not only in terms of frequency but also in terms of distribution patterns. Consequently, it becomes difficult to distinguish the role of idiosyncrasies or factors such as topic from that of writing culture. In the Italian corpus, too, there are articles in which the overall incidence of use of inclusive authorial references is particularly high (i.e. AGI5, 46 occurrences and RID5, 69 hits). For this reason, I decided to subtract AL4 and AL5 from the English corpus and, in order to preserve balance, I also excluded two papers by male writers, one per each type of text superstructure (L5 and SL2), in which the occurrence of use of *we* was considerably high. I did the same in the Italian sub-corpus (AGI5, RID5, RID7, SGI1). The resulting corpus contains 28 RAs in each sub-corpus and Table 9 summarises the results obtained from its analysis.

Discourse function	English (28 texts)	Italian (28)
Appealing to the disciplinary background	2.23	0.99
Elaborating an argument	2.51	3.85
Providing metadiscourse guidance	0.70	3.40
Making generalizations	0.79	0.76
TOT	6.23	9.00

Table 9. Inclusive authorial references across functions in English and Italian (28 texts). (Normalised figures per 10,000 words).

In English, the discrepancy between the two discourse functions of *appealing to the disciplinary background* and *elaborating an argument* is not only less evident as compared to the data in Table 4, but their reciprocal significance is inverted, with the function of *elaborating an argument* appearing as slightly more frequent. In Italian, on the other hand, the data relating to the articles omitted had an impact on total frequencies but not on the description of preferred patterns of writer-reader interaction, which remain unvaried. The data displayed in Table 9 are more likely to reflect the role of writing culture since the articles that skewed the results have been omitted.

4.4. Summary of cross-cultural findings

This exploratory study of reader engagement indicates that in both English and Italian Linguistics RAs, inclusive authorial references are most often used to engage readers in the construction of the argument, so as to make them appear active and collaborative discourse participants. Anglo-American and Italian linguists, however, differ in the extent to which they resort to inclusive authorial references to appeal to communal background and to provide metadiscourse guidance. While Anglophone scholars prefer to engage their readers to refer to shared knowledge, values, problems and expectations, Italian linguists prefer to involve readers when it comes to the construction of the text, as if sharing the responsibility for its organization, coherence and clarity.

5. Concluding remarks

The analysis of inclusive authorial references reported in this chapter has illustrated how a stratified corpus can be used in the cross-cultural study of academic discourse. The choice of a stratified corpus enables researchers to control the interference of contextual factors. A balanced corpus with uniform allocation also enables analysts to compare corpora within individual strata, as was done in 4.3.2 and 4.3.3, in order to investigate the impact of specific contextual factors on overall results. Finally, a balanced corpus facilitates the task of evaluating the influence of individual texts and uniform allocation enables researchers to fine-tune the corpus so as to arrive at a maximum grade of comparability.

Differences were found between English and Italian Linguistics RAs. Overall, Italian linguists resort to inclusive authorial references more often than Anglo-American scholars. In addition, in the two writing communities inclusive authorial references tend to be used for somewhat dissimilar discursive strategies. The analysis of the results has also shown that some articles were significantly divergent from corpus tendencies, thus skewing overall quantitative data. If on the one hand, those divergences may indicate that within the two communities more marked linguistic patterns are acceptable, on the other hand, they also suggest that in further investigations of academic writing, sub-discipline and topic should be considered as parameters to include in the sampling frame, if enough text material is available.

This issue is relevant for the last question posed in the Introduction: is quantitative descriptive research still meaningful when maximal comparability is difficult to achieve? If a contextual variable is too infrequent to lend itself to be sampled, other methods of inquiry will best suit its investigation, such as case studies or ethnographic analyses. Nevertheless, as Leech observes (2007: 143-144), if our corpus design goals cannot be fully satisfied,

> it is best to recognise that these goals are not an all-or-nothing: there is a scale of representativity, of balancedness, of comparability. We should seek to

> define realistically attainable positions on these scales, rather than abandon them altogether.

In a situation such as the one illustrated in this chapter, where several major contextual factors were controlled, the difficulty of achieving maximal comparability may be overcome if the design criteria are explicitly described so as to be able to identify areas of imbalance and their potential impact on results.

References

Burgess, Sally 2002. Packed Houses and Intimate Gatherings: Audience and Rhetorical Structure. In Flowerdew, John (ed.) *Academic Discourse*. Harlow: Longman, 196-215.

Connor, Ulla 2002. New Directions in Contrastive Rhetoric. *TESOL Quarterly* 36, 493-510.

Connor, Ulla / Moreno, Ana I. 2005. Tertium Comparationis: A Vital Component in Contrastive Research Methodology. In Bruthiaux, Paul / Atkinson, Dwight / Eggington, William G. / Grabe, William M. / Ramanathan, Vaidehi (eds) *Directions in Applied Linguistics: Essays in Honor of Robert B. Kaplan*. Clevendon: Multilingual Matters, 153-164.

Dörnyei, Zoltán 2007. *Research Methods in Applied Linguistics*. Oxford: Oxford University Press.

Ehrlich, Susan 2004. Language and Gender. In Davies Alan / Elder, Catherine (eds) *Handbook of Applied Linguistics*. Oxford: Blackwell, 304-327.

Fløttum, Kjersti / Kinn, Torodd / Dahl, Trine 2006. 'We Now Report On...' Versus 'Let Us Now See How...': Author Roles and Interaction with Readers in Research Articles. In Hyland, Ken / Bondi, Marina (eds) *Academic Discourse Across Disciplines*. Bern: Peter Lang, 203-224.

Freddi, Maria 2005. Arguing Linguistics: Corpus Investigation of One Functional Variety of Academic Discourse. *Journal of English for Academic Purposes* 4, 5-26.

Giannoni, Davide S. 2008. Medical Writing at the Periphery: The Case of Italian Journal Editorials. *Journal of English for Academic Purposes* 7, 97-107.

Hewings, Anne / Lillis, Theresa / Vladimirou, Dimitra 2010. Who's Citing Whose Writings? A Corpus Based Study of Citations as Interpersonal Resource in English Medium National and English Medium International Journals. *Journal of English for Academic Purposes* 9, 102-115.

Hunston, Susan 2008. Collection Strategies and Design Decisions. In Lüdeling, Anke / Kytö, Merja (eds) *Corpus Linguistics: An international Handbook*, Volume 1. Berlin: Mouton de Gruyter, 154-168.

Hyland, Ken 2000. *Disciplinary Discourses: Social Interactions in Academic Writing*. London: Pearson.

Hyland, Ken 2005a. Stance and Engagement: A Model of Interaction in Academic Discourse. *Discourse Studies* 7/2, 173-192.

Hyland, Ken 2005b. Representing Readers in Writing: Student and Expert Practices. *Linguistics and Education* 16, 363-377.

Hyland, Ken 2010. Constructing Proximity: Relating to Readers in Popular and Professional Science. *Journal of English for Academic Purposes* 9, 116-127.

Hyland, Ken / Bondi, Marina (eds) 2006. *Academic Discourse Across Disciplines*. Bern: Peter Lang.

Leech, Geoff 2007. New Resources, or Just Better Old Ones? In Hundt, Marianne / Nesselhauf, Nadja / Biewer, Carolin (eds) *Corpus Linguistics and the Web*. Amsterdam: Rodopi, 134-49.

Leki, Ilona 1991. Twenty-Five Years of Contrastive Rhetoric: Text Analysis and Writing Pedagogies. *TESOL Quarterly* 25/1, 123-143.

Mauranen, Anna 2001. Descriptions or Explanations? Some Methodological Issues in Contrastive Rhetoric. In Hewings, Martin (ed.) *Academic Writing in Context. Implications and Applications*. London: Continuum, 43-54.

Mei, Wu Siew 2007. The Use of Engagement Resources in High- and Low-rated Undergraduate Geography Essays. *Journal of English for Academic Purposes* 6, 254-271.

Molino, Alessandra 2010. Personal and Impersonal Authorial References: A Contrastive Study of English and Italian Linguistics Research Articles. *Journal of English for Academic Purposes* 9/2, 86-101.

Moreno, Ana I. 2008. The Importance of Comparable Corpora in Cross-Cultural Studies. In Connor, Ulla / Nagelhout, Ed / Rozycki, William (eds) *Contrastive Rhetoric: Reaching to Intercultural Rhetoric*. Amsterdam: John Benjamins, 25-41.

Mur Dueñas, Pilar 2008. Analysing Engagement Markers Cross-Culturally: The Case of English and Spanish Business Management Research Articles. In Burgess, Sally / Martín Martín, Pedro (eds) *English as an Additional Language in Research Publication and Communication*. Bern: Peter Lang, 197-213.

Mur Dueñas, Pilar 2011. An Intercultural Analysis of Metadiscourse Features in Research Articles Written in English and in Spanish. *Journal of Pragmatics* 43, 3068–3079.

Pérez-Llantada, Carmen / Ramón, Plo / Ferguson, Gibson R. 2011. 'You Don't Say What You Know, Only What You Can': The Perceptions and Practices of Senior Spanish Academics Regarding Research Dissemination in English. *English for Specific Purposes* 30, 18-30.

Sala, Michele / D'Angelo, Larissa 2009. Gender- and Authority-related Writing Performance in Academic Research Articles. *Textus* 22/1, 59-76.

Salager-Meyer, Françoise / Alcaraz Ariza, María Angeles / Zambrano, Nahirana 2003. The Scimitar, the Dagger, and the Glove: Intercultural Differences in the Rhetoric of Spanish, French and English Medical Discourse 1930-1995. *English for Specific Purposes* 22, 223-247.

Scott, Mike 2011. *WordSmith Tools* (Version 6.0). Liverpool: Lexical Analysis Software.

Swales, John M. 1990. *Genre Analysis. English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Tarone, Elaine / Dwyer, Sharon / Gillette, Susan / Icke, Vincent 1998. On the Use of the Passive and Active Voice in Astrophysics Journal Papers: With Extensions to Other Languages and Other Fields. *English for Specific Purposes* 17/1, 113-132.

Tse, Polly / Hyland, Ken 2006. Gender and Discipline: Exploring Metadiscourse Variation in Academic Book Reviews. In Hyland, Ken / Bondi, Marina (eds) *Academic Discourse Across Disciplines*. Bern: Peter Lang, 177-202.

Yakhontova, Tatyana 2006. Cultural and Disciplinary Variation in Academic Discourse: The Issue of Influencing Factors. *Journal of English for Academic Purposes* 5, 153-167.

Appendix

Italian sub-corpus

Amenta, Luisa 2006. La perifrasi aviri da /a + infinito nel siciliano contemporaneo. Analisi di un campione di dati ALS. *Rivista Italiana di Dialettologia* XXX, 59-73. [RID1]

Arcangeli, Massimo 2007. Il lessico sportivo e ricreativo italiano nelle quattro grandi lingue europee (con qualche incursione anche altrove). *Studi di Lessicografia Italiana* XXIV, 195-247. [SLI1]

Aureli, Massim 2003. Pressione dell'uso sulla norma: le relative non-standard nei giudizi degli utenti. *Studi Italiani di Linguistica Teorica e Applicata* 32, 45-67. [SILTA1]

Benati, Alessandro 2005. Ricreazione nella classe di lingua. Il modello sperimentale. *Rassegna Italiana di Linguistica Applicata* 37/2-3, 53-73. [RILA1]

Chinellato, Paolo 2003. Sintassi dialettologica e afasia non fluente. Analisi linguistica e strategie riabilitative. *Rivista Italiana di Dialettologia* XXVII, 179-196. [RID2]

D'Achille, Paolo 2005. L'invariabilità dei nomi nell'italiano contemporaneo. *Studi di grammatica italiana* XXIV, 189-209. [SGI1]

Dalosio, Michele 2006.Lingua straniera e sviluppo dei processi di memoria del bambino. *Rassegna Italiana di Linguistica Applicata* 38/2-3, 377-396. [RILA2]

De Cesare, Anna-Maria 2005. La frase pseudoscissa in italiano contemporaneo: aspetti semantici, pragmatici e testuali. *Studi di grammatica italiana* XXIV, 293-322. [SGI2]

Fresu, Rita 2006. Gli uomini parlano delle donne, le donne parlano degli uomini. Indagine sociolinguistica in un campione giovanile di area romana e cagliaritana. *Rivista Italiana di Dialettologia* XXX, 23-53. [RID3]

Grandi, Nicola 2005. I verbi valutativi in italiano tra azione e aspetto. *Studi di grammatica italiana* XXIV, 153-188. [SGI3]

Guazzieri, Annavaleria 2008. Conoscenza dell'apprendimento cooperativo da parte degli insegnanti CLIL. *Rassegna Italiana di Linguistica Applicata* 40/1-2, 91-115. [RILA3]

Lisma, Grazia M. 2007. Sul lessico marinaresco dell'Ottocento. *Studi di Lessicografia Italiana* XXIV, 165-194. [SLI2]

Lo Prete, Rosario 2005. Ricerca sullo studio della L2 nella scuola primaria. *Rassegna Italiana di Linguistica Applicata* 37/2-3, 121-138. [RILA4]

Maiocco, Marco 2004. Varietà e intensità del contatto linguistico: inerzia sintattica nel greco antico semitizzato. *Archivio Glottologico Italiano* 89/1, 84-110. [AGI1]

Marazzi, Martino 2007. Preistoria e storia di "Afro-Americano". *Studi di Lessicografia Italiana* XXIV, 249-264. [SLI3]

Mariottini, Laura 2006. La pragmatica della CMC. Strategie di cortesia linguistica nelle interazioni in chat. *Studi Italiani di Linguistica Teorica e Applicata* 35/2, 319-338. [SILTA2]

Masini, Francesca 2006. Diacronia dei verbi sintagmatici in italiano. *Archivio Glottologico Italiano* 91/1, pp. 67-105. [AGI2]

Mauri, Caterina 2006. Combinazione e contrasto: i connettivi, congiuntivi e avversativi nelle lingue d'Europa. *Archivio Glottologico Italiano* 91/2, 166-202. [AGI3]

Menegale, Marcella 2008. Tipi di domande utilizzate durante la lezione frontale partecipata e output degli studenti. *Rassegna Italiana di Linguistica Applicata* 1/2, 117-145. [RILA5]

Miotti, Renzo 2010. Importazione fonologica in spagnolo e italiano. Riflessioni sul consonantismo. *Studi italiani di linguistica teorica e applicata* 39/3, 491-521. [SILTA3]

Patrucco, Elisa 2003. Sul dialetto in internet. *Rivista Italiana di Dialettologia* XXVII, 139-174. [RID4]

Pinto, Immacolata 2004. Alcune osservazioni sul prefisso in- negativo nel sardo e in area romanza. *Rivista Italiana di Dialettologia* XXVIII, 197-218. [RID5]

Rati, M. S. 2004. L'alternanza tra indicativo e congiuntivo nelle proposizioni completive: sondaggi sulla prosa italiana del Due-Trecento. *Studi di Grammatica Italiana* XXIII, 1-59. [SGI4]

Regis, Riccardo 2006. Breve fenomenologia di una locuzione avverbiale: il "solo piu" dell'italiano regionale piemontese. *Studi di Lessicografia Italiana* XXIII, 275-289. [SLI4]

Romagno, Domenica 2006. Gradiente di transitività e codifica dell'oggetto. Dall'accusativo preposizionale al partitivo. *Archivio Glottologico Italiano* 91/2, 203-222. [AGI4]

Rossi, Giovanni 2009. Il discorso scritto interattivo degli SMS. Uno studio pragmatico del "messaggiare". *Rivista Italiana di Dialettologia* XXXIII, p.143-193. [RID6]

Rossi, Lucia 2004. Italia-Europa: una ricognizione di terminologia glottodidattica. *Studi italiani di linguistica teorica e applicata* 33/3, 453-480. [SILTA4]

Ruggiero, Raffaele Savio 2004. Il dialetto tra i giovani a Torino. *Rivista Italiana di Dialettologia* XXVIII, 11-35. [RID7]

Scarpino, Cristina 2008. Il lessico scientifico nel dizionario di John Florio. *Studi di Lessicografia Italiana* XXV, 65-95. [SLI5]

Stefanelli, Rossana 2008. E' un composto nominale. Analizzare con cautela. *Archivio Glottologico Italiano* 93/1, 3-45. [AGI5]

Telve, Stefano 2004. Vicende editoriali e normative della Gramatica ragionata della lingua italiana di Francesco Soave. *Studi di Grammatica Italiana* XXIII, 61-86. [SGI5]

Vietti, Alessandro 2009, Contatto e variazione nell'italiano di immigrati. La formazione di una varietà etnica. *Studi italiani di linguistica teorica e applicat*a 38/1, 29-53. [SILTA5]

English sub-corpus

Anderson, Kate T. 2007. Constructing "otherness": ideologies and differentiating speech style. *International Journal of Applied Linguistics* 17/2, 178-197. [IJAL1]

Barnbrook, Geoff 2005. Usage notes in Johnson's dictionary. *International Journal of Lexicography* 18/2, 189-201. [IJL1]

Bender, Emily M. 2005. On the Boundaries of Linguistic Competence: Matched-Guise Experiments as Evidence of Knowledge of Grammar. *Lingua* 115, 1579–1598. [L1]

Berns, Margie 2005. Expanding on the Expanding Circle: Where do WE Go from Here? *World Englishes*, 24/1, 85-93. [WE1]

Boecxk, Cedric 2004. Long-Distance Agreement in Hindi: Some Theoretical Implications. *Studia Linguistica*, 58/1, 23-36. [SL1]

Ellis Nick C. 2006. Selective Attention and Transfer Phenomena in L2 Acquisition: Contingency, Cue Competition, Salience, Interference, Overshadowing, Blocking, and Perceptual Learning. *Applied Linguistics* 27/2, 164-194. [AL1]

Ellis Rod 2006. Modelling Learning Difficulty and Second Language Proficiency: The Differential Contributions of Implicit and Explicit Knowledge. *Applied Linguistics* 27/3, 431-463. [AL2]

Friginal, Eric 2007. Outsourced call centers and English in the Philippines. *World Englishes* 26/3, 331-345. [WE2]

Hall, Tracy A. 2006. English syllabification as the interaction of markedness constraints. *Studia Linguistica*, 60/1, 1-33. [SL2]

Hellerman, John 2006. Classroom Interactive Practices for Developing L2 Literacy: A Microethnographic Study of Two Beginning Adult Learners of English. *Applied Linguistics* 27/3, 377-404. [AL3]

Hill, Virginia 2006. Stylistic Inversion in Romanian. *Studia Linguistica* 60/2, 156-180. [SL3]

Holmberg, Anders 2007. Null subjects and polarity focus. *Studia Linguistica*, 61/3, 212-236. [SL4]

Hughes, Shaun F. D. 2004. Was there ever a "Māori English"? *World Englishes*, 23/4, 565-584. [WE3]

Hyland, Ken 2008. Academic clusters: text patterning in published and postgraduate writing. *International Journal of Applied Linguistics* 18/1, 41-62. [IJAL2]

Jary, Mark 2008. The Relevance of Complement Choice: A Corpus Study of "Believe". *Lingua* 118, 1-18. [L2]

Larsen-Freeman, Diane 2006. The Emergence of Complexity, Fluency, and Accuracy in the Oral and Written Production of Five Chinese Learners of English. *Applied Linguistics* 27/4, 590-619. [AL4]

Lindeman, Stephanie 2005. Who speaks "broken English"? US undergraduates' perceptions of non-native English. *International Journal of Applied Linguistics* 15/2, 188-212. [IJAL3]

Martin, Elizabeth 2007. "Frenglish" for Sale: Multilingual Discourses for Addressing Today's Global Consumer. *World Englishes*, 26/2, 170-188. [WE4]

McDermott, Anne 2005. Johnson's Definitions of Technical Terms and the Absence of Illustration. *International Journal of Lexicography*, 18/2, 173-187. [IJL2]

McKenzie Robert M. 2008. Social factors and non-native attitudes towards varieties of spoken English: A Japanese case study. *International Journal of Applied Linguistics* 18/1, 63-88. [IJAL4]

Mitchell, Linda C. 2005. Johnson Among The Early Modern Grammarians. *International Journal of Lexicography*, 18/2, 203-216. [IJL3]

Moon, Rosamund 2004. On Specifying Metaphor: An Idea and its Implementation. *International Journal of Lexicography*, 17/2, 195-222. [IJL4]

Morgan, Gary 2006. "Children are Just Lingual": The Development of Phonology in British Sign Language (BSL). *Lingua* 116, 1507-1523. [L3]

Osselton, Noel E. 2006. Usage guidance in early dictionaries of English. *International Journal of Lexicography*, 19/1, 99-105. [IJL5]

Padney, Anita 2004. Culture, Gender, and Identity in Cross-Cultural Personals and Matrimonials. *World Englishes*, 23/3, 403-427. [WE5]

Pereltsvaig, Asya 2007. The Universality Of Dp: A View From Russian. *Studia Linguistica* 61/1, 59-94. [SL5]

Ramanathan, Vaidehi 2008. Redux: A Derridean Exploration of Alzheimer Life Histories. *Applied Linguistics* 29/1, 1-23. [AL5]

Seargent, Philip 2005. "More English than England itself": the simulation of authenticity in foreign language practice in Japan. *International Journal of Applied Linguistics* 15/3, 326-345. [IJAL5]

Smith Jennifer L. 2004. Making Constraints Positional: Toward a Compositional Model of CON. *Lingua* 114, 1433-1464. [L4]

Velde, John R 2005. Unifying Prosody and Syntax for Right- and Left-Edge Coordinate Ellipsis. *Lingua* 115, 483-502. [L5]

Vickers, Caroline H. 2008. Expertise, Language Competencies and the L2 User. *International Journal of Applied Linguistics* 18/3, 237-255. [IJAL6]

Walter, Catherine 2007. First- to second-language reading comprehension: Not transfer, but access. *International Journal of Applied Linguistics* 17/1, 14-37. [IJAL7]