JOURNAL OF ENGLISH STUDIES - VOLUME 5-6 (2005-2008), 309-324

TOOLS FOR ENGLISH-SPANISH CROSS-LINGUISTIC APPLIED RESEARCH[1]

ROSA RABADÁN
University of León

ABSTRACT. *Empirically-based, cross-linguistic research should have a central role to play in offering solutions to applied problems. However, this role remains largely unexploited and the transformation of contrastive descriptive findings into useful applications has received little attention. Applied professionals looking for useful and reliable aids to assist them in their cross-linguistic routines see their needs ignored as supplying solutions is not generally considered as part of the research process. Part of the problem derives from the inadequacy of the existing tools for turning declarative knowledge into performative knowledge to serve particular applied purposes. This paper aims at re-defining the role of both new and already existing tools in terms of their contribution to applied research.*

1. INTRODUCTION

Applied linguistic research means different things in different contexts. In the context of the real world what applied users require, and indeed desire, are ready-to use aids built on reliable findings. In the academic context, however, the research community tends to dismiss applied needs as either beyond their concern or to be sorted out by the applied users themselves. Researchers do not consider supplying solution(s) as part of their duty and applied users are expected to derive the aids themselves from state-of-the-art research (Sinclair 2004).

1. Research for this paper has been funded by the research grants HUM2005-01215/FILO (Spanish Ministry of Education and the ERDF) and CO03/102 (Castile and León Regional Government). Further information about the ACTRES (*Análisis Contrastivo y Traducción Inglés-Español* / Contrastive Analysis and Translation English-Spanish) research program at http://actres.unileon.es

Proposals that seek to bridge this gap between researchers' and users' expectations are generally linked to corpus-based work. In terms of methodology, corpus linguistics has played a central role in the construction of tools which (assumedly) serve to improve and upgrade user performance. Nevertheless, rather than actual applications, most of the new possibilities involve corpus work by the final user (Bowker and Pearson 2002). Applied users tend to consider the rapid access to empirical evidence and the immediate feedback that may be obtained as aids. But corpora, of whichever type, do not provide applied users with answers and/or ready-made solutions (Rabadán in press). In addition there is the problem that lies in the unpredictability of the results of these searches when the user is an applied professional (Bernardini 2000).

In many respects, these issues belong in the domain of *cognetics*, i.e., cognitive engineering. Cognetics is a kind of 'ergonomics of the mind' and its purpose is to take into account the capabilities and limitations of the human mind when designing a user interface (Foraker Design 2002-2005) or, as in this case, tools for English-Spanish applications. Two of the more general and central concepts in cognetics are usefulness and usability.

Usefulness is a performance indicator associated with the extent to which tools (technological, conceptual or otherwise) are actually relevant to the actual needs of a user. When research has an applied goal, not every single phenomenon which is interesting from a descriptive point of view is necessarily relevant, however those which are tend to be associated with frequent problems in cross-linguistic practice.

In addition to being useful, descriptive findings need to work as an efficient tool for applied purposes. This requirement is known as *usability*. This parameter refers to (a) the way the users' abilities and limitations have/have not been considered when building a product and (b) the user satisfaction it arouses in terms of learnability, effectiveness and efficiency (Byrne 2006: 154-156).

This paper sets out to identify the role(s) and assess the contribution to cross-linguistic applied research of a set of tools in terms of their usefulness and usability.

2. IDENTIFYING THE TOOLS AND THEIR CONTRIBUTION

This cross-linguistic applied research uses four types of tools: two technical, namely, computerized corpora and statistics; one conceptual, i.e., cross-linguistic labelling; and one evaluative, which comprises a control group of representative users.

The empirical comparable data come from two large monolingual general corpora, the BoE (http://www.collins.co.uk/books.aspx?group=153) and the CREA

(http://corpus.rae.es/creanet.html), which serve as source corpora. Here 'source' means that the chosen monolingual corpora are used as a starting point to build a comparable corpus designed to address cross-linguistic problems (English-Spanish). The result is two subcorpora, one English, the other Spanish, that include the same type of textual materials and that have some 30 million words each. The role of the comparable corpus is to supply empirical data related to the correct grammatical usage in both English and Spanish of each phenomenon. Evidence obtained from this source brings the prescriptive component into the process, which is at the core of everything applied (Toury 1995).

P-ACTRES is a do-it-yourself (DIY) parallel corpus containing original texts in English and their translations into Spanish (McEnery, Xiao and Tono 2006: 71). The role of P-ACTRES is to contribute empirical diagnostic data that are to be contrasted with those obtained from the comparable corpus. When applied to translation data 'diagnostic' means 'instrumental' in that they are not the object and/or goal of the study but a means to obtain complementary information about language use and cross-linguistic interpretation.

In producing cross-linguistic analysis quantitative and qualitative data are required, and the need to interpret both types of findings arises. Statistics help in keeping the analysis of corpus-based materials within manageable limits. Furthermore they attest the representativeness of the samples to be analyzed. Inferential statistics are an invaluable aid in interpreting the results from the corpus-based contrast and constitute the bridge between quantitative and qualitative analysis. This is so because inferential tests tell us whether statistical significance between proportions of different rank (original vs. translated language) exists.

The conceptual tool consists of a set of labels relevant for cross-linguistic discrimination of grammatical meaning English-Spanish. The role of the labels is to help identify the common ground (or the opposite) between English and Spanish during the experimentation process, i. e., to act as *tertium comparationis* (Krzeszowski 1990). Without them it would not be possible to collate the data, as there would be no systematic relationship between English and Spanish phenomena.

The informants are prototypical applied users that act as a 'control group' in providing feedback concerning the relevance and usefulness of findings as well as usability recommendations. Their suggestions function as working assumptions or on-the-road evaluative comments throughout the research process.[2]

2. Suggestions from the 'control group' do not preclude extensive assessment and evaluation of future applications derived from this research.

These four tools can be used profitably in different sequential combinations, which are designed to yield useful and usable results for English-Spanish applied activities (Rabadán 2007).

3. MONOLINGUAL CORPORA AS 'SOURCE' CORPORA: BUILDING COMPARABLE CORPORA ENGLISH-SPANISH

One important decision when setting out to research cross-linguistic problems (English ↔ Spanish) is to take advantage of already existing corpora whenever their usefulness is pertinent. Since corpora are always designed for a particular purpose, judging whether ready-made resources could actually be useful sources for our particular aim was crucial. There is only one foolproof measure for this: a consideration of the research question(s). Here 'source' means that the chosen monolingual corpora are used as a starting point to build a comparable corpus designed to address cross-linguistic problems (English-Spanish). Taking this need for flexibility into account a number of procedural decisions were made regarding three primary concerns: availability and representativeness in English and in Spanish; qualitative and quantitative comparability; and usability of the chosen corpora.

While there are abundant corpus resources for English, in Spanish the choice is more restricted.[3] For this reason the selection process commenced with a consideration of the possibilities in Spanish and their potential 'matches' in English as well as their usability. For Spanish it soon became clear that the most promising candidate was the CREA; for English both the British National Corpus (BNC) and the BoE were considered. Key factors in opting for the BoE corpus, however, were: the classification of mode and domain; and size of the texts (samples in the BNC) and their chronology. A "general corpus" typically means that the corpus is balanced with regard to the varieti(es) of the language and with regard to genres and domains (McEnery, Xiao and Tono 2006: 59). Both the BoE and CREA are large corpora (524 and 170 million words respectively) which contain different varieties of English (British and American mainly) and Spanish (European, Andean, Caribbean, Central, Chilean, Mexican and Rioplatense) respectively. The Collins Wordbanks *Online* English corpus is the commercially

3. For English, a good and balanced overview in McEnery, Xiao and Tono (2006: 59-70); for Spanish see http://liceu.uab.es/~joaquim/language_resources/lang_res/Corp_text_esp.html which covers CREA, CUMBRE and LEXesp; http://www.uam.es/proyectosinv/woslac/DOCUMENTS/CEDEL2%20-%20call%20for%20collaboration.pdf for CEDEL and http://www.bds.usc.es/ for ARTHUS. All visited November 2006. The listings are meant to be indicative, and do not claim to be comprehensive.

available section of the BoE which has been used for this enquiry. It is composed of 56 million words distributed, as in the CREA, into contemporary written and spoken text. For both corpora, the subcorpora selected are: "books", "newspapers", "magazines" and "ephemera/miscellaneous". Since the final results of our research aim at building applications for written language activities, the "spoken" language subcorpora have not been considered.

A key feature of these ready-made corpora is the possibility offered by CREA to select the chronology of the materials. As the BoE does not offer the chronological restriction feature, a primary corpus building strategy has been to start from the (default) English corpus and then go on to use the chronological selection feature in CREA so as to obtain a statistically comparable volume of materials in Spanish (Rabadán 2005: 60-61). This results in two subcorpora, one English, the other Spanish, that include the same type of textual materials and that have some 30 million words each (resulting from a trimming down of the original 'source' corpora. Qualitative comparability has been achieved by using the 'geographical variety' and the 'domain' features, both of which are present in the BoE and in CREA. These corpus selection strategies satisfy the mutual suitability of the English and Spanish language materials for the purposes of this study in terms of both quantity and quality.

The final parameter influencing procedural decisions is usability. For researchers as users of the BoE and CREA much of this work has been done, since each 'source' corpus supplies its own browser and navigation styles. An informal 'control group' of researchers reported satisfaction with corpora content and organization, which help effectiveness, whereas the same group expressed on-and-off disappointment with the software tools, which occasionally detract from efficiency. Yet, on the whole, the benefit of using available corpora was deemed to outscore the arduous task of building an English-Spanish bilingual comparable corpus from scratch.

4. THE NEED FOR DIAGNOSTIC DATA: BUILDING P-ACTRES[4]

The principal factors when planning P-ACTRES were the research questions to be addressed and a number of practical considerations concerning corpus size and hence the availability of textual materials.

---

4. The acronym stands for *parallel ACTRES*, which in turn is the Spanish-English mixed language acronym for the research group (and long-term research project) 'Contrastive analysis and translation English-Spanish' (*Análisis contrastivo y traducción inglés-español*). It is the name of a general language bilingual corpus compiled by group members M. Izquierdo, B. Labrador, R. Rabadán and N. Ramón

The corpus was intended as a source of diagnostic data in contrastive grammatical analyses English-Spanish. Although an independent corpus, P-ACTRES is intended to be used mainly as a diagnostic tool in combination with the phenomenon-specific comparable bilingual corpora English-Spanish sampled from the BoE and CREA. A third verification (or control) research protocol contemplates its use in combination with the CREA subcorpus used for each particular analysis (as part of the comparable bilingual corpus).

It was decided to use the sampling scheme of both the BoE and CREA so as to comply with the requirements of balance and representativeness, two corpus features that are best interpreted loosely, as noted by Hunston (2002: 28-30). The standards of balance, and, representative of what, are questions that do not lend themselves to an easy answer and, if there is one, it will most likely be debatable. In this case balance means that P-ACTRES reflects the qualitative composition of both the BoE and CREA, which is taken to be an indication of balance according to the adopted models.[5]

Representativeness is an important consideration when building a corpus and is closely connected to two further features: size and sampling. There is no standard size for a corpus to be representative, only recommendations that are always open to question. While a very influential proposal is the convention of considering the threshold of representativeness at 1 million words per language (Biber 1993), the discussion chiefly concentrates on whether to use large or small corpora and for what purposes.

Following the slogan 'no data like more data' large, multi-million word corpora, have been the normal practice in corpus analysis for lexical purposes. Yet, smaller corpora have proven useful and representative for purposes other than lexical analysis (Flowerdew 2001, Connor and Upton 2004). What can be considered as adequate size is relative to the purpose(s) the corpus serves as well as to practical issues.

---

and used for researching different problem-trigger aspects of grammatical contrast English-Spanish. Another group member, K. Hofland (AKSIS, University of Bergen) has contributed expert advice and help concerning computing issues.

5. Corpus balance is largely 'a matter of faith' as there is no reliable scientific measure for it. For a most authoritative source, (McEnery, Xiao and Tono 2006: 16). Yet seeking to achieve it by adopting an already existing (assumedly balanced) corpus as a model is an accepted and acceptable procedure in practice. P-ACTRES builders have considered this strategy convenient and methodologically sound for their research goals. Corpus data must be treated with caution and conclusions and/or generalizations derived from them are not to be seen as hard-and-fast rules (if such a thing can be said of language production), but rather as deductions or tendencies drawn from empirical information.

P-ACTRES has been designed to work effectively for application-oriented, cross-linguistic grammatical research. The rate of recurrence is higher for grammatical than for lexical phenomena as the former constitute a closed, finite set, which means that a given grammatical phenomenon will occur more often than one particular lexical item in the corpus regardless of its size.

Since it obviously limits corpus size, the availability of materials (machine-readable and/or paper-based) is a primary consideration. As observed by Zanettin (2000), this is particularly true in the case of parallel (or translation) corpora, as it is not unusual to have a very unbalanced situation across languages and types of text. The conversion of paper-based materials to electronic form has to be taken into account as it continues to be costly and necessarily involves a higher degree of errors due to human intervention. A further issue concerning available materials is copyright and related legal implications.[6] This can seriously affect the size of the parallel corpus since the corpus builder(s) are responsible for seeking copyright clearance from the copyright holders of both English and Spanish materials. It also determines whether the corpus can be made freely and/or commercially available.

The P-ACTRES corpus features over 2 million words distributed among the same types of textual material contained in the monolingual subcorpora (see Table 1 below). All the translated materials are reviewed for "threshold quality" before becoming part of the corpus. The "threshold quality test" reviews two aspects: overall intelligibility in Spanish and "degree of semantic match" between the original and the translation. P-ACTRES is an open corpus and its copyrighted materials cover the period 2000-2006.[7] Due to legal restrictions P-ACTRES includes chunks of about 15,000-20,000 words each rather than complete texts. The vast majority of the materials for the subcorpora 'books fiction', 'books non-fiction' and 'miscellaneous' are paper-based, those for 'newspapers' and 'magazines' were downloaded from their respective Internet sites.[8]

6. As a general rule, in the EU texts are copyrighted for their authors' lifetime plus 70 years. For more information see http://ec.europa.eu/internal_market/copyright/docs/docs/1993-098_en.pdf. For copyright regulations in the USA see http://www.copyright.gov/circs/circ01.pdf. In Canada refer to http://laws.justice.gc.ca/en/C-42/index.html.

7. There are two exceptions to this chronology dated 1995 and 1998 respectively.

8. About electronic data capture, cleaning and conversion tools see McEnery, Xiao and Tono (2006: 73-74).

P-ACTRES CORPUS	ENGLISH	SPANISH	TOTAL
Books Fiction	396,462	421,065	817,527
Books Non-Fiction	494,358	553,067	1,047,425
Magazines	111,770	117,828	229,598
News	132,006	147,967	279,973
Miscellaneous	40,178	49,026	89,204
Total words	1,174,774	1,288,953	2,463,727

Table 1. *Composition P-ACTRES corpus (June 2007).*

P-ACTRES adheres to the mark-up scheme of the Text Encoding Initiative (TEI) (Sperberg-McQueen and Burnard 2004) so as to facilitate exchange of information and usability. This (opaque) declaration of intentions indicates that contextual and organizational information has been encoded in the corpus so as to instruct its users on how the contents of the text are presented. P-ACTRES has ensured this is as basic as possible: the texts conform to XML[9] and the body of each text is tagged for structural elements (Izquierdo, Hofland and Reigem forthcoming). The software used is XML editor Oxygen 5.0 (SyncRO Soft Ltd 2002-2007). Traceability is also warranted so each text is assigned an ID code (P-ACTRES browser interface figure 1, 'text' window).

When considering corpus annotation, again the guiding criteria were usefulness, usability, and, of course, feasibility. On these grounds, it was decided to keep annotation as neutral as possible using generic part of speech (POS) tagging. The advantages of this decision are evident, as the corpus materials are used to address different research questions concerning different cross-linguistic grammatical issues that would have called for different corpora, if restrictiveness rather than broad coverage, had been the choice. P-ACTRES is then POS tagged using traditional grammatical categories by means of the IMS[10] Tree Tagger (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/), a language independent tagger that also includes lemma information, with a success rate of

9. eXtensible Markup Language, an improved offshot of the Standard Generalized Markup Language (SGML: ISO 8879).

10. IMS stands for *Institut für Maschinelle Sprachverarbeitung,* the Institute for Natural Language Processing at the University of Stuttgart, Germany.

96.4%, which makes it quite reliable. The POS labels can be accessed by scrolling down the POS boxes in the P-ACTRES browser interface (Fig.1).

The alignment program used in P-ACTRES is a new and refined version of the Corpus Translation Aligner (Hofland and Johansson 1998), created as the alignment tool for the English-Norwegian Parallel Corpus (ENPC) (http://www.hf.uio.no/ilos/forskning/forskningsprosjekter/enpc/). It automatically matches sentences, although it is possible (if necessary) to correct and expand the proposed correspondence manually. The program is language independent and simultaneously uses three main criteria: i) a list of anchor words; ii) proper nouns and iii) 'dice score' (cognates in McEnery and Wilson 2001). The alignment process yields two types of results: i) XML textual pairs containing both source and target texts and their identification tags, the function of which is to register and record equivalent relations across language and textual boundaries, and ii) a series of new TXT paired documents. These TXT documents are in turn converted into a HTML single bi-textual document. This third 'aligned' product is fed to the browser and is then ready to be searched in its usable format.

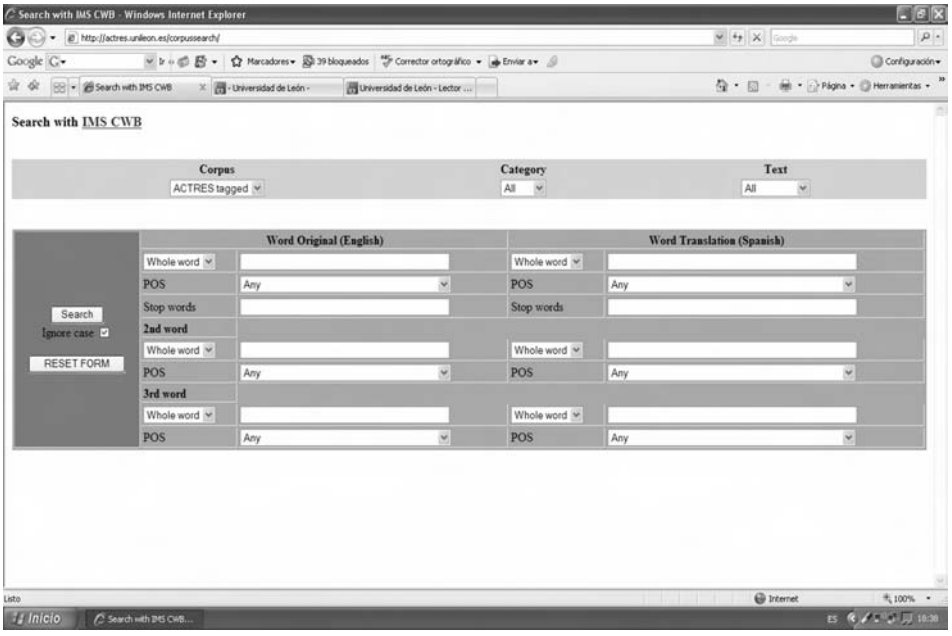

Fig 1. *Corpus Workbench interface.*

At this stage, P-ACTRES has acquired the added-value of an aligned, annotated, bilingual corpus. All the extra information that has been encoded in the corpus materials will allow for smooth, convenient querying strategies. In order to extract information from a corpus, a browser is of capital importance.

The choice for P-ACTRES has been the IMS *Corpus Workbench* (CWB) (http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/), a corpus query processor that is the final users' tool. As shown in fig. 1 the browser's interface offers two selection parameters, 'category' and 'text' that tell us the code tags for the subcorpora that can be browsed – e.g., 'F' for fiction, 'M' for miscellaneous - and for the aligned pair, a chosen part it belongs to - e. g., (FCJ1E.s573) + (FCJ1S.s557). It also contains 3 box sets allowing up to three word queries. Each set features a row for the item to be searched and another for the POS tag selection. Searches can be made in both directions, English ↔ Spanish.

The reasons why the TEI recommendations (mark-up), the Tree Tagger (annotation), the Corpus Translation Aligner (alignment) and the CWB (browser) have been the tools chosen for building P-ACTRES are quite straightforward: (1) they are adequate for our purposes, that is, useful, and (2) the technical expertise they require is accessible for different user types, including researchers, which means they are usable.

5. TERTIUM COMPARATIONIS: CROSS-LINGUISTIC LABELS

While computer tools allow us to organize raw materials in the corpus-building stage, it is important to explicitly recognize the instrumentality of both comparable and parallel corpora in the process and focus again on the research question(s). In order to find answers to those questions, the empirical information offered by the corpora must be collated profitably, and, to do so effectively, conceptual tools are necessary.

Any attempt at cross-linguistic analysis needs some criteria for comparison to weigh the extent of the similarities and the differences. This tool is generally referred to as *tertium comparationis* (Krzeszowski 1990: 15) and in this proposal it is a set of useful labels. Since their purpose is to help identify and represent features which are relevant for cross-linguistic purposes, they do not follow the usual conventions of descriptive linguistics, rather they show different statuses as they can mark grammatical, pragmatic, semantic and, even, interlanguage information (Chesterman 1998: 27-40). Furthermore, the labels and the terminology are tested for usability so that they can be accessible to final users as a basic tool in a number of applied activities, including revision.

As our corpora do not offer semantic tagging,[11] this part of the analysis cannot be solved (semi)automatically, and the meaning labels have to be assigned to each utterance individually, as shown below.

HABIT-IN-THE-PAST [HBP] incorporates the feature 'continuous, repeated action in the past', as in examples 1 and 2

From that day on, whenever Coward came to London, Greenwell would go round to the Savoy before the performance. [HBP_145]

Desarmaba literalmente las cuestiones que le eran planteadas, ***miraba*** *a su interlocutor con su característica mirada seria y escéptica y formulaba respuestas perfectas para la impresión*. [HBP_40]. [He would literally dissect the questions that were asked of him, stare at his listener with his skeptical serious eyes and give answers that could perfectly go straight into print].

6. USING STATISTICS: QUANTITATIVE AND QUALITATIVE ANALYSIS

In producing cross-linguistic analysis (objective) quantifiable and (intersubjective) qualitative data are required, and the need to interpret both types of findings is apparent.

Statistics are most useful at two stages in the experimentation process, when selecting raw data and when reporting results. The first is traditionally dealt with by means of random sampling; the second involves stating whether your results are statistically significant. In both situations, real differences need to be represented and distinguished from random variability. Statistical stringency serves in safeguarding against any potential mistake(s).

Sampling is necessary at both the corpus building and selection stages. Applying stratified random sampling according to needs is a safe way of 'avoiding the possibility of obtaining/generating a sample that does not include interesting rare items in the population' (McEnery, Xiao and Tono 2006: 21). In other words, we ensure that the sample is a perfect mirror of the distribution of the different classes (or strata) identified for the population. In our research these strata are the different meanings, functions or uses for which empirical evidence has been found.

Frequency information is another aspect where statistics have a role to play. Such information may be readily available from corpora, but this does not mean

11. There are some results in this direction, and a type of semantic tagging has been developed at Lancaster University (Piao, Ryson, Archer and McEnery 2005). This type of (necessarily) restrictive and highly formalized tagging does not seem to be particulary useful for our ultimate applied goal(s), although evidently it does have implications for future developments.

that being the most frequent (or typical) equals being the most important and that being less frequent justifies the exclusion from the study of certain phenomena. As noted by Kennedy (1998: 290) the value of frequency is to add one more criteria to those already being used in the research frame and its importance should not be overvalued.

While not all types of tests in statistics are useful or apply to all corpus-based findings, some stringency measure to collate the quantitative data is needed. It is precisely our concern with being as inclusive as possible that has influenced the type of tests used in this study. When analyzing data the goal is 'to arrive at the strongest possible conclusion from limited amounts of data' and to try to reach conclusions that extend beyond the sample we have analyzed. To this end, recourse to inferential statistics proves useful (Lowry 1999-2007).

Our test choice must be informed by the characteristics of our empirical data. In cross-linguistic work we very often deal with fairly large samples distributed in very irregular patterns. Furthermore, the data to be compared tend to come from different populations and different sampling processes, indicating that they qualify as independent proportions. Under these conditions, it is appropriate to use statistical 'hypothesis testing' ('hypothesis test for two independent proportions'). Two helpful indicators in our study are the *p-value* and the *z-test* (or *z-score*).

A result is said to be statistically significant when the p-value is less than a preset threshold value called α value. By convention, *p* (also known as 'observed significance level') tends to be set to 0.05 when the confidence interval is calculated for 95% confidence. In theory, confidence intervals can be computed for any degree of confidence and the α value will change accordingly; in practice it is traditionally almost always set to 0.05. Once the appropriate calculations have been made (Orris 2006), if the p-value is less than the threshold ($p < 0.05$), the difference is 'statistically significant'. If it is greater ($p > 0.05$), the difference is 'not statistically significant'.

Statistical significance is also sometimes misinterpreted as signifying an important result, but it only indicates whether the data show no difference between the 'known' (original, non-translated) and the 'new' (translated) data. In our applied cross-linguistic research the 'control' data are the original non-translated results obtained from the comparable data analysis and the 'new' data are the diagnostic, translated results. However, it should be kept in mind that in assessing the relevance of statistically significant results what is important is not the size of the significant results, but their effect and consequences on language use.

The z-test (or z-score) measures the difference between the data and what is expected under the null hypothesis (that translated and non-translated original

grammatical usage are identical, i.e., the samples or populations have the same mean). The z-score is compared to a Z table, which contains the percent of area under the normal curve between the mean and the z-score. For a significance level of 0.05 ($\alpha = 0.05$) and a level of confidence of 95% the normal curve happens between ±1.96. If the computed z-score belongs in this region, the null hypothesis is rejected and the result is therefore statistically significant.

Statistical significance obtained by these calculations is, then, another parameter to add to the researcher's toolkit. It can be particularly interesting when interpreting results and it provides a welcome link between quantifiable and qualitative empirical evidence as it helps to focus on those uses or functions that trigger cross-linguistic problems. Yet, quantitative data by themselves do not supply application-building information. Results have to be filtered and their representativeness and suitability for the purposes of the study qualitatively assessed. Feedback from prototypical users constitutes one of these filters.

7. REPRESENTATIVE USERS / INFORMANTS

As an additional tool, a group of informants can play a key role when analyzing for meaning, establishing the cross-linguistic labels and assessing usability. In the case of this proposal there are 10 informants. Their sociolinguistic profile can be defined as 'university educated speaker', 'middle class' and '25-50 years'. Five of them have some variety of English as their first language; the other five are native speakers of European Spanish. All the informants have had some training in linguistic analysis, although only two in each group are professional linguists. In each subgroup there is at least one person who cannot communicate both in English and in Spanish. The rest can at different levels of proficiency. Their professional profile is that of 'applied language professionals', including language services providers, EFL teachers, speech therapists, creative writers, translators, etc.

While strict descriptive linguists tend to prefer non-contaminated, monolingual speakers of the language(s), prototypical applied users were considered a better choice and more useful in this case because (a) they are the intended final users of our proposals, (b) they contribute to testing usability at each stage of the process and (c) they do not require particular training in order to perform as informants.

8. CONCLUSION

We have clarified the contribution and the suitability of each of the tools to the research process. Arguments supporting the convenience of using already available monolingual corpora have been given and the reasons for choosing the

BoE and CREA as 'source' corpora justified. P-ACTRES was born from the need to supplement comparable data with diagnostic data in cross-linguistic analyses. Parameters in DIY corpus building and the choices made for P-ACTRES on the basis of their suitability for the goals of this study have been discussed in detail. Qualitative data are analyzed by means of our conceptual tool, the cross-linguistic labels, which provide the organizing criteria for the English-Spanish contrast. The transition from raw quantitative data to useful information usable in our research frame is provided by statistical significance tests. These add rigor and help to ward off what is known as 'confirmation bias' on the part of the researcher (i.e., the tendency to search for interpretations that confirm his/her unverified view). Usability is warranted by the contribution of a 'control group' of representative users who act as informants.

While tools occupy the front stage in any empirical study, this fact may obscure the purpose of the research process. Therefore, a clear understanding of the research questions being addressed and the need to find solutions for them are essential. This involves a significant degree of flexibility on the part of the researcher when selecting, adopting and adapting useful and replicable procedures capable of functioning as an effective means to particular applied ends.

REFERENCES

Bank of English. http://www.collins.co.uk/books.aspx?group=153. Visited May 2007.

Bernardini, S. 2000. "Systematising serendipity: Proposals for concordancing large corpora with language learners." *Rethinking Language Pedagogy from a Corpus Perspective: Papers from the Third International Conference on Teaching and Language Corpora.* Eds. L. Burnard and T. McEnery. Frankfurt am Main: Peter Lang. 183-190.

Biber, D. 1993. "Representativeness in corpus design." *Literary and Linguistic Computing* 8 (4): 243-257.

Bowker, L. and J. Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora.* London: Routledge.

Byrne, J. 2006. *Technical Translation. Usability Strategies for Translating Technical Documentation.* Dordrecht: Springer.

Chesterman, A. 1998. *Contrastive Functional Analysis.* Amsterdam and Philadelphia: Benjamins.

Connor, U. and T. A. Upton. 2004. "The genre of grant proposals: A corpus linguistic analysis." *Discourse in the Professions: Perspectives from Corpus Linguistics.* Eds. U. Connor and T. A. Upton. Amsterdam and Philadelphia: Benjamins. 235-255.

Corpus de Referencia del Español Actual. http://corpus.rae.es/creanet.html. Visited July 2007.

English-Norwegian Parallel Corpus (ENPC). http://www.hf.uio.no/ilos/forskning/forskningsprosjekter/enpc/. Visited July 2007.

Flowerdew, L. 2001. "The exploitation of small learner corpora in EAP materials design." *Small Corpus Studies and ELT*. Eds. M. Ghadessy, A. Henry and R. L. Roseberry. Amsterdam and Philadelphia: Benjamins. 363-380.

Foraker Design. 2002-2005. Usability First. http://www.usabilityfirst.com/. Links to "Introduction to Usability." Visited December 2006.

Ghadessy, M., A. Henry, and R. L. Roseberry. 2001. *Small Corpus Studies and ELT*. Amsterdam: Benjamins.

Hofland, K. and S. Johansson. 1998. "The Translation Corpus Aligner: A program for automatic alignment of parallel texts." *Corpora and Cross-linguistic Research: Theory, Method, and Case Studie*. Eds. S. Johansson and S. Oksefjell. Amsterdam: Rodopi. 87-100. Also at http://khnt.hd.uib.no/files/align3.pdf. Visited December 2006.

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

Institut für Maschinelle Sprachverarbeitung (IMS). Visited July 2007. http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/.

Institut für Maschinelle Sprachverarbeitung (IMS). Visited July 2007. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.

Izquierdo, M., K. Hofland, and O. Reigem. Forthcoming. "The ACTRES Parallel Corpus: an English-Spanish Translation Corpus".

Kennedy, G. D. 1998. *An Introduction to Corpus Linguistics*. London and New York: Longman.

Krzeszowski, T. P. 1990. *Contrasting Languages. The Scope of Contrastive Linguistics*. Berlin and New York: Mouton de Gruyter.

Lowry, R. 1999-2007. *Concepts and Applications of Inferential Statistics*. http://faculty.vassar.edu/lowry/webtext.html/. Visited November 2006.

McEnery, T. and A. Wilson. 2001. *Corpus Linguistics. An Introduction*. Edinburgh: Edinburgh University Press.

McEnery, T., R. Xiao, and Y. Tono. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. London and New York: Routledge.

Orris, J. B. 2006. Megastat®. http://blue.butler.edu/~orris/megastat/index.html. Visited July 2007.

Piao, S. S., P. Rayson, D. Archer, and T. McEnery. 2005. "Comparing and combining a semantic tagger and a statistical tool for MWE extraction." *Computer Speech and Language* (Special issue on *Multiword Expressions*) 19 (4): 378-397. Elsevier. doi:10.1016/j.csl.2004.11.002.

Rabadán, R. 2005. "The Applicability of Description. Empirical research and translation tools." *Contemporary Problematics of Translation Studies*. Ed. C. Toledano. Special issue of *Revista Canaria de Estudios Ingleses* 51. 51-70.

Rabadán, R. 2007. "Divisions, description and applications: The interface between DTS, corpus-based research and contrastive analysis." *Doubts and Directions in Translation Studies. Selected contributions from the EST Congress, Lisbon 2004*. Eds. Y. Gambier, M. Shlesinger and R. Stolze. Amsterdam and Philadelphia: Benjamins. 237–252.

Rabadán, R. In press. "Refining the idea of "applied extensions"." *Beyond Descriptive Translation Studies. In Homage to Gideon Toury*. Eds. M. Shlesinger, D. Simeoni and A. Pym. Amsterdam and Philadelphia: Benjamins.

Sinclair, J. 2004. *How to Use Corpora in Language Teaching*. Amsterdam and Philadelphia: Benjamins.

Sperberg-McQueen, C. M. and L. Burnard. 2004. *Guidelines for Electronic Text Encoding and Interchange. XML-compatible edition*. http://www.tei-c.org/P4X/index.html. Visited July 2007.

SyncRO Soft Ltd. 2002-2007. http://www.oxygenxml.com/. Visited July 2007.

Toury, G. 1995. *Descriptive Translation Studies and Beyond*. Amsterdam and Philadelphia: Benjamins.

Zanettin, F. 2000. "Parallel corpora: Issues in corpus design and analysis." *Intercultural Faultlines*. Ed. M. Olohan. Manchester: St. Jerome. 105-118.