Ed. María Sánchez **Salford Working Papers in Translation and Interpreting** **Vol. 1 (2012)**

Towards professional uptake of DIY electronic corpora in legal genres

Juliette Scott, *University of Portsmouth*

Abstract

This paper will describe current progress in a project which develops a DIY electronic corpus procedural methodology to assist legal translators. A particular focus is placed on corporate-related and court-related translation rather than legislation.

The methodology involves the targeted retrieval of authoritative legal texts from Internet repositories and other sources by professional legal translators themselves, according to their specific needs or those of specific projects. Results so far seem to indicate that compilation of such corpora can be achieved in 30-45 minutes, thus in line with users' expressed criteria.

Having carried out a pilot study with some initial testing with professional legal translators in certain legal genres, the current phase of the project involves performing more in-depth trialling of different language pairs and other legal genres, and looking at the issues of corpus representativeness in this context, target audience and translator expectations, the availability of corpus material, and quality assessment, whilst bearing in mind at all times feasibility in the workplace.

The theoretical framework currently being leveraged is genre and supergenre analysis in professional applications, and contributions will be drawn from, *inter alia*, the fields of corpus linguistics, corpus use for learning applications, and jurilinguistics.

Keywords: legal translation, terminology, DIY corpora, legal genres and subgenres, collocations

Introduction: Project background

Legal translation is subject to considerable difficulties, in part due to the highly entrenched nature of its sublanguages. Lawyers are taught legal language at law school as part of their studies, and paralegal staff undertake training, as attested by the plethora of courses on legal writing. Translators are, however, rarely trained in the law, although this may become a trend in the future – a number of specific legal linguistics and legal translation programs are currently being developed.

Whilst legal translators have a wide range of general tools at their disposal, such as dictionaries, glossaries, termbases, and online fora, target sublanguage conventions and appropriate collocations may escape them due to the dearth of legal thesauri and legal collocation dictionaries. Parallel corpora, generally used by professionals in the form of translation memories, may make a contribution, but availability is limited to the genre of legislation – very little parallel material can be found in corporate legal genres or for court proceedings and submissions.

Although corpus use in translation as a general field has been extensively researched since first highlighted by Baker (1993, 1995), Kenny (2001) and Olohan (2004), amongst others, professional uptake is considerably more limited, as demonstrated in the EU-funded MeLLANGE project survey (MeLLANGE, 2006) and in my own survey (Scott J, 2011:7-8) and discussed by Bowker (2004) and Bernardini (2006). The MeLLANGE results indicated that

58.2 % of professional translators did not use corpora in their work, while my own survey, based on the MeLLANGE questionnaire, found a figure of 61 %.

The project seeks to examine whether and to what extent electronic corpora created in legal subgenres by translators themselves according to their needs can assist them in producing translations that are closer in line with target audience expectations. The project also aims, in a distinct strand, to gather information from receivers of legal translations to better understand their requirements.

1. Legal genres

Bhatia (1987:227) provides a structure distinguishing the main legal genres according to their 'communicative purposes' and the 'lexico-grammatical, semantico-pragmatic, and discoursal resources' used in different legal contexts. Systematising such a complex supergenre is clearly valuable. However, concerning written genres, he differentiates between 'frozen legal documents like contracts, agreements, insurance policies, etc.' and 'formal' documents such as 'legislation, rules and regulations, etc.'; classifying them both under the term 'legislative', which seems somewhat equivocal.

Other classifications of the legal supergenre have been proposed, for example by Trosborg (1997:20) according to 'situation of use'. Kurzon (1997:120) differentiates between 'legal language […] employed when people talk about the law' and 'the language of the law' that he defines as institutional laying down of the law. A further classification has been made by Mattila (2006:4-5) according to the sub-groups of legal professionals. I will be offering an alternative structure aimed at addressing use within legal translation practice.

2. Legal language

Legal language may also be referred to as legal discourse(s) or legal linguistics, with the use of the terms 'register', 'variety' and 'sublanguage'.

In Legal Translation Explained (2002), Alcaraz Varó and Hughes provide a

detailed practical account for translators. The book deals with ‘legal English as a linguistic system’ (pp. 1-18), legal systems and branches of law, the translation of written and oral genres, and gives in-depth descriptions of legal translation problems and solutions. In particular (pp. 154-165), the authors establish three classes of lexis: ‘exclusively legal vocabulary’; polysemic ‘semi-technical vocabulary’ – words having one meaning in a legal context and another conventional meaning; and lastly words which maintain their usual meaning but are widely used across different legal genres and are ‘contextually bound’.

Tiersma (1999) catalogues the range of features characterising legal language, including sentence length, binomial expressions, doublets and triplets, the avoidance of pronouns, nominalisations, archaisms, as well as homonymy and polysemy in legal terms.

These linguistic attributes are examined by Mattila (2006) from a comparative perspective, in which differences between legal systems and concepts are brought to the fore.

The works by Šarc (1997) and Cao (2009) are exclusively dedicated to legal translation, but unfortunately both concentrate on legislation, and devote little space to the corporate texts and court proceedings that compose the majority of freelance translators’ work.

The importance of one key aspect of legal discourse, collocations, has been highlighted by many. In particular, Biel (2010) has highlighted their importance in translator training.

To summarise this section, the entrenched nature of legal discourse and its features have been thoroughly typified by various authors. In addition to alignment with these precepts, the differences occurring at system, genre and concept levels also represent a major challenge for translators in this field.

3. Corpora and Language for Special Purposes

Bowker and Pearson are the authors of a comprehensive work – Working with

Specialized Language: A practical guide to using corpora (2002). It covers the compilation of corpora and how they can be exploited in the field of language for special purposes (LSP): for specialised translation, technical writing and by LSP learners. It includes a chapter on electronic corpus processing tools, and despite having been published in 2002, remains extremely relevant today, possibly because these particular tools have not been radically upgraded (i.e. WordSmith Tools, Scott, 2010) or replaced, and possibly because the authors concentrate on techniques, output and why they can be used rather than aiming to provide a manual for a particular software package.

The book gives thorough guidelines for LSP corpus design covering the various criteria to be taken into consideration. Key term extraction and the building of glossaries are also dealt with. Readers are thus equipped to understand why specialised corpora may be used, how to compile and process them, and which applications can be envisaged.

Flowerdew (2004) put forward arguments to support the use of specialised corpora in understanding professional discourse. She noted (p. 26) that 'the compiler may well have to rely on a combination of *judgement and convenience* sampling as the data may only be available from a limited number of sources'. Flowerdew also broaches the subject of tagging and annotation in this context (p. 26), citing Tognini-Bonelli (2001:73) who endorses not approaching data in this fashion so as not to *insulate* it in any way and to achieve thorough exploitation. Furthermore, in the context of this project, where professional translators are the compilers, it would seem unrealistic to imagine that they might take the time to perform this process.

Corpus tools to contribute to language prowess as a part of professional training for fledgling lawyers was examined by Hafner and Candlin (2007), in a study carried out in Hong Kong. Some issues were observed with the use of the tools in this data-driven approach, leading the authors to conclude that further trials were required.

4. DIY corpora

A considerable amount of work has been carried out on DIY corpora – small corpora produced by a translator-compiler for a specific purpose. However, studies almost exclusively involve students rather than professionals. The only study involving professionals that I have been able to identify to date was carried out by Jääskeläinen and Mauranen (2006) with translators employed in the Finnish timber industry. Their study examined the levels of usage of electronic tools by translators in business, and also involved a trial of a corpus compiled by the researchers, made available to nine translators in a test group and queried using the software WordSmith Tools. Among the conclusions were opportunities for 'highly specialised corpora […] compiled for limited use by translators or companies themselves', stressing the need for guidelines on compilation and training for translators.

Nomenclature is not standardised – such 'DIY' corpora (Bowker, 2002, Zanettin, 2002) are also known as 'disposable' (Varantola, 2002), 'virtual' (Corpas Pastor and Seghiri, 2009), 'ad hoc' (Aston, 1999, Fletcher, 2004), 'do-it-yourself, disposable, specialised mini' corpora (Maia, 2002), and, directed at ELT use, 'quick and dirty' (Tribble, 1997), amongst other appellations.

Varantola (2002) examines 'disposable corpora as intelligent tools'. She explains the concept of 'intelligence' by referring to electronic dictionaries that 'adapt to users' needs and allow user profiling' (p. 171). She seeks to make tools 'more usable and user-friendly [...] without a great deal of irrelevant and redundant data' (p. 171). Regarding the disposable nature of the corpora concerned, she states however that 'disposable material can be recycled and refined to form part of a more permanent collection' (p. 175).

She examines the reasons for translators consulting corpora and also discusses 'serendipitous' finds (p. 178) in which translators may discover solutions to problems of which they were previously unaware, a concept also discussed by Wilkinson (2007). She concludes by expressing a proviso regarding the competence of users in manipulating the tools.

Varantola's study concerns the language pair Finnish-English and involved

non-native translators. Importantly, she does not analyse the issues of quality and representativeness of corpus material, and elected not to discuss the time required for compilation (p. 181).

5. Corpus 'representativeness' or 'reliability'?

McEnery, Xiao and Tono (2010) offer an extensive guide to the use of corpora for language studies, covering multiple aspects of their production and exploration, reproducing a selection of key literature in the field of corpus linguistics and several case studies. Definitions of corpus representativeness are discussed, differentiating general and specialised corpora in this regard and introducing the notions of 'closure' and 'saturation' (p. 15) to describe the point beyond which a language feature might no longer vary in a specialised corpus.

Corpas Pastor and Seghiri (2007, 2009) present software called ReCor to assess corpus representativeness with respect to minimum size in terms of number of words and of documents. They cite, for example, 275 documents and 2.5 million words for an English corpus in their study (2009:90). This may not be practicable in a professional context, and is at odds with work carried out by many other authors in the DIY corpora field who suggest far lower figures. Bowker and Pearson (2002), for example state: 'In our experience, well-designed corpora that are anywhere from about ten thousand to several hundreds of thousands of words in size have proved to be exceptionally useful in LSP studies'.

Unlike Varantola, who used authentic texts, these authors chose to include European Union legislation as a large part of the corpora produced in English and Spanish in their 2009 paper (p. 82). This may have been to increase its size in view of the above. We would challenge this decision with the quality-not-quantity argument put forward by many researchers (e.g. Bowker and Pearson, 2002:45-46). European legislation may contain translation errors or inconsistencies (see in this regard *Studies on translation*, 2012), as well as artificial terms – indeed Corpas Pastor and Seghiri themselves cite an example (2009:94-95) regarding a Spanish word that was replaced in general use by a

term with a less specific meaning owing to a mistranslation of a Council Directive. Such a source, albeit voluminous, may thus not be reliable. Additionally, it is often difficult to identify whether these documents are in fact originals or translations, due, *inter alia*, to 'hub language' practices used in translation at the European Commission.

Bhatia, Langton and Lung (2004), looking at corpus linguistics and language teaching and learning in legal contexts, cite Trosborg's work (1997) on a one-million word corpus of statutes and contracts (2004:207), reminding us that she 'found a remarkable degree of convergence, implying that even a smaller corpus would have given equally effective results'. The authors conclude, quoting Tribble (2001), *inter alia*, that 'the use of genre-based small corpora will be much more useful that large corpora covering a complete register of law' (2004:215). They further hold that 'legal discourse is so conservative in its construction, interpretation and use that it often does not require a large corpus to determine its linguistic frequencies' (2004:207).

Tribble had already examined this issue (1997) stating that 'if one wishes to investigate the lexis of a particular content domain, a specialist micro-corpus can often be much more useful than a much larger general corpus'. In a recent personal communication (February 11, 2012), he has suggested that a highly specialised DIY corpus could be referred to using such terms as an 'authoritative purposive corpus'.

It has been suggested to me by Tony McEnery (personal communication, July 15, 2011) that the issue of representativeness is a false argument since, by their very nature, such corpora are not and do not need to be representative of language as a whole.

In order to address the issue of 'big is best' versus 'small is beautiful' in this regard, I shall therefore focus on testing and demonstrating 'reliability', 'fitness for purpose' and 'authoritativeness'.

6. Translation practice, receivers' needs and functionalism

In a work devoted to the interaction between translation theory and practice, Chesterman and Wagner (2002) raise three key issues: differentiation of translators' service by offering several levels of quality to suit a range of purposes; 'maintaining a resemblance to the intended genre' (e.g. p. 96) when translating; and increasing the use of target language resources in the profession.

In the legal domain, Šarčević (2000) outlines the new focus placed by Reiss and other German scholars from the 1970s onwards on target audiences, to enable translators to select appropriate translation strategies, discussing *skopos* theory in particular. She also considers differentiation of purpose, and the corresponding variation in strategies. She cites Kelsen (1979:40) in classifying different groups of addressees or receivers as direct or indirect, and further develops approaches and advice for specific receiver groups.

In 1997, Nord, a member of the above German school, laid out a clear and detailed description of *Skopostheorie,* and functionalism in translator training, in which the subject of the translation brief is also explored. I have carried out some initial testing of brief templates with translation agencies, and hope to further develop and investigate this area as a possible contribution to quality improvement for legal translations. A number of questions in the receiver survey described below are devoted to that issue.

7. Work to date

In terms of electronic project management tools, I have used the iPad apps SG Project (FourthFrame) for project planning, iThoughtsHD (Scott C, 2011) for mind-mapping and the cloud-based Evernote (Evernote Corporation) for notes and journal-keeping.

7.1. Pre-study

As part of an MA thesis (Scott J, 2011), I carried out a pre-study for this project, involving an earlier version of the methodology in which the only substantial difference was the breaking down of the workflow into eight steps rather than the present five.

Testing was carried out first on myself, with the production of five DIY corpora in the subgenres of Memoranda and Articles of Association, Non-Disclosure Agreements, joint venture agreements, private banking conditions, and terms and conditions of employment. The subgenres were mainly chosen as a result of actual translation commissions.

The above tests led to an initial step of producing comparable corpora in order to contrast textual macrostructure being abandoned, despite its usefulness, due to the time required to carry it out (one day for a small corpus). The first software package used, WordSmith Tools (Scott, 2010), was replaced by AntConc (Anthony, 2010) after the first two corpora, since it proved to be far more user-friendly, with faster access to data, and is also free of charge.

Once the methodology had been self-tested, I then performed trials with three professional translators using interviews. This led to the production of DIY corpora for software agreements, power purchase agreements, court proceedings relating to personal injury, and codes of ethics. In total six language pairs were tested: EN>FI, FI>EN, IT>DE, FR>EN, IT>EN, and EN>IT.

The interviews seemed to indicate that the time investment required could be workable and acceptable to the respondents, and that the specialised corpora produced enabled concrete translation issues to be resolved. A significant level of interest in such corpora among professional translators was also identified during surveys carried out using the LinkedIn professional network and the SurveyMonkey platform, with 88 respondents at the time of publication, and a further 21 at a face-to-face translators' meeting in Milan, Italy. Acceptable times cited by respondents as regards corpus creation (collection of material and document conversion were between 30-45 minutes, with a maximum of half a day (one respondent only, and on condition that the project in question was large). I have continued with these trials, and am consistently able to create useful DIY corpora within a duration of 30 minutes.

7.2. The current DIY corpus methodology

I have produced a training pack to assist translators participating in the project

when compiling corpora, consisting of an illustrated handbook as well as a video guide in the form of a commented screencast slide show, filmed using the software ScreenFlow (Telestream). These multimedia tools can be downloaded from a simple project website that I created to streamline the administration of respondents. The website is password-protected in order to control participation in the study and avoid undue dissemination.

The methodology is defined as a process consisting of five steps: defining corpus parameters, collecting corpus material, manual assessment, file conversion and use in translation.

In the training pack, I have tried to stress the importance of defining clear parameters for the corpus, as laid out by Bowker and Pearson (2002), in particular: language; geographical perimeter, due to differences between UK/US English or Portuguese in Brazil and Portugal, for example; type of document (the subgenre); file type, since PDFs may be more freely available than .doc files in the legal domain and may show signatures or official stamps thus attesting to their authoritative nature; and date of publication, for instance where terminology has changed following a certain event or piece of legislation.

As regards corpus collection, study participants are free to leverage their own archives, and are also given advice on legal electronic document repositories, both with free access and on a subscription basis. A step-by-step process for advanced Google file search is also outlined.

The importance of a brief but indispensable manual assessment is strongly emphasised, to check that the texts are authoritative, include good quality language (are not obviously translations or badly written), and whether they comply with the criteria defined. This is feasible since the number of texts in a corpus for a highly specific legal subgenre remains manageable.

Since, at the present time, corpus query software such as AntConc (Anthony, 2010) and WordSmith Tools (Scott, 2010) can only accept .txt files, advice is then given on batch file conversion software. The importance of clear

labelling of files is stressed in the training pack, as highlighted by Maher, Waller and Kerans (2008) to enable sources to be identified at the corpus consultation stage. A suggested file name might include a language code, type of document (subgenre) and name(s) of party(ies). Some early participants in the study have also used MemoQ (Kilgray) to consult their corpora, which does not require this conversion stage. This needs to be further investigated.

In terms of corpus consultation and use in translation, the attention of study participants is drawn to the Key Word in Context (KWIC) concordance function, keyword lists, and cluster/collection functions. The possibility of consulting multiple subgenre corpora once compiled is also offered, where relevant for a translation.

7.3. Recruitment, data collection and management

In terms of recruitment of professional translators, to date I have used the following sources:

- a pool of 81 translators that had expressed interest in further research when I carried out the pre-study as part of a Masters' thesis;
- 106 translators that were present at a legal translation conference where I ran a workshop on the methodology;
- a notice posted on a private legal translators' group on Yahoo;
- a notice posted on an open legal translators' group on LinkedIn.

This first recruitment drive for the pilot study took place in early January 2012.

Clearly the above electronic sources may lead to certain type of 'internet-savvy' participant being foregrounded in the study, but I weighed this up with the advantages of obtaining global reach and thus a wider range of language pairs. By the very nature of their working methods today, translators should also tend towards this profile in any case. Wright (2006) carried out a study of the strengths and weaknesses of researching Internet-based populations and online survey research, in which he concludes that researchers may save considerable time using online survey tools, but must be aware of 'issues related to sampling frames, response rates, participant deception, and access to populations' in respect of their research aims.

These initial contacts have led to some inadvertent ‘snowballing’ (Denscombe, 2007:26-27, Noy, 2008) in the form of a post on a widely read translation technology blog, and contacts with two national translators’ associations on behalf of their members.

So far, as a result of the above recruitment drive, 37 translators have registered to participate in trialling the methodology (as at 15 February 2012). A wide range of language pairs is represented. Insufficient data is available at present to review the subgenres of corpora being compiled.

When contacting the translators I used the software MaxBulk Mailer (Max Programming) in order to send personalized emails to large numbers of potential respondents in a single action. Contact details can very easily be imported into this application from a Microsoft Excel or .csv file.

For data collection purposes, the Wufoo electronic survey platform has been leveraged. Wufoo is significantly less costly than SurveyMonkey, its sister company from the same group, which was the deciding factor in my choice. In terms of data security, the survey platform is fully compliant with the US-EU Safe Harbor Framework as set forth by the U.S. Department of Commerce regarding the collection, use, and retention of personal information from European Union member countries. This was considered as satisfactory by the University ethics review committee.

Wufoo offers more flexibility in graphic interfaces than SurveyMonkey whilst offering similar services in terms of question types. A full comparative study between the two platforms is, however, beyond the scope of this paper. Wufoo enables automatic email alerts to be sent each time a respondent completes a feedback form, which has proved to be extremely useful in administering the trial promptly.

Three forms are being used in the pilot study. A registration form, setting out the appropriate ethical information, is used to collect data relating to the translator’s profile. Upon receipt of this form, a participant code is issued,

enabling the remaining data to be anonymised. Participants are asked to provide feedback using two separate forms – one relating to corpus compilation, and another on corpus use in translation.

The platform can also generate reports but there are some issues which the Wufoo support team have informed me are not likely to be resolved in the foreseeable future. Some of the drawbacks encountered with the reports include:

- Limited number of graphics can be generated per form;
- Longer question wording cannot be made visible;
- WYSIWIG PDFs cannot be produced so graphic representations can only be visualised on-screen or using screen grabs;
- Likert scale data cannot be represented;
- Two or more language versions of the same form cannot be merged in a report.

The above reporting disadvantages should be easy to overcome by using another package or platform, since Wufoo allows the data to be exported in Excel format. In this respect I am currently looking at other, more powerful software such as NVivo or SPSS, and the online data analysis platform Dedoose (SocioCultural Research Consultants).

Data collection is at a very early stage at present, and I shall therefore not give details here.

7.4 Identifying the expectations and requirements of receivers of legal translations

As mentioned earlier, a distinct strand of the project aims at clarifying the expectations and requirements of the target audience. A survey is thus to be carried out of lawyers, process servers, in-house counsel, law firms, judges, court officials, legal translation agencies, and translation quality departments dealing with legal texts for multinational firms. A first pilot study is under way, and the main study is due to be launched later this year.

Although online recruitment has been attempted using LinkedIn lawyers'

groups, this had not led to any results. With this group, it has appeared preferable to leverage personal contacts and proactive snowballing. This has been possible due to my ‘action research’ position in this project as a legal translator.

The International Standing Conference of University Institutes of Translating and Interpreting (CIUTI) held on 26-27 January 2012 in Geneva proved to be a good source of contacts, and led to 5 survey replies from outside the world of academia, despite the title of the conference.

Initial data from this pilot study has been of high quality, from a wide range of respondents both in terms of their roles and geographical spread. As at 17 January 2012, 18 completed questionnaires have been received. Examples of those replying are: the president of an international association, a judge, two in-house counsel working for international groups, translation quality managers for three multinational corporations, and directors of high-profile legal translation agencies. The quality of the data and level of interest expressed by participants seem to suggest that it might be valuable to carry out interviews at a later stage, to examine more closely the points raised in questionnaires received.

The Wufoo platform, as described above, is being used for data collection, with a single form comprising ethics information, respondent profile details and the questionnaire itself.

8. The next stage of the project

8.1. DIY corpus testing

Data collection will be continued, and completion of the pilot stage is planned for April 2012. According to the results, adjustments may be made to the collection methods and/or the methodology. The main study should be launched before the summer. Due to the way in which translators have been signing up for the project, which has been piecemeal, rather than in waves when calls for participation are issued, it is possible that the move from the pilot to the main study may take place in a fairly seamless way. This will obviously depend on the number of changes required in the methods.

In terms of additional recruitment, several avenues are available, including, but not limited to, the legal groups of three professional translators' associations, and supplementary online groups.

In the light of data collected, the initially planned interview stage for translator respondents will be reassessed. If it is to be carried out, use will be made of remote desktop sharing software such as TeamViewer (TeamViewer GmbH), and videoconferences or audioconferences will likely take place using Skype rather than webconferencing platforms such as Yuuguu (YG Technologies) or Watchitoo (Watchitoo Inc.), which have proved to be less than satisfactory in exploratory tests.

Subsequently, a follow-up questionnaire is planned to examine uptake levels in the long term.

8.2 Extension of the receiver survey

Completion of the pilot survey of receivers is planned for April 2012. Results are to be presented at a conference on language and the law in May, and it is hoped that this will lead to further interest from the legal profession.

The main study should be launched in July 2012. Adjustments may be made to the online questionnaire, although at present this does not appear necessary. Despite the fact that this strand of the project was planned as a questionnaire only at the outset, the data collected so far (see above) seems to indicate that more in-depth interviews may be warranted.

In addition, it is hoped that attendance at two other conferences on language and the law will lead to other avenues for the recruitment of receivers.

The major difficulties with the population of legal professionals are access and obtaining replies. It transpires from first contacts that those who have a personal interest in language seem much more likely to respond. Attempts will be made to employ the snowballing technique as described by Denscombe (2007:26-27) and Noy (2008) in order to increase the number of respondents. Ventures to involve official entities, such as Bar Councils, have so far proved unsuccessful.

8.3 Refinements to the methodology

Since software is evolving constantly, the packages being used, such as corpus query applications, optical character recognition (OCR) programs, and file converters, will be reappraised at regular intervals.

Web-based corpus platforms such as Sketch Engine (Lexical Computing) and BootCat (Baroni, Zanchetta and Shaoul, 2011) will also be compared with the current stand-alone methodology, bearing in mind, however, the issues of confidentiality that necessarily apply to the legal field.

Other tools such as Wmatrix (Rayson, 2012) and ApSIC Xbench are also to be assessed, although approachability by translators is a primary concern here. Input from translator respondents will be taken into account, for example the use of a function of the translation memory package MemoQ (Kilgray) to consult the corpora and act as a concordancer, which has been already raised.

Further research will be carried out to identify free access electronic document repositories where legal texts may be found, for as many language pairs as possible. Text retrieval as a discipline is to be explored for any relevant applications that might bear upon the project.

Corpus linguistics theory will be further investigated for ways to establish, within acceptable time constraints, optimum corpus size ranges providing reliable answers to translators' terminology queries. The use of stop lists and reference corpora to generate keyword lists will also be more fully evaluated. In particular, as a result of advice proffered by experts in the field at a workshop where the project was presented, comparisons between the effects of different types of reference corpus will be made, including the impact of a large corpus constituted of exclusively legal texts.

9. Conclusion

It is hoped that this working paper has enabled the reader to have an overview of a project aimed at establishing whether the riches offered by corpora, long present in the world of academia, can at last be adopted by professional translators to their benefit, and, by extension, afford improvements to the quality of legal translations, rendering them closer to receivers' expectations and requirements.

References:

Alcaraz Varó, Enrique, and Hughes, Brian (2002). *Legal translation explained.* Manchester: St Jerome.

Aston, Guy (1999). Corpus use and learning to translate. *Textus*, *12*(2), pp. 289–314.

Baker, Mona (1993). Corpus linguistics and translation studies. Implications and applications. In M. Baker, G. Francis and E. Tognini-Bonelli (Eds.), *Text and Technology: In Honour of John Sinclair* (pp. 233-50). Amsterdam: John Benjamins.

Baker, Mona (1995). Corpora in translation studies: An overview and some suggestions for future research, *Target 7*(2), pp. 223–43.

Bernardini, Silvia (2006). Corpora for translator education and translation practice Achievements and challenges*. Third International Workshop on Language Resources for Translation Work, Research & Training*, 17-22. Retrieved August 13, 2011 from http://hnk.ffzg.hr/bibl/lrec2006/workshops/W17/proceedingsLR4TransIIIey.pdf#page=23

Bhatia, Vijay K. (1987). Language of the law. *Language Teaching 20*, pp. 227-234.

Bhatia, Vijay K., Langton, Nicola M., and Lung, Jane (2004). Legal discourse: Opportunities and threats for corpus linguistics. In U. Connor, T. A. Upton (Eds.), *Discourse in the professions. Perspectives from corpus linguistics* (pp. 203-231). Amsterdam: John Benjamins.

Biel, Lucja (2010). Corpus-based studies of legal language for translation purposes: methodological and practical potential. In C. Heine and J. Engberg (Eds.), *Reconceptualizing LSP: Online proceedings of the XVII European LSP Symposium 2009*. Aarhus: Aarhus School of Business, Aarhus University. Retrieved August 13, 2011 from http://www.asb.dk/fileadmin/www.asb.dk/isek/biel.pdf

Bowker, Lynne (2002): 'Working together: A collaborative approach to DIY corpora'. In E. Yuste-Rodrigo (Ed.), *Language Resources for Translation Work and Research, LREC 2002 Workshop Proceedings, Las Palmas de Gran Canaria*,

29–32. Retrieved August 28, 2011 from http://www.lrec-conf.org/proceedings/lrec2002/pdf/ws8.pdf

Bowker, Lynne (2004). Corpus resources for translators: academic luxury or professional necessity? *TradTerm, 10*, pp. 213-247.

Bowker, Lynne and Pearson, Jennifer (2002). *Working with specialized language: a practical guide to using corpora*. London: Routledge.

Cao, Deborah (2009). *Translating law*. Clevedon: Multilingual Matters Ltd.

Chesterman, Andrew, and Wagner, Emma (2002). *Can theory help translators?: a dialogue between the ivory tower and the wordface*. Manchester: St Jerome.

Corpas Pastor, Gloria and Seghiri, Miriam (2007). Specialized corpora for translators: A quantitative method to determine representativeness. *Translation Journal*, 11(3). Retrieved August 13, 2011 from http://accurapid.com/journal/41corpus.htm

Corpas Pastor, Gloria and Seghiri, Miriam (2009). Virtual corpora as documentation resources: Translating travel insurance documents (English-Spanish)*. In A. Beeby, P. Rodríguez Inés and P. Sánchez-Gijón (Eds.), *Corpus use and translating: corpus use for learning to translate and learning corpus use to translate* (pp. 75-107). Amsterdam: John Benjamins.

Denscombe, Martyn (2007). *The good research guide for small-scale social research projects*. Maidenhead: McGraw-Hill Open University Press.

Fletcher, William H. (2004). Facilitating the compilation and dissemination of ad-hoc web corpora. In G. Aston, S. Bernardini and D. Stewart (Eds.), *Corpora and Language Learners* (pp. 273–300). Amsterdam: John Benjamins.

Flowerdew, Lynne (2004). The argument for using English specialized corpora to understand academic and professional language. In U. Connor, T. A. Upton (Eds.), *Discourse in the professions. Perspectives from corpus linguistics* (pp. 11-33). Amsterdam: John Benjamins.

Hafner, Christoph, and Candlin, Christopher (2007). Corpus tools as an affordance to learning in professional legal education. *Journal of English for Academic Purposes*, *6*(4), pp. 303-318. doi: 10.1016/j.jeap.2007.09.005

Jääskeläinen, Riitta and Mauranen, Anna (2006). Translators at work: a case study of electronic tools used by translators in industry. In G. Barnbrook, P.

Danielsson, and M. Mahlberg (Eds.), *Meaningful texts: the extraction of semantic information from monolingual and multilingual corpora* (pp. 48-53). London: Continuum International.

Kenny, Dorothy (2001). *Lexis and creativity in translation: a corpus-based study*. Manchester: St. Jerome.

Kurzon, Dennis (1997). 'Legal language': varieties, genres, registers, discourses. *International Journal of Applied Linguistics, 7*(2), pp. 119-139.

Maher, Ailish, Waller, Stephen and Kerans, Mary E. (2008, July). Acquiring or enhancing a translation specialism: The monolingual corpus-guided approach. *The Journal of Specialised Translation, 10*. Retrieved August 13, 2011 from http://www.jostrans.org/issue10/art_maher.php

Maia, Belinda (2002). Do-it-yourself, disposable, specialised mini corpora – where next? Reflections on teaching translation and terminology through corpora. *Cadernos de Tradução IX - Tradução e Corpora*, *1*(9), pp. 221-236. Retrieved August 13, 2011 from http://www.periodicos.ufsc.br/index.php/traducao/article/view/5987/5691

Mattila, Heikki E.S. (2006). *Comparative legal linguistics*. Aldershot: Ashgate.

McEnery, Tony, Xiao, Richard, and Tono, Yukio (2010). *Corpus-based language studies: an advanced resource book*. London: Routledge.

MeLLANGE (Multilingual eLearning in LANGuage Engineering) (2006, April 20). *Corpora & e-learning questionnaire results summary*. Retrieved January 13, 2011 from http://mellange.eila.univ-paris-diderot.fr/

Nord, Christiane (1997). *Translating as a purposeful activity*. Manchester: St Jerome.

Noy, Chaim (2008). Sampling knowledge: the hermeneutics of snowball sampling in qualitative research. *International Journal of Social Research Methodology 11*(4), pp. 327-344.

Olohan, Maeve (2004). *Introducing corpora in translation studies*. London: Routledge.

Šarčević, Susan (1997). *New approach to legal translation*. The Hague: Kluwer Law International.

Šarcevic [sic], Susan (2000, February 17-19). *Legal translation and translation theory: a receiver-oriented approach*. Paper presented at Legal translation, history, theory/ies, and practice. Retrieved August 29, 2011 from http://tradulex.org/Actes2000/sarcevic.pdf

Scott, Juliette R. (2011). *DIY corpora: a pearl in the legal translator's sea of tools*. Unpublished masters dissertation, University of Portsmouth, Portsmouth.

Studies on translation and multilingualism: Quantifying quality costs and the cost of poor quality in translation (2012). Luxembourg: Publications Office of the European Union.

Tiersma, Peter M. (2000). *Legal Language*. Chicago: The University of Chicago Press.

Tribble, Christopher (1997, April 12-14). *Improvising corpora for ELT: Quick-and-dirty ways of developing corpora for language teaching*. Paper presented at Practical Applications of Language Corpora, Lodz, Poland. Retrieved August 13, 2011 from http://www.ctribble.co.uk/text/Palc.htm

Trosborg, Anna (1997). *Rhetorical strategies in legal language: discourse analysis of statutes and contracts*. Tübingen: Gunter Narr Verlag Tübingen.

Varantola, Krista (2002). Disposable corpora as intelligent tools in translation. *Cadernos de Tradução IX – Tradução e Corpora*, *1*(9), pp. 171-189.

Wilkinson, Michael (2007, January). Corpora, serendipity & advanced search techniques. *The Journal of Specialised Translation, 7*. Retrieved August 28, 2011 from http://www.jostrans.org/issue07/art_wilkinson.php

Wright, Kevin B. (2006). Researching Internet-based populations: advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. *Journal of Computer-Mediated Communication 10*(3). doi: 10.1111/j.1083-6101.2005.tb00259.x

Zanettin, Federico (2002). Corpora in translation practice. In E. Yuste-Rodrigo (Ed.), *Language resources for translation work and research*, *LREC 2002 Workshop Proceedings*, *Las Palmas de Gran Canaria*, 10-14. Retrieved August 29, 2011 from http://www.lrec-conf.org/proceedings/lrec2002/pdf/ws8.pdf

Software and electronic tools

Anthony, Laurence (2010). AntConc (Version 3.2.0m) [Computer software]. Tokyo: Laurence Anthony. Retrieved June 7, 2010 from http://www.antlab.sci.waseda.ac.jp/antconc_index.html

ApSIC Xbench (Version 2.9) [Computer software]. Barcelona: ApSIC S.L.

Baroni, Marco, Zanchetta, Eros and Shaoul, Cyrus (2011). BootCaT [Computer software]. Bologna: Scuola Superiore di Lingue Moderne per Interpreti e Traduttori.

Dedoose.com. Manhattan Beach, California: Sociocultural Research Consultants, LLC.

Evernote (Version 3.0.1) [Computer software]. Mountain View, California: Evernote Corporation.

MaxBulk Mailer (Version 8.3.5) [Computer software]. Denia: Max Programming LLC.

MemoQ [Computer software]. Budapest: Kilgray Translation Technologies.

Rayson, Paul (2012). Wmatrix. Lancaster: Paul Rayson. Retrieved February 18, 2012 from http://ucrel.lancs.ac.uk/wmatrix2.html

Scott, Craig (2011). iThoughtsHD (Version 4) [Computer software]. York: Craig Scott.

Scott, Mike (2010). WordSmith Tools (Version 5.0) [Computer software]. Oxford: Oxford University Press. Retrieved June 5, 2010 from http://www.lexically.net/wordsmith/index.html

ScreenFlow (Version 3.0.1) [Computer software]. Nevada City, California: Telestream.

SG Project (Version 3.5) [Computer software]. Dublin, Ohio: FourthFrame.

Sketch Engine (2011). Brighton: Lexical Computing Ltd.

SurveyMonkey.com. Palo Alto: SurveyMonkey.com, LLC.

TeamViewer (Version 4.1.8780) [Computer software]. Göppingen: TeamViewer GmbH. Retrieved January 12, 2011 from http://www.teamviewer.com/download/index.aspx

Watchitoo.com. New York, New York: Watchitoo Inc.

Wufoo.com. Palo Alto: SurveyMonkey.com, LLC.

Yuuguu.com. Manchester: YG Technologies Ltd.