A Corpus of Sworn Translations – for linguistic and historical research

Francis Henrik Aubert
Stella E. O. Tagnin
University of São Paulo

1. Introduction

In Brazil, all and any documents and papers, when in a foreign language, need to be translated by public sworn translators if they are to be used for any official purposes. Such documents will normally range from school papers needed for a student's transfer from one country to another, birth, marriage or death certificates for naturalization, marriage or inheritance purposes, up to contracts, powers-of-attorney, promissory notes and articles of incorporation or other commercial documents for international transactions. In principle however, any text may be submitted to a "sworn" translation procedure if such text, for any reason, is to be processed by government authorities, at any level, or by the Courts. Thus, for instance, a love letter may well have to be translated by a public sworn translator if such letter is used as evidence in a divorce suit. A translation of a play may give rise to a suit for copyright infringement, and, as a consequence, even a literary text may have to be dealt with by a sworn translator, back-translating it into the initial source language, as documentary evidence in the records of the suit. (See also Aubert 1996).

Public translators must register these translations (i.e., true and unabridged copies of each translation delivered to the clients) in a special 'book of records', conforming to a certain number of rules. Upon their retirement or death, these books are turned in to the Board of Trade of the state in and for which they are qualified (the *Junta Comercial*).

The Board of Trade of the State of São Paulo (*JUCESP*) has approached the University of São Paulo, where a large corpus for teaching and translation purposes is being built – the COMET – offering this material for inclusion in the corpus. It covers a period of over one hundred years and is written in more than 20 languages. However, this offer is not without problems: a large part of the material will require a long and strenuous work of restoration, which, in turn will demand substantial fundings.

To bring the project off the ground in a relatively short period of time, it has been decided to first concentrate our efforts on the books of the last thirty years with translations into and out of Portuguese, English, German and Spanish, the languages presently addressed by the COMET.

This material will not only allow for research in linguistic and translation matters, such as, contrastive stylistics, lexicography, legal terminology, translation norms and translationese, but also in studies of a more historical nature, especially Brazil's immigration waves, periods in which there is a high demand for public translations to comply with the legal aspects of immigration. We can envisage historians and sociologists as interested researchers.

The article will discuss the relevance of this unique material, the design and structure of the corpus, its population, the audience it is aimed at, the problems expected in the preparation of the texts, the header, and possible research areas.

2. The COMET project

This project grew out of two experiments in teaching English into Portuguese translation to students at the Specialization in Translation course at the University of São Paulo (Tagnin 2002a; 2003b to

appear). The students built small bilingual corpora ranging between 100,000 and 200,000 words in each language in several technical areas[1] over a period of two years, 2000-2001. This material was put together on a CD-ROM and made available for their terminological research. The resulting glossaries are available at http://www.fflch.usp.br/citrat As other corpora were being built to inform masters and doctoral studies, it was decided to bring all this material together under a common project: COMET – A Multilingual Corpus for Teaching and Translation. The COMET consists of three subcorpora: a Technical Corpus, a Learner Corpus and a Translation Corpus. The Technical Corpus favours mainly three areas in which a significant lack of terminological sources has been identified by professional translators: Commercial Law, Computing and Orthodontics. This means that regular work is being carried on to enlarge these corpora systematically. Nevertheless, all technical corpora produced by student work or otherwise at the University of São Paulo (USP) will eventually be hosted here.[2]

As corpus work became more evident at USP, we joined the Br-Icle project, which was being conducted by Tony Berber Sardinha at the Catholic University of São Paulo. This was brought to the attention of other scholars interested in language teaching/acquisition and it was subsequently decided to construct a multilingual learner corpus at USP as the Department of Modern Languages is comprised of five different areas: English, French, German, Italian and Spanish. At the moment of writing, English, German and Spanish have teamed up to pursue this project (for more details see Tagnin 2003a).

The Translation Corpus initially consisted of student literary translations: 9 American short stories and 20 Canadian short stories. The latter have been published in book format, along with a bio-bibliography of each author (Tagnin 2002b).
To these will be added parallel (original/translation) literary texts collected in the course of several theses and dissertations in translatoloty, including short stories by Edgar Allan Poe (for which several renderings into Brazilian and European Portuguese are available), the translation into English of the classic "Os Sertões" ("Rebellion in the Backlands"), by Euclides da Cunha, to mention but a few.

Inclusion of uncorrected student production has been considered so as to form a Translation Learner Corpus to function as a source for research in the pedagogy and practice of translation.

As mentioned above, JUCESP has approached USP to donate the books of records of deceased and retired sworn translators to be included in COMET's Translation subcorpus. The sections below will provide the details of the project.

3. The JUCESP subproject

The corpus is meant for research primarily in the fields of legal translation, lexicology and terminology. It will initially be fed with JUCESP texts translated from English, German and Spanish into Portuguese or out of Portuguese into any of these foreign languages.

The audience envisaged is both translation students and teachers, lexicographers and terminologists. However, as all texts will be included in their full form they will also make valuable material for researchers interested in discourse analysis.

---

1 The areas in which these corpora were built are: Biotechnology: transgenic foods; Cooking: Spices; Computing: Security; Fashion: Clothes; Veterinary: Bovine diseases; Ecology: Biodiversity; Dentistry: Orthodontics; Automation: Safety Locks; Business: Brazilian Financial Market; Tourism: Ecotourism; Genetic Engineering: Genoma.

2 Part of this corpus will be fed into the Lácio-Web Project (Aluísio et al., 2003). In exchange, all corpus tools developed within that project are being made available to the COMET Project.

As explained below, parallel texts (originals and their translations) will be rare so that the corpus will be above all a comparable one. It will also be an open-ended corpus as new material may be inserted when new material is processed or more books of records are made available due to the death or retirement of other sworn translators.

A relevant feature is that the material is of public domain so that no copyright restrictions apply.

The sworn translations on file at the Board of Trade are organized in 'books of records', identified by the name of the sworn translator and by the foreign language from or into which the translation was carried out. Thus, e.g., a sworn translator qualified for English, French and Spanish, will have three series of volumes (of up to 400 pages each), one for each foreign language. If a given original contains textual material in more than one foreign language, the translation will normally be recorded in a volume corresponding to the prevailing language in the original text, or, alternatively, in the volume corresponding to the official language, if any, of the country in which the original document was issued.[3]

The major foreign languages represented (as source or as target languages) in the material are English, Spanish, French, Italian and German (roughly in this order). A certain amount of material is also to be found for Arabic, Dutch, Greek, Hebrew, Hungarian, Japanese, Korean, Latin, Norwegian and Russian, though for these languages the actual volume has yet to be assessed.

3.1 Conversion of the texts into electronic format

One initial major difficulty will be the conversion of the material into electronic format. Up to the early 50's, most translators copied the translations into the book of records by hand. From the mid 50's and up to the early 70's, the copies found are mostly carbon copies. Photostatic reproduction became common only in the 1980's, and, even here, the quality of the copies is not always such that it will permit electronic scanning to be performed without involving an unreasonable amount of revision. It is therefore expected that, to a very large extent, the material will have to be transcribed, an operation which – as one knows from the times of Medieval copyists – is fraught with the risks of errors, slips of the finger, lapses and the conscious or subconscious desire to "improve". The fact that, over the relevant period (1902-2002), Brazilian Portuguese has undergone two major orthographical reforms, does not render the task any easier. Nevertheless, the variety of the material and potential rewards for research are such that the effort involved (including strict supervision of the electronic transcripts) is felt to be well worth the while.

Due to the volume of material donated, it will be processed in batches:

1. texts covering the last 30 years: 1972-2002;
2. texts covering the period 1935-1971;
3. texts covering the period 1902-1934.

---

[3] The linguistically hybrid texts are indeed fairly common. A daughter company of a US holding, organized in the Grand-Duchy of Luxemburg, will have its Articles of Association drawn up in English, followed by a French version, and with its notarization also in French, unless one of the signatories is resident in Italy, in which case one of the notary public acknowledgments may be worded in Italian. In a more extreme case, a bill of lading was found to have been printed in English, its different boxes filled out in a blend of Portuguese and Spanish (in all likelihood, in a not very successful attempt to produce Portuguese); the rubber stamp of the carrier was worded in German, the address of the shipping company appeared in Norwegian, but the notarization had been conducted in the Canton of Ticino, Switzerland, and was thus formulated in Italian. Since the basic text (the starting point) was in English, this translation was inserted in the book of records for the English language of the relevant translator.

As soon as a substantial amount of the first batch has been processed, it will be made available on the Web as a pilot corpus. It will then be updated regularly as new texts are prepared and ready to be read electronically.

3.2 Text types

Although any text can, in a given situation, be subjected to a "sworn translation" procedure (see Introduction), the major part of the translations available in the material can be divided into four main groups: (a) personal documents (identity documents, birth, marriage, divorce and death certificates, school documents, and the like); (b) corporate documents (articles of association or incorporation, corporate deliberations, minutes of shareholder meetings, secretary's certifications, etc.); (c) financial documents (bills of lading, agreements in general – purchase and sale, rental, licensing of trademarks, technology transfers –, promissory notes and other securities); and (d) legal documents (petitions, letters rogatory, court decisions). Obviously, groups (b), (c) and (d) intersect, to a large extent, in terminology and phraseology, although their specific purposes and the actors involved in the respective communicative processes are somewhat different. Texts dealing with matters directly related to industrial technology are relatively rare, save for patent registrations and exhibits to technology transfer agreements, but even these contain, to a greater or lesser extent, terms and phrasings which pertain to the legal and commercial specialty languages.

3.3 The structure of the corpus

Because of the comparative and contrastive studies envisaged, structuring of the material by language will take precedence. The next level will be by date and then by text type. For instance:

- English
  - 2002
    - Personal Documents
      - Birth, Death, Marriage and Divorce Certificates
      - Transcripts
    - Financial Documents
      - Balance Sheets
      - Bills of Lading
      - Promissory Notes
    - Corporate Documents
      - Articles of Incorporation/Association
      - Minutes of Shareholder and Board Meetings
    - Legal Documents
      - Petitions, pleas and similar acts of bringing suit
      - Powers of Attorney etc
- German
  - 2002
    - Personal Documents
      - Birth, Death, Marriage and Divorce Certificates
      - Transcripts
    - Financial Documents
      - Balance Sheets
      - Bills of Lading
      - Promissory Notes
    - Corporate Documents
      - Articles of Incorporation/Association
      - Minutes of Shareholder and Board Meetings
    - Legal Documents
      - Petitions, pleas and similar acts of bringing suit
      - Powers of Attorney etc

3.4 Header

Each text will be identified by a code indicating language direction, date, text type and text number. Thus GP902BC0001 means: German into Portuguese, 1902, Birth Certificate, text nr. 00001 in that category. This will allow up to 99,999 texts in each category. The potential total will probably never be reached for most categories, but if we are to plan ahead and hope to process most of the material, we may well come close to that number for at least a few categories, especially in English, which is, by far, the most prevalent source language.

A list will be drawn up of all text types and their corresponding abbreviations, which will be part of the identifying code of each text (see next section).

A header, based on the international standards devised for the Translational English Corpus (TEC)[4], developed at the UMIST (University of Manchester Institute of Science and Technology), will provide more detailed information as to the direction of translation (into mother tongue, into foreign language), language of source text, language of target text, the date of the translation, the name of the translator, his/her sex and nationality (if available), the text type, the extent of the text, and the subject.

This procedure requires careful analysis of the material so that the header is completed correctly as the data therein will serve as pointers for selecting the texts to be submitted to research with the aid of electronic search tools, i.e., the texts will be searchable by language, by text type, by date, by translator, by subject etc.

3.5 A brief analysis of the JUCESP texts

Legally speaking, sworn translations are not "independent" texts. A sworn translation is not used officially in Brazil *in lieu of* the original; rather, it ensures that the original text[5] may be put to official use, and the original and the translation are therefore jointly submitted to the public office, agency, court or other institution (schools and academies, banks, insurance companies, traffic departments, etc.) to which they are destined. This, in principle, suggests that sworn translations will tend to adhere more closely to the original texts, that, in a sense, they will be more "literal" than common unofficial translations. This, indeed, is – supposedly – a defining feature of sworn translations in general: the intent of such translations is to assist the recipient in understanding the text within its source cultural setting, and not to propose solutions which would have been appropriate if the text had been originally produced in the target language setting. In Venuti's (1995) terms, sworn translations adopt – or ought to adopt – translation strategies or procedures which are "foreignizing" rather than domesticating. The material to be included in the corpus will, it is expected, afford the possibility of testing this hypothesis, and, more specifically, assist in identifying if and to which extent other factors beyond the "sworn translation mode" as such tend to stimulate or check the "foreignization rule" (e.g. text typology, translational direction – from/to Brazilian Portuguese –, subject matter, etc.).

But here the material available presents a specific problem. Although the sworn translators must retain a full copy of each translation they have produced in their official capacity, there is no equivalent requirement concerning copies of the original texts. Thus, the books of records which will be fed into the COMET project will not provide a strict parallel corpus, and the original texts can only be inferred from the translations

This limitation will, at times, pose a problem in that it will not allow a comparison with the original. An odd passage in the translation might, of course, indicate an error or mistake of the translator.

---

4 www2.umist.ac.uk/ctis/research/TEC/tec_home_page.htm

5 or a certified copy thereof, but never a non-certified copy or other reproduction (facsimile, electronic printout, etc.).

Alternatively, however, it could be indicative of an attempt to reproduce an oddity existing in the original, given that a sworn translation is supposed to "mirror" the original, not to improve on it (although, not infrequently, the original texts would probably have benefited from such improvements) (see Aubert, 1996, op.cit).

3.6 Research possibilities

Despite the limitations referred to in the preceding paragraphs, the material is expected to provide a wide range of information relevant to linguistic and historical research.

A cursory review of the material seems to indicate that approx. ¾ of the translations have been made into Portuguese, the remaining ¼ corresponding to translations from Portuguese and into the relevant foreign languages. Despite the marked difference in distribution, the sheer quantity of the material available (some 3,000 volumes – i.e., approx. 1.200.000 pages – covering the entire 20th century) is expected to provide sufficient textual samples for a thorough linguistic, stylistic, translational, intercultural and historical investigation in both translation directions (from/into Portuguese), and afford relevant comparisons.

For terminological purposes, the corpus of sworn translations is expected to provide a vast range of real-life translational situations, disclosing both the underlying strategies and the actual solutions provided. The fact that different source and target languages are involved will also afford a reasonable degree of comparison, and the possibility of verifying to what extent the specific source or target language exerts, as such, any influence on the terminological options made by the translators.

The terminological studies afforded by the corpus will not only derive from the actual translated texts. Some sworn translators have, in the course of their careers, set up their own personal glossaries, and one such glossary (containing close to 2,000 terms) has already been made available to the Centre for Translation and Terminology at USP, by the heirs of a recently deceased sworn translator for the German language. The textual material to be included in the COMET will serve the purpose of validating, reviewing and/or varying the solutions proposed, by comparing the glossary with the actual usages and corresponding contexts found in the said translator's recorded sworn translations. Under a subsequent stage, the terminological solutions validated for the production of this specific translator can be compared to those found in the records of other translators as well as those proposed/recommended by Chambers of Commerce, monolingual glossaries in German and Brazilian Portuguese, and so on.

Much the same expectations are relevant to phraseology, although, since phraseology is intimately related to stylistics and to the idiomatic features of each language, one might expect that the subcorpus of translations into the foreign languages will afford a safer ground for comparative analysis than translations into Brazilian Portuguese from several different foreign languages.

Consider, for instance, the more or less set phrase which usually closes the preamble to a standard Brazilian contract: "*As partes qualificadas supra têm entre si justo e acordado o que segue*" (discursively equivalent to "*Now therefore, in view of the promises and mutual undertakings and obligations contained herein, the parties agree to be bound by the following terms and conditions*"). Although the actual original texts will not be available, the predictability of this set phrase is such that it can be readily inferred in the several translations to be analysed. It can be expected to reappear in a number of different translations, into a variety of languages, with varying solutions, and will thereby provide a basis for a typical "*stylistique comparée*" investigation, much as originally conceived by Vinay and Darbelnet (1958).

In the absence of the corresponding original texts, a direct observation of the translation procedures involved cannot be conducted in any systematic fashion[6]. Yet, the observation of the lexical and syntactical structures and their frequencies in the translated material, as compared to their corresponding frequencies in authentic original texts and in common (i.e. "not-sworn") translations is expected to unveil the degree of structural "contamination" (or literal shift) of sworn translations from and into Brazilian Portuguese. The from/to distinction is here of a certain relevance. In fact, one of the initial hypotheses is that the translational strategies will not be the same, but will vary according to the translational direction.[7] Also of a certain interest is the observation of translational solutions which become standard "translationese" for translations from Brazilian Portuguese into foreign languages, as markers of Brazilian cultural, legal and institutional specificities, e.g. the widespread use of "quotaholder" as a translation for "quotista" (a reference to the shareholder of a Brazilian limited liability company, legally termed "*sociedade* ***por quotas*** *de responsabilidade limitada*") or of the official name of the country as "Federative Republic of Brazil" (derived from *"República Federativa do Brasil*", although "Federal Republic of Brazil" would probably be a more idiomatically adequate solution in English).[8]

As already mentioned, the first stage of the corpus will cover a period corresponding to the last thirty years (1972/2002). At later stages, the books of records of sworn translators from earlier times will also be included, covering a full century. Even the initial stage, however, will provide sufficient elements for diachronic investigations; probably not in terms of linguistic structure, but most certainly in terminology and, very possibly, in translation procedures and strategies. Here, a number of relevant extralinguistic factors are likely to have exerted marked influences: the redemocratization process, after a long period of military rule, culminating with the 1988 Federal Constitution (and the new institutions and legal concepts arising therefrom); the first free presidential elections in a generation, in 1989; the opening of Brazilian economy as from 1990; the intensification of the commercial and cultural exchanges with other Latin American countries, specially within the framework of the Mercosur; the growing assertion of intellectual property rights by actions in court or otherwise (which, in turn, tend to require that translations be more closely tied to the original texts, so as to avoid the risk of infringing on such property rights by the intromission of a more explicit – and uncalled for – co-authorship); and, obviously, the translators' move from the typewriter to the personal computer as a tool for writing, editing and proofreading, for terminological research, and for exploiting new strategies, including intersemiotic translation. If these factors are indeed relevant to the production of public sworn translators, one might reasonably expect to observe their reflections and refractions on the translated material, by comparing the translations produced in the mid-70's with those produced in the second half of the 90's.

The preceding considerations point to other research possibilities beyond the realm of language studies as such. The close connection between public sworn translation and the political, institutional,

---

6 Occasionally, sworn translation clients request that the original text be set up together with the actual translation, in two parallel columns. Also, a significant volume of sworn translations involve standardized original texts (e.g. passports, driving licenses), in which the wording is basically the same, save only for the personal data of the actual bearer. Here, it will suffice to have access to one such original standardized text in each language in order to conduct direct observation of the translation procedures applied by the different translators.

7 In her doctoral dissertation, Sonia T. Gehring (1998), working with a different text and translation typology (social sciences), provides statistical evidence that translations from English into Brazilian Portuguese are in a sense more "literal" (or "foreignizing"), whilst equivalent texts translated from Brazilian Portuguese into English tend to be "freer" (or more "domesticated").

8 Evidently, the same concern holds good for the opposite translational direction. It is interesting to observe that a US "county" is usually translated as "condado", although the Brazilian institutional system has a reasonably close correspondent in "comarca". And, for reasons unknown and which would bear further investigation, the Brazilian consulates abroad, when in the US or Canada, tend to identify "notary publics" as "notários" but, elsewhere (including the UK), as "tabeliães".

economic and legal spheres suggests that material covering a full century will also bear marks of the historical processes that the target community of these translations has undergone: the initial republican regime, after the abolition of monarchy in 1889 and the subsequent dictatorships, alternating with civilian rule; the two world wars and Brazil's participation therein; the several waves of immigration, specially from Spain, Portugal, Italy, Lebanon, Japan and, more recently, from Korea; the budding industrialization of the country and the shift from a rural to an urban society, accelerated as from the mid-1950's; the ups and downs of the economy. At this point, the project branches out to a promising inter- and transdisciplinary co-operation with historians, sociologists and anthropologists.

4. Conclusion

This paper has reported on the creation of a Multilingual Corpus of Sworn Translations (MCST) at the University of São Paulo, Brazil.

The MCST will initially consist of sworn translations into Portuguese and out of English, German and Spanish, as well as translations out of Portuguese into these foreign languages, covering a 30-year period (1972-2002), extracted from the complete works of deceased and retired sworn translators in the state of São Paulo over a period of one hundred years (1902-2002).

Due to the unique character of this material it is believed that even the "small" part currently under consideration will offer enough material for relevant studies in the areas of translational, lexical, syntactic, terminological, discursive and even historical and sociological research.

References

Aluísio S M, Pinheiro G M, Finger M, Nunes M G V, Tagnin S E O 2003 The Lácio-Web Project: overview and issues in Brazilian Portuguese corpora creation. In *Proceedings of Corpus Linguistics 2003*, Lancaster, UK.

Aubert F H 1996 Translation Typology: the Case of 'Sworn Translations'. In Coulthard M, De Baubeta P A O (org) *Theoretical Issues and Practical Cases in Portuguese-English Translations,* Edwin Mellen Press.

Gehring S T 1998 *As modalidades de tradução inglês/português: correlações bidirecionais*. Unpublished Doctoral Dissertation. University of São Paulo.

Tagnin S E O (ed.) 2002a *Lá do Canadá,* São Paulo: Olavobrás.

Tagnin S E O 2002b Corpora and the Innocent Translator: How can they help him. In Thelen M (ed.) *Translation and Meaning, Part 6***,** Proceedings of the Lodz Session of the 3rd Maastricht-Lodz Duo Colloquium on "Translation on Meaning", Lodz, Poland, September 22-24, 2000, Maastricht: Universitaire Pers Maastricht, 489-496.

Tagnin S E O 2003a A multilingual learner corpus in Brazil. In the *Proceedings of the Workshop on Learner Corpora* at Corpus Linguistics 2003, Lancaster.

Tagnin S E O 2003b Os *Corpora*: instrumentos de auto-ajuda para o tradutor. To appear in the special issue *Translation and* Corpora, Cadernos IX - 2002/1, University of Santa Catarina.

Venuti L. 1995 *The translator's invisibility.* London, Routledge.

Vinay J, Darbelnet J P 1958 *Stylistique comparée du français et de l'anglais*. Paris, Didier.