A multilingual learner corpus in Brazil

Stella E. O. Tagnin
University of São Paulo

1. Introduction

One of the problems with textbooks used in Brazil for teaching a foreign language is that most are written by foreign authors unacquainted with Brazilian students' difficulties. It is a known fact that a learner corpus can provide useful data to detect such specific difficulties and consequently inform the production of pedagogic material to address these problem areas (Leech 1998).

Until recently, the only learner corpus under construction in Brazil was the Br-Icle (Berber Sardinha 2001), the Brazilian Portuguese part of the ICLE project (Granger 1993, 1994), which at the time of writing contains 40,000 words of argumentative texts collected at the Catholic University of São Paulo (PUC-SP). As of early 2002 the University of São Paulo (USP) has joined the project.

The PUC-USP partnership in the Br-Icle project triggered an interest in extending the collection of texts to include all genres in which there is student production at our Department of Modern Languages, which is composed of five areas: English, German, Spanish, French and Italian.

This article will give an overview of foreign language teaching in Brazil, with a special focus on the state-of-the-art at the University of São Paulo. It will discuss the PUC-USP partnership and how this project motivated teachers from other languages to build their own learner corpora, which will be all brought together in the Multilingual Learner Corpus (MLC). This corpus is part of a larger project – COMET, a Multilingual Corpus for Teaching and Translation, which is also being built at the University of São Paulo under my coordination (Tagnin 2002a, 2002b). Next, it will discuss the design and structure of the corpus. The last part will focus on the research possibilities envisaged by the MLC.

2. Foreign language teaching in Brazil

One of the problems with textbooks used in Brazil for teaching a foreign language is that most are written by foreign authors unacquainted with Brazilian students' difficulties. The other is that they take no account of Brazilian culture or the students' interests.

2.1.1 The curriculum at the University of São Paulo

The Department of Modern Language is divided into five areas, one for each "modern" foreign language taught at the undergraduate level: English, French, German, Italian and Spanish. The curriculum is composed of eight semesters during which students follow courses addressing both the language and the literature components. With the exception of English, where it is taken for granted that students have a quite good command of the language upon entering the course, the other languages start teaching "from scratch" as these languages are not taught regularly in secondary school whereas English is part of the compulsory curriculum.

Due to the high demand for foreign language courses in general, the Department also offers extracurricular courses, mainly aimed at the academic community. These go by the name of *English on Campus*, *Español en el Campus* etc. and extend from five to ten semesters, depending on the language. They are taught by post-graduate students under the supervision of a coordinator, who is a regular teacher in the language area. In section 4.1 we will go into more detail about these courses and how they can contribute to the corpus.

3. Learner corpora in Brazil

To our knowledge, the only learner corpus under construction in Brazil is the Br-Icle. To date it is composed of 40,000 words compiled at the Catholic University of São Paulo under the coordination of Tony Berber Sardinha. In line with ICLE requirements it is restricted to argumentative texts and

should reach 200,000 words upon completion. As of early 2002 USP has joined the project, which is expected to be completed by the end of 2004.

Although there is no notice of any other "formal" learner corpus, several materials involving the teaching of FL have been assembled by various researchers. A few are in electronic format (diskettes, CD-ROMs), but most are probably not. In any case, the material is not prepared for investigation with the aid of electronic search tools.

The area of German has been working on a Contrastive German-Portuguese Grammar project for which it has collected different types of student production. Most of this material is recorded on CD-ROMs or diskettes but is only available for internal use:

- **Verbs of transportation. Vol. 1. CAPLE - Corpus of German and Portuguese as Foreign Languages** (Blühdorn et al. 1997).

This material was collected in several schools engaged in foreign language teaching. In Brazil it came from 3rd and 4th year undergraduate students at USP and intermediate and advanced students at the Goethe Institut in São Paulo; in Germany from undergraduate Germanistic and students and learners of Portuguese at the University of Erlangen-Nürnberg.

It consists of three types of production: a) sentences in which students were required to use verbs of transportation, b) translations of 17 sentences into the foreign language, and c) description of the stories presented in six different sequences of cartoons.

- **Compositions in German – Vol. 2 CAPLE − Corpus of German and Portuguese as Foreign Languages.** (Blühdorn et al. 1999)

This material was collected between 1996 and 1998 from 342 informants at three German-Brazilian secondary schools and is divided into three categories: a) Brazilian learners of German, b) Brazilian learners who have both German and Portuguese at home, and c) German native speakers living in Brazil.

- **Contrastive Analysis Corpus of Mistakes in Portuguese and German as Foreign Languages** (Glenk & Stanich, 2000).

This material was collected from 2nd to 3rd year undergraduate learners of Portuguese at the Universities of Erlangen-Nürnberg (1997) and University of Vienna (1999 and 2000), and from 3rd and 4th year undergraduate learners of German at USP (1998 and 1999). It consists of descriptive texts based on the cartoons used in the research referred to above, narrative texts and essays.

Other non-contrastive materials are:

- **Corpus of letters exchanged between learners of German in Fès (Morroco) and São Paulo (University of São Paulo)** (Blühdorn 1997).
- **Studentenzeitung** (Blühdorn 1999). A newspaper written by 3rd year German learners at USP during the 2nd semester 1999.
- **Student writing, different typologies: criteria for text production** (Nomura, in preparation). Material produced by 2nd year German learners at USP.

In English and Spanish there are scattered collections of texts as a result of individual researches by post-graduate students, mainly meant for contrastive studies. However, they are not in a format that makes them searchable by corpus tools.

4. The Learner Corpus at USP

When the PUC-USP partnership was established, several teachers at USP became interested in corpora studies. Because the Br-Icle requires only argumentative texts, the English teachers decided to build a corpus with the other types of texts produced by their undergraduate students, mainly narrative and

essays. However, once the goal of 200,000 words of argumentative texts has been reached for the Br-Icle, this type of text will also be included in the USP Multilingual Learner Corpus.

The areas of German and Spanish have already joined the MLC project. French and Italian have shown some interest but no official contact has been made as yet. Nevertheless, the project is underway. and it will also be fed with texts from the *on campus* courses.

4.1 The *on campus* courses

The participation of the *on campus* courses at USP opens up other possibilities, such as including other genres of texts and gathering the production of another type of students.

As these courses are aimed mainly at the academic community, that is, students, teachers and other employees, one gets a fairly varied audience, both in terms of age and cultural background, which is certain to affect the content of their production. Quite a few undergraduate students attend these courses to "catch up" with the rest of their class, that is, as remedial work to improve their linguistic performance. The teaching is more informal than in the regular undergraduate courses and students feel they have a lighter chip of responsibility on their shoulders when it comes to passing or failing the course.

The *English on Campus* (EOC) course has a total of 10 semesters: Basic 1, 2 and 3, Pre-Intermediate 1 and 2, Intermediate 1 and 2, Advanced 1 and 2 and a semester of Conversation. The coursebooks used are **New Interchange**: from **Introduction** up to volume 3, Part B at the Basic, Pre-Intermediate and Intermediate levels, and **Passages** Parts A & B at the advanced level. Topics for the written assignments are suggested according to the grammar points addressed. For instance, *Write a conversation with a friend in which you describe your apartment or house and ask about his or her living place* when the focus is on the Simple Present, short answers; questions with *how many* and answers with *there is, there are*.

Due to the high number of students at the *English on Campus* courses – approximately 700 – and considering about two assignments per student, it would be possible to collect around 1,000 texts per semester, that is, as long as most learners agree to sign the permission to have their assignments inserted into the corpus.

The picture is slightly different for the *German on Campus* (GOC) course. They offer a five-semester course: Basic 1, 2, 3 and 4 and a semester of Conversation. A German coursebook **Moment mal!** is used at the first three levels. Basic 4 and Conversation use material prepared by the teachers and based on other German books. The content of the Conversation course varies each semester and students may take it more than once as it is mainly aimed at giving undergraduate students an opportunity to exercise their oral production.

As there are approximately 100 students enrolled, the number of possible texts for inclusion each semester would be around 200, again considering two texts per student.

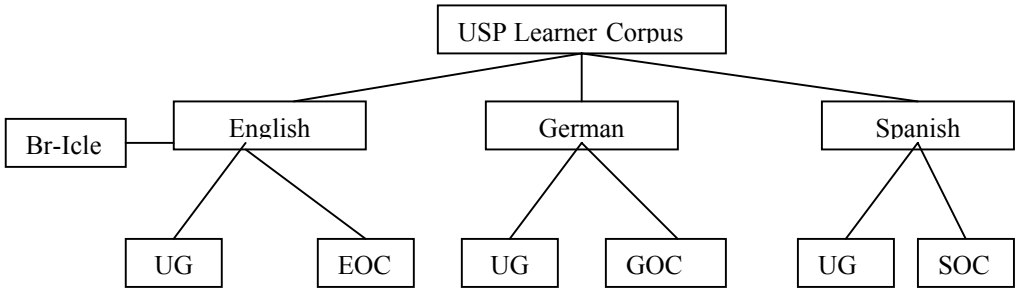
The levels at *Spanish on Campus* (SOC) are Basic 1, 2, Intermediate 1, 2 and Advanced 1, 2 Three enhancement modules are also offered, each one semester-long: a) Culture, b) Literature, and c) Conversation. A Grammar module is in preparation.

As opposed to the previous courses, Spanish relies on material prepared by their own teachers only and is based on the profile and a needs analysis of their students. There are currently about 460 students enrolled in their courses, which would give a total of 920 texts per semester.

The course has also been a source for post-graduate research on exclusion and self-exclusion factors; methodological and ideological analyses of textbooks; error analysis; evaluation of assessment procedures; and theories of language acquisition focusing especially on how or how much learning can contribute to acquisition. One such research took a contrastive approach comparing how the simple and compound past tenses are used by learners of Spanish and of English. This study was informed by data from both the *English on Campus* and *Spanish on Campus* courses.

5. The USP Multilingual Learner Corpus

To date the **USP Multilingual Learner Corpus** (MLC) will be composed of texts produced by their undergraduate (UG) and *on Campus* learners in the areas of English, German and Spanish. As mentioned above, English undergraduate texts of the argumentative type will be fed into Br-Icle until it has reached its goal of 200,000 words. See diagram below:



Each student will be identified by a code and a profile with basic information as to his course, level, year of attendance, age, sex etc. Each text will be stored in its full form and preceded by a header with information as to text type, grammar point covered, topic of assignment, coursebook (or other materials) in use, etc.

6. Possible areas of research

Learner corpora in various parts of the world have already produced a wealth of research (Granger 1998b, Granger 2002, Granger et al 2002) but to our knowledge there is no multilingual learner corpus, that is, learners with a common mother tongue learning different foreign languages. This is in contrast with the ICLE project in which one has learners with different mother tongues learning a common language.

With this design, and considering that each subcorpus is also a self-standing contrastive learner corpus, in that it allows comparison between productions originating in the two distinct courses offered at USP, it is envisaged that the corpus will not only allow for horizontal studies, comparing student production originating in the same class or in the same level, but also studies on the vertical axis, assessing student development over a period of time, either individually or collectively (cf. Kaszubski 2000, Lenko-Szymanska. 2000, among others). Research on student writing strategies like paraphrasing, the (over/under)use or avoidance of certain syntactic structures, vocabulary items, collocations and formulas (cf. Altenberg & Tapper 1998; Altenberg 2002, Berber Sardinha 2001, De Cock 1998, Granger, 1998a, 1998b and many others) will also be possible.

More interesting perhaps is the possibility of cross-linguistic studies, like detecting problems common to learning a foreign language or problems common to Brazilian learners. Another contrastive area made possible by the design of the corpus lies in the field of methodology as it will enable researchers to evaluate the effectiveness of different methodologies or materials at both the undergraduate and/or the *on campus* courses.

7. Conclusion

The Multilingual Learner Corpus under construction at the University of São Paulo is currently being fed with student production from two types of courses: the regular undergraduate courses and the extracurricular *on campus* courses offered by the areas of English, German and Spanish, at the Department of Modern Languages. The MLC is not only a promising project in terms of the array of possible research areas, but it has also integrated the different languages taught at the Department by bringing them together to work under a common project.

References

Altenberg B 2002 Advanced Swedish learners' use of causative *make*. A contrastive background study**.** In Granger et al

Altenberg B, Tapper M 1998 The use of adverbial connectors in advanced Swedish learners' written English. In Granger, S. (ed.) *Learner English on Computer*. Addison Wesley Longman, London and New York, pp 80-93.

Berber Sardinha T 2001 O Corpus de Aprendiz Br-Icle, http://lael.pucsp.br/~tony/2001bricle-interc.pdf.

Blühdorn H, Evangelista G, Reckziegel M C (orgs.) 1999. *Redações em alemão*. Vol. 2. *CAPLE - Corpus em alemão e português como línguas estrangeiras*. São Paulo, USP.

Blühdorn H. (org.) *Sagen und Legenden, Tänze, Festtagsbräuche und Kochrezepte aus Brasilien. Eine interkulturelle Korrespondenz zwischen Deutsch-Studenten in São Paulo, Brasilien, und Fès, Marrocos*

Blühdorn H (org.) 1999 *Studentenzeitung***,** São Paulo, USP.

Blühdorn H., Dias Moreira L F, Silva R F (orgs.) 1997 *Verbos de transporte*. Vol. 1. *CAPLE - Corpus em alemão e português como línguas estrangeiras*.

De Cock S 1998 A Recurrent Word Combination Approach to the Study of Formulae in the Speech of Native and Non-Native Speakers of English. *International Journal of Corpus Linguistics vol 3(1)*: 59-80.

Glenk, E, Stanich K (orgs.) 2000 *Corpus de Análise Contrastiva de Erros em Português e Alemão como Línguas Estrangeiras***.** São Paulo, USP, setembro de 2000.

Granger S 1993 The International Corpus of Learner English. In Aarts, J., de Haan, P. and Oostdijk, N. (eds) *English Language Corpora: Design, Analysis and Exploitation.* Rodopi, Amsterdam, pp 57-69.

Granger S 1994 From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In Aijmer, Karin et alli (eds.) *Languages in Contrast – Papers from a Symposium on Text-based Cross-linguistic Studies*, Lund 4-5 March 1994, Lund: Lund University Press, pp 37-51.

Granger S 1998a Prefabricated patterns in advanced EFL writing: collocations and formulae. In Cowie A. (ed.) *Phraseology: theory, analysis and applications*. Oxford University Press, Oxford, pp 145-160.

Granger S (ed.) 1998b *Learner English on Computer*. Addison Wesley Longman, London & New York.

Granger S 2002 A Bird's-eye View of Computer Learner Corpus Research. In Granger et al 2002.

Granger S, Hung J, Petch-Tyson S (eds) 2002. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Benjamins, Amsterdam and Philadelphia.

Kaszubski P 2000 Lexical profiling of English (learner) corpora: can we measure advancement levels? In Lewandowska-Tomaszczyk, B. and Melia, P. (eds) *Łódź studies in Language, Vol. 1: PALC'99: Practical Applications in Language Corpora [Papers from the International Conference at the University of Łódź, Poland, 15-18 April, 1999]*. Peter Lang, Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Wien, pp 249-86.

Leech G 1998 Learner corpora: what they are and what can be done with them. In Granger S (ed.) 1998b, pp. xiv-xx.

Lenko-Szymanska A. 2000 *How to trace the growth in learner's active vocabulary. A corpus-based study.* Paper presented at the 4th International Conference on Teaching and Language Corpora. Graz, 19-23 July 2000.

Nomura M (org.) (in preparation). *Redação de textos de diferentes tipologias: critérios de produção de texto.*

Tagnin S E O 2001 COMET – A Multilingual Corpus for Teaching and Translation. In: PALC '01 – International Conference on Practical Applications in Language Corpora, Lodz, Poland. September 07-09, 2001. To appear in the Proceedings.

Tagnin S E O 2002 Taking off in Brazil: COMET – A Multilingual Corpus for Teaching and Translation. Paper presented at ICAME 2002 – The Theory and Use of Corpora – The 23rd International Conference on English Language Research on Computerized Corpora of Modern and Medieval English, Gothenburg, Sweden, May 22- 26, 2002.