TOWARDS THE CONSTRUCTION OF A MULTILINGUAL, MULTIFUNCTIONAL CORPUS: FACTORS IN THE DESIGN AND APPLICATION OF CORDIALL

Adriana Pagano / Célia Magalhães / Fábio Alves*

ABSTRACT: This paper describes the rationale for the design of the CORDIALL corpus, developed at the Núcleo de Estudos da Tradução (NET) at the Faculdade de Letras, Federal University of Minas Gerais, Brazil. It focuses on aspects of the construction and use of CORDIALL as a resource for the study of discourse and cognitive issues in an interdisciplinary approach drawing on insights from corpora studies, translation studies, cognitive studies, discourse analysis, and cultural studies.

KEYWORDS: corpora design, translation studies, discourse analysis, cultural studies, cognitive studies.

RESUMO: Este artigo apresenta os fundamentos teóricos para elaboração do corpus CORDIALL, desenvolvido pelo Núcleo de Estudos da Tradução (NET) da Faculdade de Letras da UFMG, focalizando aspectos da sua construção, bem como sua utilização para o estudo de características discursivas e cognitivas por meio de uma abordagem interdisciplinar que congrega subsídios dos estudos de corpora, dos estudos da tradução, dos estudos da cognição, da análise do discurso e dos estudos culturais.

UNITERMOS: desenvolvimento de corpora; estudos da tradução; análise do discurso; estudos culturais; abordagens cognitivas.

* Federal University of Minas Gerais, Belo Horizonte, Brazil.

Introduction

One of the challenges for corpora studies in the decades to come is to explore new ways of retrieving more discourse-friendly data and reconcile data-driven studies with speculative theoretical approaches that have traditionally rejected quantifications, even when numbers give way to qualitative analyses on a second stage.

The interface between text and discourse analysis, and corpora studies is no doubt well established, especially through research based on M. A. K. Halliday's functional grammar and genre and register analysis (Stubbs, 1996; Biber et al, 1999; Ghadessy et al., 2001). Drawing on these theoretical perspectives, most studies rely on some kind of manual corpus annotation in order to allow for automatic retrieval of discourse relevant units. Things get infinitely more complex, however, when it comes to dealing with patterns around and beyond the clause, particularly when annotation would be almost impracticable even in small corpora. The articulation between cultural theories and corpora studies equally demands a great deal of effort not only to be acknowledged as an authorized practice in academic circles but also to be able to rely on text compilation and data likely to be interpreted from the perspective of discourse features. A further challenge to corpora studies is the development of computer tools to deal with language process data, instead of product data fed into the databank through text files. Process data involves, for instance, files stored in learner corpora produced by software that records the writing and/or translating processes.

These issues and others have been present ever since the CORDIALL corpus began to be compiled at the Núcleo de Estudos da Tradução, at Faculdade de Letras, Federal University of Minas Gerais, Brazil, particularly because of its projected design hosting both product and process data and its multifunctional applications seeking interfaces between cultural studies, discourse analysis, and corpora studies, on the one hand, and cognitive and discourse approaches to language use and translation on the other. This article aims at introducing aspects of the design of CORDIALL and presents some of the major ongoing projects that draw on its data.

The design of CORDIALL

As previously mentioned, CORDIALL – Corpus of Discourse for the Analysis of Language and Literature – is a corpus developed by researchers at NET – Núcleo de Estudos da Tradução – hosted at Faculdade de Letras, Federal University of Minas Gerais, Brazil. The corpus began to be compiled in 1999 and has already surpassed the milestone of one million words. It consists of computerized texts gathered according to specific criteria that relate to the research subprojects implemented by the researchers at NET.

At present CORDIALL combines four subcorpora, namely:

- A multilingual translation corpus – a subcorpus of original and translated texts in English, German, Portuguese, and Spanish. This includes single and multiple translations of the same original for the purposes of comparative analyses regarding translators' retextualization and the investigation of historical parameters in the production of translated texts;
- A comparable corpus of Brazilian Portuguese – a subcorpus of original texts in Brazilian Portuguese and texts translated into Brazilian Portuguese, compiled according to the criteria of genre and publication date;
- The processual corpus CORPRAT, Corpus on Process for the Analysis of Translations, a subcorpus of novice and expert translators' texts including log files with translation process data recorded online (i.e. keystroke logging through TRANSLOG©), audio files gathered by means of concurrent and retrospective verbal protocols, as well as their corresponding transcriptions, image files recorded with proxy monitoring systems, and text files with the translation product;
- A specialized corpus of academic and journalistic texts – a subcorpus of texts selected on the basis of discourse and language features compiled for the purpose of examining rhetorical patterns and lexical choices. This subcorpus does not include translated texts.

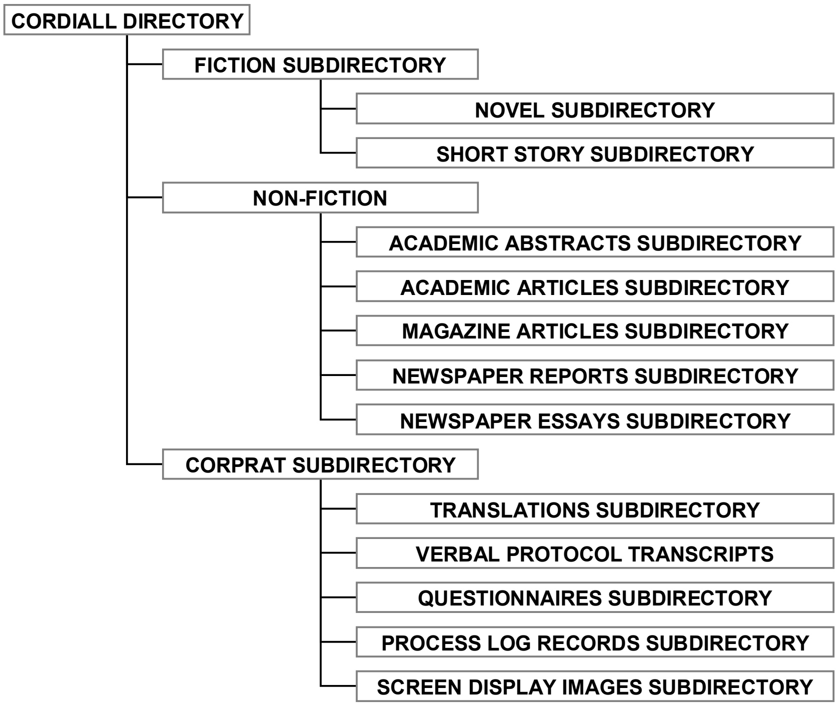
Hence, CORDIALL is a multilingual corpus covering so far Brazilian Portuguese, English (including North American, Canadian, and British varieties), Spanish (Argentine and Iberian), and German. It is also worth mentioning that CORDIALL hosts the production of multiethnic and multilingual authors, in which case they are classified in terms of the dominant language they use (e.g. English in Salman Rushdie's novels); observations are made as to the hybridity of forms in their language.

Despite the complex issues involved in obtaining permissions for text reproduction, **full texts** rather than samples are used, so that the corpus can be used for specific projects, such as the study of translators' style and idiosyncrasies, and the analysis of recurrent patterns as characteristic of particular genres. Two main categories of text are represented in CORDIALL: **fiction** and **non-fiction. Published** and **non-published** texts are used, the latter being translations produced both by **novice** and **expert translators** for the CORPRAT subcorpus.

Two genres are included in fiction: **novels** and **short stories**; and four genres are represented in non-fiction: **newspaper reports and essays, magazine articles**, **academic articles,** and **abstracts.**

Text preparation includes: scanning and/or typing; production of full bibliographic records for CORDIALL's info databank; annotation of titles, headlines, major divisions such as chapters and sections, genres within major genres (e.g. poems within novels). Wordsmith Tools is the most frequently used software for text analysis. Specific software for automatic recognition and parsing of clauses is being developed.

CORDIALL is made up of subdirectories that correspond to the genre classification in the corpus. All files are ASCII text files. The diagram below outlines the file structure of CORDIALL.


CORDIALL DIRECTORY
FICTION SUBDIRECTORY
NOVEL SUBDIRECTORY
SHORT STORY SUBDIRECTORY
NON-FICTION
ACADEMIC ABSTRACTS SUBDIRECTORY
ACADEMIC ARTICLES SUBDIRECTORY
MAGAZINE ARTICLES SUBDIRECTORY
NEWSPAPER REPORTS SUBDIRECTORY
NEWSPAPER ESSAYS SUBDIRECTORY
CORPRAT SUBDIRECTORY
TRANSLATIONS SUBDIRECTORY
VERBAL PROTOCOL TRANSCRIPTS
QUESTIONNAIRES SUBDIRECTORY
PROCESS LOG RECORDS SUBDIRECTORY
SCREEN DISPLAY IMAGES SUBDIRECTORY

The texts included in CORDIALL are all **written texts from written sources,** except for written transcriptions of concurrent and retrospective verbal protocols in the CORPRAT subcorpus. Image files are also part of the databank and correspond to both scanned covers of novels and screen display image files recording results in the CORPRAT subdirectory.

CORDIALL works along two time dimensions: a **synchronic** one covering originals and translated texts produced from the 1990s onwards, and a **diachronic** one including originals and translated texts produced in the period ranging from the 1930s to the 1950s. Texts produced in different historical periods can thus be compared. Since selection of texts is carried out on the basis of translated texts, which comply with the historical criteria defined above, some of the originals fall outside those time parameters, as is the case of Mark Twain's *The adventures of Huckleberry Finn,* published in 1883 but translated into Brazilian Portuguese and published throughout the 1940s-1990s period.

The time parameters mentioned above are actually those set by ongoing projects and there is a clear potential for further expansion. The databank is continuously fed with new files for the texts incorporated into CORDIALL by new research projects.

The texts included in CORDIALL are recorded in information files in a databank. Entries include basic identification data, historiographical data when applicable, and file attributes. Identification data comprise title, author/translator, language, status (original, translation) and genre. Historiographical data include place, date, and publishing house of first edition and place, date, and publishing house of the edition in CORDIALL, when the latter differs from the former. Details about bibliographic collections or series are considered, as well as data about paratextual features, such as covers, blurbs, prefaces, and footnotes. As regards authors and translators, biographic details, including sexual orientation and race/ethnicity, are entered. File attributes such as name, extension, number of words, and location are also recorded.

A word on terminology is important here. Investigations in many fields of knowledge conducted with the support of corpus linguistics are still in their infancy; much has to be done for a consistent ground to be achieved in terms of the adequate use of theoretical concepts. This applies to two expressions commonly used with respect to the methodology of corpus research: the "corpus-based approach" and the "corpus-driven approach". Tognini-Bonelli (2001) is illuminating in this matter. This is how she distinguishes both approaches:

> (...) corpus based is used to refer to a methodology that avails itself of the corpus mainly to expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study (...) In a corpus-driven approach the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be comprehensive with respect to corpus evidence. The corpus, therefore, is seen as more than a repository of examples to back pre-existing theories or a probabilistic extension to an already well defined system (p. 65, 84)

Hence, for Tognini-Bonelli (2001), corpus has the primacy in the corpus-driven approach, but it is taken for granted that, throughout the process, the analysis is mediated by the linguist. This entails that the linguist's theoretical knowledge, his/her attitudes and life experience will continue to guide him/her at every stage of corpus research.

Corpus-driven investigation in translation is usually recognized by the international research community as integrating a field of translation studies that emerged in the 1990s out of a combination of the theoretical foundations of the Descriptive Translation Studies (DTS) and the methodology of Corpus Linguistics. This kind of study is best represented in Baker (1993, 1995, 1996, 2000), Laviosa (1997a, 1997b, 1998), and Kenny (2001), among others. However, there is also the kind of research which draws on different theoretical traditions, like the work carried out in a contrastive linguistic perspective using COMPARA (Santos, 1998) or investigations with parallel corpora based on Halliday's functional grammar, e.g. Maia (1998). In Brazil, a cross-linguistic research about collocation and colligation patterns in English and in Portuguese is being conducted within COMET (Tagnin, 2000) with a view to publishing a corpus-based bilingual dictionary of word combinations. Corpus-driven research with CORDIALL is developing a new interdisciplinary tradition inasmuch as it attempts to interrelate theories of different fields of knowledge for the investigation of translation products, namely cultural studies, critical discourse analysis, and systemic functional grammar, using enhanced corpus-analytical procedures through the methodological resources of corpus linguistics.

One of the focuses of CORDIALL has been to support research on translation and discourse analysis, drawing on studies with small, rather than large corpora. The reason for this lies in objectives pursued by CORDIALL researchers: to interrogate the corpus with a view to elucidating discursive and cognition-related aspects of the texts. In language and translation studies alike, the beginnings of corpus-based research was marked by the intention of examining large corpora in order to focus on language specificities from naturally occurring data. However, researchers have increasingly pointed to the need to consider the results of corpus research in a perspective that transcends claims of objectivity and exhaustiveness, questioning requirements for corpus size. Within the field of translation studies, Tymoczko (1998: 654) reminds us that *"corpora are to be seen as products of human sensibility, connected with human interests and self-interests"*, which implies that insights gathered from corpus studies are relevant regardless of corpus size, provided we are aware of our research agenda and its projection onto the data to be examined. Thus, we agree with Tymoczko when she reinforces the idea that *"the value of corpora in translation and of a CTS approach to translation theory and practice does not rest on the claim to 'objectivity' and the somewhat worn philosophical claims of this type presuppose"* (Tymoczko, 1998: 654). Speaking from the standpoint of a linguist and a corpus researcher himself, John Sinclair (2001) warns us about the historical contingency involved in the labeling of corpora as "large" and "small", the difference between which should be seen in methodological rather than quantitative terms. In fact, small corpora allow for the choice of methodological paths, such as partly automatic or manual annotation, which can render discourse-relevant data to be interpreted in an interdisciplinary perspective that draws on critical discourse analysis, cultural studies, and cognitive studies, among other disciplines. This is precisely what CORDIALL seeks to do through the use of now available computational tools, as well as other potential developments that might offer the possibility of more sophisticated, discourse-friendly queries.

CORDIALL covers the production of several female and male authors and translators. As previously mentioned, selection of texts, particularly authors, titles and topics, is related to the different projects drawing on CORDIALL's data. Similarly, text annotation is oriented towards particular analyses carried out using the databank. An examination of some of the major projects under way will give us a better picture of the corpus.

Research project: hybridity in (con)texts of a parallel corpus

One of the research projects being carried out within CORDIALL aims at analyzing a small parallel corpus (English/ Portuguese; Portuguese/English) of novels and short stories collected on the basis of the textual and cultural hybridity of these genres. It is informed by cultural studies, critical discourse analysis, and a version of corpus-based translation studies, mainly based on Halliday's systemic linguistics.

The focus on a macrotextual dimension is the representation of hybrid identities and power relations through hybridization of texts and its connection to the sociocultural context of production both of source and target texts (Fairclough, 1992). On a microtextual dimension, lexical cohesive networks (Halliday & Hasan, 1976; Halliday & Hasan, 1989; Koch, 2002) and their role in constructing identities and power relations are investigated. The main objectives of this research are: the investigation of racial representations in a parallel corpus of novels and short stories and the study of relations between macro- and microtextual aspects of discourse – hybridity of cultural identities and hybridization of genres and lexical cohesion – in texts.

Works being supervised within this project comprise:

- investigation of the role of lexical cohesion, mainly reiteration in the construction of women characters in a corpus of short stories by Clarice Lispector (*Laços de Família*) and their Italian translations;
- investigation of the role of lexical cohesion in the construction of the phantasm in *Beloved*, by Toni Morrison, and its Brazilian Portuguese translation;
- study of the role of lexical cohesion in the representation of the vampire in *Interview with the vampire*, by Anne Rice and its Brazilian-Portuguese translation by Clarice Lispector;
- study of lexical cohesion in the construction of racial identity in *The color purple*, by Alice Walker and its Brazilian-Portuguese translation;

- analysis of the role of lexical creativity in the construction of the trickster figure in *Macunaíma*, by Mário de Andrade and its English translation;
- analysis of the role of lexical cohesion in the construction of transcultural identities in a parallel corpus (Canadian and Brazilian-Portuguese) of short stories;
- investigation of lexical cohesive devices in the construction of the slave character in *Adventures of Huckleberry Finn,* by Mark Twain and its three Brazilian-Portuguese translated versions by Monteiro Lobato, Alfredo Ferreira and Sergio Flaksman.

This project also comprises two additional investigations:

- study of racialization of crime, focusing on reported speech, in a monolingual single (Brazilian Portuguese) corpus of news reports (from Brazilian newspapers *Folha de São Paulo* and *O dia*);
- analysis of patterns of reported speech in a monolingual comparable corpus (translated Brazilian-Portuguese versus non-translated Brazilian-Portuguese) of short stories and news reports.

The main concern of the research carried out within this subcorpus of CORDIALL is, thus, the investigation of lexical cohesion as a discourse aspect of text structure. This is likely to be related to the hybridization of genres as another discourse aspect of text structure. Both discourse aspects are ultimately interpreted as traces in the configuration of racial identities in source and target texts. Since WordSmith Tools is designed for analyses of keywords, collocations, and clusters, analysis of cohesive networks should be carried out semi-automatically, on the basis of representative samples while the development of software for the purpose of this investigation is assessed.

Research project: a corpus-based analysis of the strategies used by literary translators during the Brazilian and Argentine publishing boom of the 1930s-1950s

A subcorpus of American and English novels and short stories translated in Brazil and Argentina during the 1930s, 1940s, and 1950s is the basis for the investigation of the so called *golden age* of translation in these two countries. This is part of a corpus-based historical approach to translation in Latin America.

Historiographic approaches to translation have traditionally relied on paratextual and contextual features of translated texts. Translators' notes, prefaces, letters, statements and other types of documentary evidence of their work make up the sources researchers frequently use to explore different paths in their historical reconstruction of translators' praxis. Translated texts themselves have also informed researchers' surveys albeit less frequently, usually constituting a corpus subject to close observation for selected occurrences to be counted and monitored by the analysts, based on their perception and skill to do that. The emergence of technological resources such as computer storage facilities and software to analyze text databanks has brought new potentialities to historiographic accounts of translation in that intuition-based appraisals give way to corpora-based insights about recurrent patterns found in the translated texts. Thus, consistently built parallel corpora made up of originals and their translations into another language and designed on the basis of historical criteria can offer interesting raw data for discourse analysis of cultural variables which play a role in translation activity. This is the rationale for the investigation of 1930s-1950s translation boom in Brazil and Argentina. The main objectives of this investigation are:

- to integrate historiographical data (publication and translation in two Latin American countries – Brazil and Argentina – during the 1930s-1950s period) and textual insights provided by corpus-based studies as a means to further contextualize the study of translated texts;

- to investigate which strategies were used by the translators and how their choices could be explained in terms of the contextual aspects of their work, such as representation of the prospective readers of the translated texts, and also in terms of the role those texts of popular fiction were envisaged to play at that time;
- to correlate the results obtained through corpus-based studies with data obtained through text analysis and historiographic documentation;
- to explore the socio-cultural and historical contexts in order to frame the perspective adopted to interpret micro- and macrodiscursive features of the translated texts.

The criterion for selection of texts for this subcorpus is related to the translators of the novels and short stories into Brazilian Portuguese and Argentine Spanish, who are chosen because they represent well known writers and translators who have left substantial paratextual evidence of their adherence to a certain poetics and ideology of translation during the 1930-1930s period in Brazil and Argentina. Hence, the subcorpus consists of texts translated by Érico Veríssimo, Monteiro Lobato, Mário Pedrosa, Lino Vallandro, and Rachel de Queiroz in Brazil, and Jorge Luis Borges, Adolfo Bioy Casares, Julio Cortázar, and Juan Rodolfo Wilcock in Argentina.

Halliday's functional grammar is used to analyze the texts, with a special focus on thematic progression both in originals and translations. In order to retrieve data concerning thematic development, the texts are annotated in terms of theme classes defined on the basis of the categories drawn by Ghadessy & Gao (2001), which are adapted to account for specificities of the Portuguese and Spanish languages. A second object of analysis is reporting verbs in direct and indirect speech as used in novels and short stories.

It must be born in mind that patterns identified through corpus-based analyses of translated texts may be equally ascribed to the translator, his/her style, idiosyncrasies, and to the influence of the original language on the target text. Besides, these factors cannot be separated from the socio-historical context in which their production takes place. As Baker (2000: 258) puts it,

> Identifying linguistic habits and stylistic patterns is not an end in itself: it is only worthwhile if it tells us something about the cultural and ideological positioning of the translator, or of translators in general, or about the cognitive processes and mechanisms that contribute to shaping our translational behaviour.

In this sense, recurrent features in the translator's praxis accessed through corpus-based analyses of translated texts can be correlated with paratextual evidence, that is, translators' statements, notes, prefaces, and letters revealing aspects of their translation projects.

Research project: a corpus-driven analysis of the translation process with a focus on patterns of inferential processing, strategic planning, problem solving, and decision making

Within the CORPRAT subcorpus studies are carried out in order to investigate the acquisition of translation competence, the role of inferential processes, problem solving and decision making in translation contexts, cognitive profiles of novice and expert translators, and to scrutinize different subprocesses encompassing the translation process, such as the size and scope of translation units, the role of memory and effort, and the cognitive interfaces between the translation, reading, and writing processes.

Building on the process-oriented approach to translation, which goes back to the works of Krings (1986), Séguinot (1989), Tirkkonen-Condit (1991), and Lörscher (1991), research within CORPRAT innovates by fostering the use of triangulation as a methodological alternative to cross-analyze translated data. CORPRAT aims at observing process-related phenomena from four complementary perspectives, i.e. log files, audio files, image files, and target text files. Differently from traditional works based on the process-oriented approach, studies within CORPRAT not only focus on process data but also use the product of target texts as a means to inform analyses trying to account for particular traits and features of the translation process. The rationale behind CORPRAT sees target texts as data-driven input which can provide further insights into the translation process and consubstantiate additional evidence to validate process-related hypotheses.

Baker (1993) and Laviosa (1998), among others, can be seen as a source for the empirically based treatment of translated texts, both from quantitative and qualitative standpoints. Their works constitute a product-oriented approach to translation which bear similar methodological characteristics to those favored by process-oriented research on translation, namely an inductive, descriptive approach to translation studies. Within CORPRAT, this combined approach, which incorporates product-driven and process-driven data, aims at empirically validating working hypotheses about the product-process interface among translated texts.

This is the main goal of CORPRAT – Corpus on Process for the Analyses of Translations. Created in 2002, CORPRAT is a subcorpus of CORDIALL (www.letras.ufmg.br/cordiall) which is also linked to the Núcleo de Tradução da Faculdade de Letras (NET-FALE) of Federal University of Minas Gerais. CORPRAT stores data about translation process in Brazilian Portuguese, English, German, and Spanish collected among novice and expert translators. Portuguese is always the target language. English (including American and British varieties), German, and Spanish (including Latin American and Iberian varieties), in turn, feed CORPRAT as source languages.

For instances related to translation competence, research within CORPRAT has a multi-component approach, along the lines developed by PACTE (2000). This includes analyses of linguistic, strategic, psycho-physiological, and affective factors involved in the process of translation. The issue of expertise also emerges as a key factor in multi-component models; it should be considered as an individual ability which matures with time end experience and not as a potentially acquirable skill.

For issues related to inferential processing and problems of pragmatic contextualization, research within CORPRAT builds on the notion of CORT – Competence Oriented Research on Translation – as proposed by Gutt (2000) and considers translation as an act of communication between and across languages. Still drawing on Gutt and on Relevance Theory (Sperber & Wilson, 1986/95), translation inferential research within CORPRAT attempts to provide a cause-effect framework for understanding complex cognitive processes in translation, noting that the cause-effect notion here is mental rather than socio-cultural.

As medium and long term goals, CORPRAT hopes to gather a substantial body of corpus-driven data on the translation product-process interface and, as a result, meet the call made by Fraser (1996), i.e. that research into the translation process should gradually move away from the case study level to more general claims made on the basis of the comparison of different case study results, a procedure which would ultimately allow for macro analysis of relevant issues regarding the translation process between and across language pairs.

Final remarks – Future in*corpora*tions

No corpus is *per se* designed, once and for all, inasmuch as no research is bound to an object of study envisioned as a permanent, fixed goal. Approaches and methods reveal new research objects, which themselves demand new approaches and methods. CORDIALL is a clear example of a corpus continuously built and restructured according to new objects of research revealed by the corpus itself through the application of different approaches and methodologies. As stated and shown in this article, CORDIALL privileges studies based on small corpora within the large corpus and thus allows for both a discursive and a cognitive approach that builds on methodologies developed for the specific subcorpora.

The challenges that lie ahead of CORDIALL project are certainly different from those pointed out for the discipline of translation studies by Baker in 1996. One of the most demanding challenges that CORDIALL and similar projects will have to face in the future is the discourse-cognition interface, particularly because theoretical attempts to integrate both perspectives are still very timid and not free of objections.

Among the few theorists who have attempted a joint approach to discourse and cognition, Edwards (1997: 17) proves particularly insightful when he suggests that one should *"elaborate a conception of discourse as an activity, which does not rely on the idea of message transmission between minds."* Researchers at CORDIALL are aware of the complex intertwining of cognitive and discursive issues at stake in translated text production. Building on Edwards (1997), the CORDIALL project proposes a view of language (processing) as activity, as discourse. Thus, CORDIALL explores discourse construction and discursive formations through both process- and product-driven data, in order to reconcile two perspectives traditionally pursued separately (see, among others, Olk, 2002; Alves & Magalhães, in this issue). This means that data collected through the use of empirical methods of elicitation is analyzed as discourse on the same basis as data gathered from textual products with a view to mapping the cognitive and discourse-oriented characteristics of language and text processing in translation.

We propose to add a new dimension to the ongoing debate as to whether traditionally different approaches can be reconciled, trusting that the methodological scope opened up by new technological resources in the case of process data and by corpus-driven studies in the case of product data may possibly lead to a reappraisal of the theoretical viability of a discourse-cognition interface.

References

ALVES, F. & MAGALHÃES, C. (this issue) Using Small Corpora to Tap and Map the Process-Product Interface in Translation.

BAKER, M. (1993) Corpus Linguistics and Translation Studies: implications and applications. In: BAKER et al. (ed.) *Text and technology: In honour of John Sinclair.* Amsterdam/Philadelphia, p. 233-50.

BAKER, M. (1995) Corpora in Translation Studies: an overview and some suggestions for future research. *Target,* v. 7, n. 2, p. 223-43.

BAKER, M. (1996) Corpus-based translation studies: The challenges that lie ahead. In: SOMERS, H. (ed) *Terminology, LSP and translation: studies in language engineering in honour of Juan C. Sager.* Amsterdam/Philadelphia: John Benjamins Publishing Company, p. 177-86.

BAKER, M. (2000) Towards a methodology for investigating the style of a literary translator, *Target,* v. 12, n. 2, p. 241-66.

BIBER, D. et al. (1999) *Longman grammar of spoken and written English.* Pearson.

EDWARDS, D. (1997) *Discourse and cognition.* London: Sage.

FAIRCLOUGH, N. (1992) *Discourse and social change.* Cambridge: Polity Press.

FRASER, J. (1996) The translator investigated: Learning from translation process analysis. *The Translator* 2/1. p. 65-79.

GHADESSY, M. et al. (ed.) *Small corpus studies and ELT: Theory and practice.* Amsterdam: John Benjamins.

GHADESSY, M., GAO, Y. (2001) Small corpora and translation: comparing thematic organization in two languages. In: GHADESSY, M. et al. (ed.) *Small corpus studies and ELT: Theory and practice.* Amsterdam: John Benjamins. p. 335-62.

GUTT, E.A. (2000) *Translation and relevance: cognition and context.* London: Blackwell.

HALLIDAY, M. A. K. (1985) *An introduction to functional grammar.* 2. ed. London/New York/Sidney/Auckland: Edward Arnold.

HALLIDAY, M. A. K. e HASAN, R. (1989) *Language, context and text: aspects of language in a social semiotic perspective.* Oxford: Oxford University Press.

HALLIDAY, M.A.K. e R. HASAN. (1976) *Cohesion in English.* London: Longman.

KENNY, D. (2001) *Lexis and creativity in translation: a corpus-based study.* Manchester, UK & Northampton MA: St. Jerome Publishing.

KOCH, I. V. (2002) *A coesão textual.* 17ª edição. São Paulo: Editora Contexto.

KRINGS, H.P. (1986) *Was in den Köpfen von Übersetzern vorgeht. Eine empirische Untersuchung der Struktur des Übersetzungsprozesses an fortgeschrittenen Lernern,* Tübingen: Gunter Narr.

LAVIOSA, S. (1997a) How comparable can 'comparable corpora' be? *Target*, v. 9, n. 2, p. 289-319.

LAVIOSA, S. (1997b) Investigating simplification in an English comparable corpus of newspaper articles. In: KINGA KLAUDY-JÁNOS KOHN (ed.) *Transferre Necesse Est*. Proceedings of the 2nd International Conference on Current Trends in Studies of Translation and Interpreting, Budapest, Hungary, 5-7 September, p. 531-40.

LAVIOSA, S. (1998) Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta – The corpus based approach*, 43: 4, p. 557-70.

LÖRSCHER, W. (1991) *Translation performance, translation process, and translation strategies. A psycholinguistic investigation.* Tübingen: Gunter Narr.

MAIA, B. (1998) Word Order and the First Person Singular in Portuguese and English. *Meta*, v. 43, n. 4, p. 589-601.

OLK, H. (2002) Critical discourse awareness in translation. *The Translator,* v. 8, n. 1, p. 101-10.

PACTE (2000) Acquring translation competence: Hypotheses and methodological problems of a research project. In: BEEBY, A., D. ESINGER & M. PRESAS (ed.) *Investigating translation*. Amsterdam: John Benjamins. p. 99-106.

SANTOS, D. (1998) Perception verbs in English and Portuguese. In: JOHANSSON, S., OKSEFJELL, S. (ed.) *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*. Amsterdam – Atlanta, GA: Rodopi, p. 319-42.

SÉGUINOT, C. (ed.). (1989) *The Translation Process*. Toronto: H. G. Publications.

SINCLAIR, J. (2001) Introduction. In: GHADESSY, M. et al. (ed.) *Small corpus studies and ELT: Theory and practice.* Amsterdam: John Benjamins.

SPERBER, D. & WILSON, D. (1986/95) *Relevance: communication and cognition.* London: Blackwell.

TAGNIN, S. (2000) Collecting data for a bilingual dictionary of verbal collocations: From scraps of paper to corpora research". In LEWANDOWSKA-TOMASZCZYK, B., MELIA, P. (ed.*) PALC '99: Practical Applications in Language Corpora. Articles from the International Conference at the University of Lodz, 15-18 April 1999*; Frankfurt am Main: Peter Lang GmbH, p. 399-407.

TIRKKONNEN-CONDIT, S. (ed.) (1991) *Empirical research in translation and intercultural studies*. Tübingen.

TOGNINI-BONELLI, E. (2001). *Corpus Linguistics at Work*. Amsterdam/ Philadelphia: John Benjamins.

TYMOCZKO, M. (1998) Computerized Corpora and the Future of Translation Studies. *Meta*, v. 43, n. 4, p. 652-9.