Comparing Non-native and Translated Language:

Monolingual Comparable Corpora with a Twist

Federico Gaspari & Silvia Bernardini

University of Bologna, Italy

1. Introduction and paper overview

This paper presents CONTE (the COrpus of Non-native and Translated English), a new type of resource which is being used in an ongoing research project looking into the salient features of two interfacing forms of mediated discourse, namely non-native and translated written language. This work focuses on the language pair Italian-English within the framework of translation universals, and adopts a novel approach hinging on a monolingual comparable perspective which diverges from research paradigms traditionally used in corpus-based translation studies.

After briefly reviewing the theoretical and methodological background to the project as a whole (Section 2), the paper explains the advantages offered by the novel research design adopted, describes the set-up of our first corpus, i.e. CONTE, the English monolingual comparable corpus (MCC) of non-native and translated texts, and provides an overview of the steps involved in designing and creating it (Section 3). With the aim of illustrating the kinds of insights that our corpus can provide, both at the descriptive and methodological levels, a case study is presented focusing on the adverbial "therefore" (Section 4). We conclude (Section 5) by discussing the potential of CONTE as a research resource and outlining some of the applications and future developments planned.

2. The project: monolingual comparable corpora with a twist

The creation of the corpus that we describe in this paper was motivated by an effort to approach the debate on translation universals from a new perspective. Research in this field was pioneered in the 1990s by Baker (1993) and Laviosa (1998; 2000; 2003), who produced groundbreaking results extending intuitions which had been gradually taking shape since the previous decade (Blum-Kulka, 1986). Within this body of work it was hypothesized, and to some extent shown, that translations (mainly into English) tend to display features such as "explicitation", "normalization", "levelling out", "simplification", etc. on which a general consensus seems to have emerged within the translation studies community (see e.g. the overview in Laviosa, 2002: 43-78).

However, some dissenting voices have been raised questioning the assumptions underlying this research agenda both in terms of objectives and methodology: criticism along these lines comes for example from Bernardini & Zanettin (2004), who focus on the limitations of the corpus-based perspective, and Salsnik (2007) who, following Toury (2004) and Chesterman (2004), takes issue with the misleading application of the notion of "universal" to translational practice, which necessarily leads to generalizations suffering from weak empirical support (cf. also Malmkjær, 2005).

Inspiration for our research project also came from work in the areas of Second Language Acquisition and English as a Second or Other Language: studies such as those presented in Færch & Kasper (1983), Blum-Kulka & Levenston (1983) and House & Blum-Kulka (1986) examine the properties displayed by non-native language production giving prominence to the notion of "interlanguage" (Selinker, 1972), which gained currency in particular throughout the

1970s and 1980s.

Since the 1990s, researchers working within the corpus-based paradigm have revived the interest in interlanguage-related issues, focusing especially on English. In particular, several studies have looked at the patterns of L2 written production by two different categories of language users with a range of mother tongues: learners at different levels of proficiency on the one hand, and people using the L2 for professional purposes on the other. The former investigations are conducted on the basis of learner corpora (e.g. Altenberg & Granger, 2001; Nesselhauf, 2004), while the latter make use of English as a Lingua Franca (ELF) corpora (e.g. Seidlhofer, 2001; Mauranen, 2003).

While studies that attempt to bring together these two research areas are few and far between, the hypothesis has been put forward that translation and non-native production may indeed share some common features. As early as the mid-1980s, Blum-Kulka (1986) suggested in a seminal article that explicitation strategies may be used both when translating and when writing a text in a foreign language; more recently Cardinaletti (2005: 60) has compared features of translated language with those typical of "language attrition", hypothesizing that the source text affects the translator's target language use like the L1 affects L2 production.

Within the corpus-based paradigm, research into translation universals (inspired by Baker, 1993) has focused exclusively on translated texts vs. native speaker usage in the attempt to uncover phenomena that can be interpreted as translation universals. The aim of our project is to extend this paradigm, attempting a systematic search for common traits shared by translation and interlanguage. If such similarities were to be found, we would be in a position to extend Baker's hypotheses about translation universals to language contact settings in general, and claim that such features are better explained in terms of *mediation* (rather than translation)

universals. On the other hand, if no similarities were apparent, Baker's claim that translation has its own unique properties would be reinforced.

Our way of testing Baker's hypotheses involves A) comparing translations with direct (non-mediated) L2 written production to see if similar features manifest themselves in non-native writing and translations for a given language; and B) considering both directions of a specific language combination (more on the corpus structure in Section 3). The methodology and research design employed by Baker and her followers is thus extended by adding the extra dimension of L2 writing to the experimental setup, and refined to focus only on a specific language pair, thus avoiding the bias introduced by direction-specific effects within the language pair in question as well as potential effects due to different source languages.

This research agenda requires appropriate corpus resources, both for English and Italian. In the remainder of the paper we focus in particular on CONTE, a MCC which consists of non-native and translated English texts. This is the first component of the pool of English/Italian corpus resources that we plan to set up and use in the context of our research into potential mediation universals.

3. The CONTE corpus

3.1 Design and construction

The CONTE corpus is part of a larger and more composite set of MCC resources, including non-native (NN) and translated (TR) texts in English and Italian, alongside (native) benchmark (BM) or reference corpora for the two languages (see Table 1; notice that the Italian component of the corpus is currently still under construction and will not be discussed in this paper).

	CONTE (English)		CONTI (Italian)	
	macro-typology	**domain**	**macro-typology**	**domain**
TR	translations from Italian into English	financial statements and reports	translations from English into Italian	TBA
NN	direct written production in English by Italian native speakers	working papers in economics	direct written production in Italian by English native speakers	TBA
BM	written production by native speakers of English	commerce & finance, economics (BNC)	written production by native speakers of Italian	TBA

***Table 1*: Overall corpus structure (Italian MCC component currently under development)**

A number of design considerations guided the planning stages and the initial steps taken in the actual compilation of CONTE. When we started our work, we did not have in mind a specific text type or domain to focus on, as it appeared fairly difficult to identify a priori domain-matched texts in English which were translations from Italian and, on the other hand, others that had been written by native speakers of Italian directly in English.[1] As a result, we carried out extensive

[1] We recognize that the strict differentiation between the notions of "native" and "non-native" speaker is an idealization, and we accept that that the validity of these notions as rigorous descriptive categories is rightly questioned in linguistics and translation studies. However, due to space constraints and given the variables that we aim to isolate, in this paper we have to rely on a broadly accepted intuitive understanding of these notions.

explorations of the World Wide Web looking for data that could be used in our research, and eventually decided to focus on texts related to economics, finance and business, which seemed to be easily available in substantial quantities for the non-native and translated English sub-corpora.

However, our extensive searches on the Web showed that no single text typology/genre exists which provides substantial amounts of freely available texts for both the non-native and the translated sub-corpora. Comparability was therefore established at the macro-topic level (economics/finance), and it was decided that the impact of genre/text type similarities and differences (if any) would be evaluated empirically through experience with the corpus. With regard to document availability, texts were sought that were meant to be widely circulated without being subject to particular restrictions in terms of copying, storing and further processing for research purposes, so as to limit copyright problems. Lastly, we were also keen to avoid the complications and the time investment entailed by the need to scan in paper documents, and therefore limited our data search only to online texts that were already available in digital format.

3.2 The non-native (NNENG) sub-corpus within CONTE

The non-native (NNENG) sub-corpus within CONTE features texts downloaded from the RePEc (Research Papers in Economics) online database,[2] a collaborative initiative which provides working papers, journal articles and software tools, currently listing over 500,000 documents and counting over 17,000 registered contributors, which claims to be the largest of its kind available on the Web. This resource seemed particularly attractive for our purposes because all RePEc materials are freely available, and are accompanied by links to the authors' institutional and personal home pages, alongside a list of contact details and information on their affiliation which

[2] Available at `http://ideas.repec.org/i/eall.html` [last accessed 21 July 2008].

proved helpful in establishing their language background. Only working papers made available through the database were included in the corpus, while articles published in academic and scientific journals were disregarded because of likely copyright-related problems. Working papers also seemed promising since they represent repositories of professional and academic writing which the authors themselves voluntarily circulate via semi-formal channels to disseminate their research findings and to encourage feedback and comments from other members of the academic community. As such, working papers would seem to be typically less subject to polishing/editing than published articles, and thus more likely to give us a direct insight into the linguistic habits and strategies of economists and financial experts who are Italian native speakers writing in English in a professional and academic setting, while probably being less affected by the confounding effects of linguistic editing and revision by native speakers.

Authors in the database writing in English but whose names "sounded" Italian were identified and information relevant for establishing their language background, and in particular their status as native speakers of Italian, was sought on the Web. Very often consulting the authors' online CVs and résumés gave us the information we needed, if their working languages or those with which they were familiar were stated, for example explicitly listing Italian as their mother tongue (possibly along with a good, excellent, etc. knowledge of English and other languages). When the explicit indication of language background was missing, we considered e.g. citizenship, place of birth, institutions where they had completed school and university education, membership of professional bodies and organizations based in Italy, articles and monographs authored in Italian, etc. If a combination of these factors indicated beyond reasonable doubt that the candidate author concerned was a native speaker of Italian, his or her working papers were included in the corpus. Doubtful cases were discarded, as there was no

shortage of suitable candidates.

When selecting the working papers for the corpus we tried to include a good spread of authors (e.g. ensuring diversity in terms of demographic and personal variables such as age, affiliation, level of academic seniority, etc.), although we did not see any problems with including more than one paper by particularly prolific writers. Regrettably, female authors are severely under-represented in the corpus (42 female vs. 153 male authors, or 21.5% of the total number of authors), as a consequence of the relatively low proportion of women writings contained in the database in the first place. All papers in the NNENG sub-corpus have single authors. Descriptive and size information is provided in Table 2.

Number of tokens	3,374,048
Number of types	149,830
Number of texts	410
Average n. of tokens per text	8,229
Number of authors	195
Authors	male: 153 female: 42
Authors with more than one text	125
Authors with ≥ 10 texts	3
Publication time span	1991-2008

***Table 2*: Details of the NNENG sub-corpus**

3.3 The translational (TRENG) sub-corpus within CONTE

Despite our efforts, we were not able to identify on the Web large enough numbers of working papers in economics which had been translated from Italian into English. For the TRENG component we therefore relied on financial statements and reports of well-known companies quoted on the stock exchange that had been translated from Italian into English and posted on the respective websites for consultation by shareholders, investors and other interested parties. These translations had one common feature: they all carried explicit notices warning their readers of their status as translations, and of the supremacy of the Italian source text for legal purposes.

Unfortunately, we were not able to document who were the translators involved in producing these English translations (it is quite possible that teams of more than one translator were employed for each translation, given the length of each document – see Table 3), as their identities are not disclosed, although in a few cases the name and contact details of the agency which took care of the translation project are provided. Ideally, we would have wanted to make sure that the translations into English had been done by native speakers of the target language, but regrettably no information is available on the texts which can help us to establish the identities or the language profiles of the translators. In the attempt to ensure as much variety as possible in our TRENG data, we selected no more than one financial statement or report per company, and as a result it is likely that these documents were translated by different (teams of) translators.

Table 3 shows some details of the TRENG sub-corpus. As can be seen this component contains fewer texts (all substantially longer) than the NNENG counterpart, and a lower number of words.

Size (number of tokens)	2,205,361
Number of types	86,027
Number of texts	39
Average n. of tokens per text	56,547
Publication time span	2000-2007

***Table 3*: Details of the TRENG sub-corpus**

3.4 Sampling strategy and data integrity

As far as the sampling strategy is concerned, for both the NNENG and the TRENG sub-corpus we opted for whole texts, following Sinclair's (1991: 19) suggestion that whole-text corpora are 'open to a wider range of linguistic studies than a collection of short samples', and widespread practice within corpus-based translation studies (e.g. Kenny, 2001; Laviosa, 1998). Omissions or deletions were avoided, as they would have been time-consuming, and might have introduced biases and potential inconsistencies. In particular, we considered and then dismissed the idea of expunging information in Italian, in-text verbatim quotations (in English) and material in the references/bibliography sections. While these might potentially represent possible sources of interference confounding the variables that we intended to investigate, the integrity of the texts under investigation was considered to be a priority; the filtering out of regularities found in parts of the texts not written by their main authors was therefore left for the corpus analysis rather than construction stage.

3.5 Benchmarking: the reference corpus

In order to support our investigations based on CONTE, we needed a reference corpus of native non-translated English for benchmarking purposes. For practical reasons, we decided to use a

sub-corpus of the British National Corpus (World Edition) (Burnard, 2007), selected so as to match as closely as possible the contents of the NNENG and TRENG sub-corpora. The 90-million-word written part of the BNC was designed according to two main concurrent criteria applied to the relevant texts, namely "domain" and "medium" (Aston, 2001: 73). Domain roughly corresponds to subject matter (e.g. imaginative, arts, belief and thought, commerce and finance, and so forth). "Medium", on the other hand, covers five classes, i.e. book, periodical, miscellaneous published, miscellaneous unpublished, to-be-spoken. Recognizing that corpus users may need finer descriptive categories, Lee (2001) provided an alternative arrangement using a more delicate categorization scheme identifying 46 "genres" for the written data. Following Lee's categories as presented in his "BNC Index",[3] we extracted the 112 texts that he grouped under the "W_commerce (commerce & finance, economics)" genre category as our reference corpus of native (British) English (BMENG, see Table 4 for more details).

[3] `http://clix.to/davidlee00` [last accessed 12 August 2008].

Number of tokens	3,759,366
Number of types	60,651
Number of texts	112
Authors	male: 54 female: 3 mixed: 6 unknown or n.a.: 49
Average tokens per text	33,565
Publication time span	1985-1994

***Table 4*: Details of the BMENG corpus (BNC commerce)**

3.6 Corpus preparation[4]

After conversion of the original pdf files to plain text format and simple cleaning procedures through batch substitutions of non-alphabetic characters,[5] the texts were POS tagged and lemmatized using the TreeTagger (Schmid, 1994). Minimal metadata were then added to the texts (as attribute-value pairs in the "text" element preceding each text in the corpus), if these were judged to be potentially useful for on-the-fly sub-corpus selection. Thus, a typical text element in the TRENG sub-corpus contains the following information:

- id="TRENG612"

[4] This section only deals with the preparation of the non-native and translational components of the English MCC, since the benchmark corpus was already available in a format adequate for the project.

[5] The data cleaning process is documented in the background information that comes with CONTE, so that its users are made aware of the (slight) interventions on the raw data.

- year="2005"

This allows users to select for searching individual texts or texts published in/before/after a given year. The NNENG sub-corpus, on the other hand, contains slightly more information, i.e.:

- id="NNENG018"
- year="2006"
- author="Antonio_Abatemarco"
- gender="Male"

Thus, the non-native corpus also allows one to select or exclude from a search texts written by a given author or by males/females. Further data about the texts (e.g., source information) are available from a separate database. The corpus was then indexed with the CorpusWorkBench (Christ, 1994), and made searchable with the associated Corpus Query Processor. At the time of writing, the corpus is available for searching via a remote Unix command line shell, though we hope to be able to make it available to the general public through a Web interface in the future.

4. Preliminary investigation: the adverbial "therefore"

Besides starting to shed new light on the hypothesized common ground between translated and non-native language (our long-term objective), initial investigations conducted on mediated English with CONTE have a methodological purpose. Given that corpus comparability is a tricky notion (Kilgarriff, 2001), especially when it gets to translated language (Bernardini & Zanettin, 2004), we believe it safer not to assume comparability "by design", i.e. based on external criteria, between our two corpora and the reference corpus we are currently employing (BNC commerce, Section 3.5). Instead, we hope that, by accumulating results and constantly evaluating them, we can develop a better idea of the ways in which the corpora we are comparing resemble or differ

from each other, and ultimately assemble data whose internal consistency or lack thereof may also tell us whether the comparability assumption is justified or not. At this stage, we take a largely serendipitous approach and look for broad trends rather than definitive evidence based on firm statistical grounds, which we leave for future more in-depth investigations.

As a first case study, we focused on the resultive adverbial "therefore", one of several linking adverbials often used in written, especially academic, discourse 'to signpost the logical and argumentative links between one part of the discourse and another' (Biber et al., 1999: 1046). We hypothesized this adverb to be potentially overrepresented in our mediated corpora with respect to native English, as a consequence of either explicitation (Blum-Kulka, 1986) or risk-avoidance (Pym, 2008). A simple search for the lemma "therefore" in the three corpora shows that the hypothesis is not supported: the normalized frequency of the adverbial in the reference corpus is intermediate between the value found in the translated corpus and the one found in the non-native corpus, as reported in Table 5.

	BNC commerce		**TRENG**		**NNENG**	
	n.	n./M words	n.	n./M words	n.	n./M words
therefore	2,397	637.6	562	254.8	3,430	1,016.5
corpus size		3,759,366		2,205,361		3,374,048

***Table 5*: Frequency of "therefore" in CONTE**

This finding is somewhat unsurprising, if one considers that the BNC sub-corpus employed for benchmarking purposes contains several different text types, including but not limited to academic prose, and that this adverbial is particularly frequent in academic language (Biber et al., 1999: 887). Thus, the much higher frequency observed in the non-native (academic) sub-

corpus and the lower frequency in the translated corpus of financial statements/reports could be due to a text-typological difference unrelated to the mediation dimension, and the data at our disposal do not allow us to rule out this possibility.

However, we can search for frequency data about patterns around "therefore", and see whether the proportion of certain patterns to the total differs in the three corpora, and in particular whether translated and non-native English are more like each other than like original English. Here we focus on two patterns, namely "[punctuation mark] + therefore" and "[verb] + therefore". We can expect a search for the first pattern to return sentence/clause initial "therefore" and (mainly) medial (post-subject or post verbal) "therefore", e.g. (Concordance 1):

wiped out during the year <. Therefore> the actual saving rate o	[NNENG]
egate saving data are not <, therefore> , sufficient evidence to	[NNENG]
affected by the mutation <; therefore> by comparing the differe	[NNENG]
d to realize new projects <: therefore> the problem is how the I	[NNENG]
pected inflation of 1.9 % <. Therefore> the real rate i equalled	[TRENG]
wo quarters . Performance <, therefore> , mirrored government fo	[TRENG]
e benefits were suspended <; therefore> , no benefit has been ta	[TRENG]
arative figures presented <: therefore> , the comparative balanc	[TRENG]
eir clearing house system <. Therefore> all banks dealing in eur	[BNC Commerce]
ersations . The telephone <, therefore> , saves time and gives y	[BNC Commerce]
ess or to create goodwill <; therefore> every letter should conv	[BNC Commerce]
the demand at that price <: therefore> even those of them who w	[BNC Commerce]

***Concordance 1*: Examples of "therefore" preceded by punctuation**

Quantitative data about the frequency of "therefore" in initial position (i.e. after a full stop, a colon or semi-colon) is shown in Table 6, while Table 7 gives results for the medial (i.e. post-comma) position. For each corpus, the third column of each table gives normalized frequency

data per million words as a percentage of the total number of occurrences of the lemma "therefore".

	BNC commerce			TRENG			NNENG		
	n.	n./M words	%	n.	n./M words	%	n.	n./M words	%
therefore	2,397	637.6	100%	562	254.8	100%	3,430	1,016.5	100%
. therefore	314	83.5	**13.0**	98	44.4	**17.4**	1,141	338.1	**33.2**
; therefore	23	6.1	**0.9**	7	3.1	**1.2**	95	28.1	**2.7**
: therefore	7	1.8	**0.2**	2	0.9	**0.3**	29	8.5	**0.8**
corpus size	3,759,366			2,205,361			3,374,048		

***Table 6*: "therefore" in initial position**

	BNC commerce			TRENG			NNENG		
	n.	n./M words	%	n.	n./M words	%	n.	n./M words	%
therefore	2,397	637.6	100%	562	254.8	100%	3,430	1,016.5	100%
, therefore	506	134.5	**21.0**	112	50.7	**19.8**	464	137.5	**13.5**
corpus size	3,759,366			2,205,361			3,374,048		

***Table 7:* "therefore" in medial position**

As can be seen, the data for the two mediated corpora show similar trends, i.e.: "therefore" used in initial position (Table 6) tends to be proportionally more frequent in mediated than non-mediated language, while medial "therefore" (Table 7) is proportionally slightly more frequent in non-mediated language.[6] Since the former is considered to be the unmarked position for linking

[6] Notice that the percentages of initial "therefore" differ substantially between translated and non-native English (17.4 vs. 33.2% for occurrences following a full stop), possibly as a result of the text typological differences discussed above.

adverbials (Biber et al., 1999: 891),[7] this observation might be explainable with reference to the normalization or (in Toury's words) "growing standardization" hypothesis, according to which '[in translation], textual relations obtaining in the original are often modified […] in favour of [more] habitual options offered by a target repertoire' (Toury, 1995: 268). In other words, this might be an instance of normalization applying to both translated and non-native texts.

Focusing on the second pattern (a verb followed by "therefore"), translated and non-native texts are characterized by lower percentages if compared to the reference corpus (again, out of the total number of occurrences of the adverbial per million words, see Table 8). While there are 225 occurrences per million words of this pattern in the reference corpus, corresponding to 35.2% of the total occurrences of "therefore", the percentage is lower in the translated sub-corpus (30.2%, or 77 occurrences per million words) and lower still for the non-native sub-corpus (17.2% or 175.4 occurrences per million words; remember that the adverb is extremely frequent in the non-native corpus).

	BNC commerce			**TRENG**			**NNENG**		
	n.	n./M words	%	n.	n./M words	%	n.	n./M words	%
therefore	2,397	637.6	100	562	254.8	100%	3,430	1,016.5	100%
[verb] therefore	846	225.0	**35.2**	170	77.0	**30.2**	592	175.4	**17.2**
corpus size		3,759,366			2,205,361			3,374,048	

***Table 8*: "verb therefore" as a percentage of total "therefore"**

Furthermore, if we observe the (normalized) frequency of "therefore" immediately following a

[7] Incidentally, in the case of "therefore" in the BNC commerce sub-corpus at least, the 'unmarked' position is not in fact the most frequent.

verb as a percentage of the corresponding normalized frequency of verbs in the corpus – rather than in comparison with the total occurrences of the adverbial, as given in Table 8 – we find that both mediated corpora have consistently lower values (see Table 9). In other words, the (slightly) higher frequency of post-verbal "therefore" in the reference corpus is not an effect of differences in verb frequency with respect to the two mediated corpora.[8]

	BNC commerce			**TRENG**			**NNENG**		
	n.	n./M words	%	n.	n./M words	%	n.	n./M words	%
total verbs	649,485	172,764.5	100%	226,067	102,507.9	100%	581,438	172,326.5	100%
[verb] therefore	846	225.0	**0.13**	170	77.0	**0.07**	592	175.4	**0.10**
corpus size			3,759,366			2,205,361			3,374,048

***Table 9*: "verb therefore" as a percentage of total verbs**

5. Conclusion and further work

In this paper we have presented a corpus-based research project whose long-term objective is the exploration of the hypothesis that non-native and translated language share similar features, and that these can be accounted for by the notion of mediation (rather than translation) universals. As a first step in this direction, the paper has described CONTE, the monolingual English component of the corpus we are building, and presented a small-scale case study to illustrate the kinds of analyses for which it can be used, focusing on the behaviour of the resultive adverbial "therefore" in non-native and translated texts. For benchmarking purposes, we used a comparable corpus of native English derived from the BNC.

---

[8] Incidentally, this case study on "therefore" has also pointed at a general tendency displayed by both mediated

While this approach can give promising results, and provide data to ascertain the validity of the corpus empirically, through accumulation and evaluation of results (Atkins et al., 1992; Hunston, 2002: 28-30), we would like to compile reference corpora to use as benchmarks that are more closely comparable by design to the corpora under investigation. In the near future we will be adding two ad hoc reference corpora to our corpus, one consisting of working papers in economics written by native speakers of English (to be used against our NNENG sub-corpus), and the other of original financial statements and reports written directly in English by native speakers (as a benchmark for the TRENG sub-corpus). We believe that this would be a worthwhile investment of time and resources, since it should be relatively easy to build these two additional tailor-made reference corpora, which would be more fine-tuned to our mediated data collections. In this way we will have a split reference corpus consisting of two parts, each of which would be closely comparable to one of the two experimental sub-corpora by design. We expect these more closely comparable corpus resources to make patterns stand out more clearly and to make searches less labour-intensive.

After analyzing a wider range of phenomena for mediated English with CONTE, of which the case study presented in this paper represents one example in terms of methodology and approach to the investigation, we intend to move on to explore if any comparable tendencies are observed in the opposite language direction. The next step in our work is therefore going to be an investigation of similar phenomena with implications for the concept of translation (and mediation) universals for mediated (non-native and translated) Italian. We intend to replicate the corpus architecture described in Section 3, using a MCC of Italian made up of the following three components: (i) translations from English, (ii) non-native texts written by authors with

---

corpora to make lighter use of modal verbs than the reference corpus, an intriguing finding deserving further study.

English as their mother tongue, accompanied by (iii) a suitable reference corpus for benchmarking purposes. For our preliminary investigations we plan to use a sub-corpus of the "La Repubblica" corpus (Baroni et al., 2004), subsequently developing, if necessary, fine-tuned reference corpora for non-native and translated production respectively.

We are currently surveying which texts are available in Italian that could offer a good level of comparability. The phenomena that one might investigate for Italian in an attempt to shed light on the "mediation universals" hypothesis include the treatment of subject pronouns (differently from English, Italian is a pro-drop language; research in this area has been recently conducted from the points of view of developmental linguistics and Second Language Acquisition, see e.g. Serratrice, 2005; 2007), the distribution of past-tense verbs (in particular the imperfect vs. the present perfect in the indicative mood), the use of definite articles and the pre- vs. post-noun positioning of attributive adjectives.

One extension that we are considering for our research would be to carry out broader analyses encompassing a parallel corpus component (clearly, this would be relevant only for the translated, not the non-native sub-component, for which no parallel texts exist). Although so far we have not taken this dimension into account, favouring a monolingual comparable approach, in the process of data collection for the translated component of CONTE we have also paid attention to source texts in Italian: whenever we were able to locate them, we kept a copy for future reference. Since the TRENG texts are translations into English of financial statements and reports of high-profile companies, parallel source texts in Italian were found and collected, though not processed at this stage. As a result, the option of using parallel data in more detailed investigations in the future is still open, particularly with a view to evaluating the impact of source-text effects on the translations.

Furthermore, for our research project we are currently restricting our focus to the English-Italian language pair in both directions, which should serve as a pilot investigation to establish the potential of our methodology. Looking at other language combinations in the future would clearly be essential to build a more accurate and comprehensive picture of mediation-related phenomena. This would involve creating corpora for other languages with structures similar to the one described in Section 3, in order to check whether findings are consistent across different language pairs and if general patterns emerge.

As we have attempted to show in this paper, the MCC of English which we have presented is a flexible research resource that can be deployed in a number of investigations adopting a variety of methodological set-ups and analytical approaches to uncover the features of mediated language. In the longer term, possible applications comprise comparing the language patterns typical of translated and non-native English in CONTE for certain phenomena against the patterns emerging in corpora of (spoken) English as a Lingua Franca (Seidlhofer, 2001; Mauranen, 2003) and learner English (Granger, 2003), so as to deepen our understanding of (the similarities and differences between) different forms of language mediation.

Acknowledgements

We would like to thank Adriano Ferraresi for very helpful comments on an earlier draft.

References

Altenberg, B. & S. Granger (2001) 'The grammatical and lexical patterning of MAKE in native and non-native student writing'. *Applied Linguistics* 22(2): 173-194.

Aston, G. (2001) 'Text categories and corpus users: a response to David Lee'. *Language*

Learning & Technology 5(3): 73-76.

Atkins, S., Clear, J. & N. Ostler (1992) 'Corpus design criteria'. *Literary and Linguistic Computing* 7(1): 1-16.

Baker, M. (1993) 'Corpus linguistics and translation studies: implications and applications'. In Baker, M., G. Francis & E. Tognini-Bonelli (eds.) *Text and Technology: In Honour of John Sinclair*. Amsterdam: John Benjamins. 233-250.

Baroni, M., S. Bernardini, F. Comastri, L. Piccioni, A. Volpi, G. Aston & M. Mazzoleni (2004) 'Introducing the *La Repubblica* corpus: a large, annotated, TEI(XML)-compliant corpus of newspaper Italian'. In *Proceedings of LREC 2004*. Lisbon: ELDA. 1771-1774.

Bernardini, S. & F. Zanettin (2004) 'When is a universal not a universal? Some limits of current corpus-based methodologies for the investigation of translation universals'. In Mauranen, A. & P. Kujamäki (eds.) *Translation Universals: Do They Exist?* Amsterdam: John Benjamins. 51-62.

Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finegan (1999) *Longman Grammar of Spoken and Written English*. London: Longman.

Blum-Kulka, S. (1986) 'Shifts of cohesion and coherence in translation'. In House, J. & S. Blum-Kulka (eds.) *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies*. Tübingen: Gunter Narr. 17-35.

Blum-Kulka, S. & E. Levenston (1983) 'Universals of lexical simplification'. In Færch, C. & G. Kasper (eds.) *Strategies in Interlanguage Communication*. London: Longman. 119-139.

Burnard, L. (2007) 'Users' reference guide to the British National Corpus (XML edition)'. Oxford: Oxford University Computing Services. Available online at `http://www.natcorp.ox.ac.uk/XMLedition/URG/` [last accessed 20 August

2008].

Cardinaletti, A. (2005) 'La traduzione: un caso di attrito linguistico'. In Cardinaletti, A. & G. Garzone (cur.) *L'Italiano delle Traduzioni*. Milano: Franco Angeli. 59-83.

Chesterman, A. (2004) 'Hypotheses about translation universals'. In Hansen, G., K. Malmkjær & D. Gile (eds.) *Claims, Changes and Challenges in Translation Studies*. Amsterdam: John Benjamins. 1-13.

Christ, O. (1994) 'A modular and flexible architecture for an integrated corpus query system'. In *Proceedings of COMPLEX'94*, Budapest, 1994. Available online at `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench` [last accessed 20 August 2008].

Færch, C. & G. Kasper (eds.) (1983) *Strategies in Interlanguage Communication*. London: Longman.

Granger, S. (2003) 'The International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research'. *TESOL Quarterly* 37(3): 538-546.

House, J. & S. Blum-Kulka (eds.) (1986) *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies*. Tübingen: Gunter Narr. 17-35.

Hunston, S. (2002) *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Kenny, D. (2001) *Lexis and Creativity in Translation. A Corpus-based Study*. Manchester: St. Jerome.

Kilgarriff, A. (2001) 'Comparing corpora'. *International Journal of Corpus Linguistics* 6(1): 97-133.

Laviosa, S. (1998) 'The English Comparable Corpus: a resource and a methodology'. In Bowker, L., M. Cronin, D. Kenny & J. Pearson (eds.) *Unity in Diversity? Current Trends in Translation Studies*. Manchester: St. Jerome. 101-112.

Laviosa, S. (2000) 'TEC: a resource for studying what is 'in' and 'of' translational English'. *Across Languages and Cultures* 1(2): 159-178.

Laviosa, S. (2002) *Corpus-based Translation Studies. Theory, Findings, Applications.* Amsterdam/New York: Rodopi.

Laviosa, S. (2003) 'Corpus and simplification in translation'. In Petrilli, S. (ed.) *Translation Translation*. Amsterdam: Rodopi. 153-162.

Lee, D. (2001) 'Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle'. *Language Learning & Technology* 5(3): 37-72.

Malmkjær, K. (2005) 'Norms and nature in translation studies'. *Synaps: Fagspråk, Kommuniksjon, Kulturkunnscap*. Bergen: Norges Handelshøyskole. 16: 13-19.

Mauranen, A. (2003) 'The corpus of English as lingua franca in academic settings'. *TESOL Quarterly* 37(3): 513–527.

Nesselhauf, N. (2004) 'Learner corpora and their potential for language teaching'. In Sinclair, J. McH. (ed.) *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins. 125-152.

Pym, A. (2008) 'On Toury's laws of how translators translate'. In Pym, A., M. Shlesinger & D. Simeoni (eds.) *Beyond Descriptive Translation Studies. Investigations in Homage to Gideon Toury*. Amsterdam: John Benjamins. 311-328.

Salsnik, E. (2007) 'Dagli universali traduttivi all'italiano delle traduzioni'. In Montella, C. & G. Marchesini (eds.) *I Saperi del Tradurre*. Milano: Franco Angeli. 101-131.

Schmid, H. (1994) 'Probabilistic part-of-speech tagging using decision trees'. *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, 14-16 September 1994.

Seidlhofer, B. (2001) 'Closing a conceptual gap: the case for a description of English as a lingua franca'. *International Journal of Applied Linguistics* 11(2): 133–158.

Selinker, L. (1972) 'Interlanguage'. *International Review of Applied Linguistics* 10(3): 209-231.

Serratrice, L. (2005) 'The role of discourse pragmatics in the acquisition of subjects in Italian'. *Applied Psycholinguistics* 26(3): 437-462.

Serratrice, L. (2007) 'Referential cohesion in the narratives of bilingual English-Italian children and monolingual peers'. *Journal of Pragmatics* 39(6): 1058-1087.

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

Toury, G. (1995) *Translation Studies and Beyond.* Amsterdam: John Benjamins.

Toury, G. (2004) 'Probabilistic explanations in translation studies: welcome as they are, would they qualify as universals?'. In Mauranen, A. & P. Kujamäki (eds.) *Translation Universals: Do They Exist?* Amsterdam: John Benjamins. 15-32.