https://doi.org/10.11606/issn.2317-9511.v37i0p10-29

Corpora, translation, terminology… and beyond: objectives and perspectives

Corpora, tradução, terminologia e mais além: objetivos e perspectivas

Belinda Maia[1*]

[*1] Centro de Linguística da Universidade do Porto, Portugal. E-mail: bhsmaia@gmail.com

Abstract: This paper will not describe any specific research in corpus linguistics. Instead, it will first reflect on the way many of us teaching languages and translation in university departments develop and use corpora in our research and teaching methodology. One of the objectives is to highlight the work by Professor Stella Tagnin and those of us with whom she has worked over twenty years, even if it does not bring anything new to the immediate area. It will go on to analyze how, apart from the didactic uses of these resources, and related research, their potential for Natural Language Processing (NLP) became increasingly important, and demonstrate how the methodology of corpus linguistics is now used in various disciplines, especially in interdisciplinary research. This analysis was prompted by involvement in a project to advise universities in two Central Asian countries on the creation of a masters' degree in computational linguistics. The languages of these countries are very different from Western European languages, which obliged a re-assessment of my experience in linguistics and NLP in the context of English and Portuguese, when considering how the world's less-resourced languages could join the mainstream of computational linguistics.

Keywords: Corpus linguistics; Translation Technology; Natural Language Processing (NLP).

Resumo: A intenção deste artigo não é descrever investigação específica em linguística de corpus. Em vez disso, pretende ser uma reflexão sobre a maneira como muitos dos que ensinam línguas e tradução na universidade desenvolvem e utilizam corpora, tanto para investigação como como metodologia de ensino. Um dos objetivos é focar o trabalho da Professora Stella Tagnin e daqueles com quem ela trabalhou durante mais de vinte anos, mesmo que isso não traga nada de especialmente novo à área. Será depois analisado como, para além dos usos didáticos destes recursos, e da investigação que eles proporcionam, o seu potencial para o Processamento da Linguagem Natural (PLN) se tornou cada vez mais importante, e como a metodologia de linguística de corpus se aplica cada vez mais em várias outras disciplinas e especialmente em investigação interdisciplinar. Esta análise provém da minha participação num projeto europeu de aconselhamento a universidades de dois países da Ásia Central para a criação de um mestrado em linguística computacional. As línguas destes países são bem diferentes das línguas da Europa Ocidental, o que me obrigou a uma reavaliação da minha experiência em linguística e PLN no contexto do inglês e do português, num contexto de criação de recursos linguísticos para línguas menos conhecidas interessadas em se juntarem ao mundo da linguística computacional.

Palavras-chave: Linguística de corpus; Tecnologia de tradução; Processamento de linguagem natural (PLN).

Introduction

In the early 20th century 'linguistics' was generally seen as a sub-field of philosophy, and it was not interested in 'real' language, which was believed to be of social and anthropological interest. Later, Chomsky (1957) and his followers argued for decades that 'real' language was irrelevant to their objectives of exploring 'competence', rather than 'performance', and this position had considerable influence on research into language, especially in the US. However, the Systemic Functionalist school, led by M.A.K. Halliday (1973 and 1985), believed that not only should real language be central to linguistics, but it should also be studied within its social context.

The interest in using empirical means to establish facts about language started in the 60s with the Brown corpus (available from several websites), and developed steadily as the power of computers to record and analyze language grew exponentially, particularly as we reached the 90s. The reasons for collecting language data are many and varied, but the areas that interest us first here are the applications of using them for teaching language, translation, and terminology. I shall then refer to certain areas of interdisciplinary research that use the corpus linguistics methodology, and end by considering how linguistic and computational interests in corpora have diverged over the years, which becomes clear when working with lesser-resourced languages today.

1. English Language Teaching (ELT) and the need for contemporary language

The importance of the US in the world since WW2, coupled with the widespread use of English in the ex-colonies of the British Empire, combined to turn English into a world *lingua franca*. In 1980, Collins publishing teamed up with Birmingham University and started the COBUILD project, led by John Sinclair, with the objective of preparing corpora of contemporary language and, from observing it, developing dictionaries and grammars that reflected modern usage. The practical results of this project were publications that would support the growing English Language Teaching (ELT) industry, and other educational publishers soon followed suit. Teachers were thus supplied with a wealth of reference books and teaching material based on 'real' language, rather than having to rely on teaching material prescribed by older usage or norms.

The areas known variously as Computational Linguistics, Human Language Technologies, Language Engineering, Natural Language Processing, and under other designations that are hotly defended by those involved, probably welcomed the funding that became available, even if they often resented what they saw as the interference of linguists from the humanities.

The British National Corpus (BNC) was developed in the early 1990s and became available to researchers interested in the computational aspects, but several university teachers of language and linguistics became increasingly involved and people like Stella Tagnin began to see the possibilities for teaching language and preparing future teachers. She and others began to develop their own small corpora, and to present papers at corpus linguistics conferences.

The BNC, complete with part-of-speech annotation, was very useful for studying the finer points of language. However, the aim of collecting raw text (without annotation) was often to show students different text conventions, and to find information on a wide variety of subjects. Creating one's own corpus was a lengthy process in the 80s and 90s, and involved typing, scanning, or begging, borrowing, and even stealing texts, as by the mid-90s a lot of textual material was available on the Internet. When I first used the expression 'do-it-yourself corpora' in a paper (Maia, 1997) at the 1997 PALC – Practical Applications of Language Corpora (LEWANDOWSKA-TOMASZCZYK & MELIA EDS. 1997), there were mutterings from those involved in serious corpora compilation, and Krista Varantola (2003) advised me to use the expression 'disposable corpora', because of the problem of copyright.

2.Corpora for teaching translation

With the globalization of trade and commerce during the second half of the 20th century, the need for translation grew. Languages and translation had long been part of the curricula of polytechnics dedicated to producing office workers. However, as the translation market expanded, universities were encouraged to develop translation specializations, usually in the Modern Languages departments, where translation was seen as a technique for learning languages and understanding the literature in the foreign language. Translation theory became a respectable object of literary and cultural studies later, but, in the meantime, the language teachers, themselves largely trained in the humanist tradition, were expected to deal with the situation.

However, graduates soon discovered that even with good language skills, the ability to translate literature was poor preparation for earning one's living in a world in which institutional, technical, and scientific translations were in much greater demand. Would-be employers complained loudly that graduates in translation were useless, and preferred to use domain specialists with knowledge of languages. University language and translation teachers struggled to improve their programs, and soon discovered the Internet as a source for texts and information.

Various conferences were organized during the 90s that attracted a wide variety of interests across the language-translation-literature-culture spectrum, but also with a focus on the professionalization of translators. The European project was committed to plurilingualism, which provided considerable impetus, and the conference organizers often had connections to corpus linguistics projects. In 1990 and 1995 the Duo Colloquium Translation and Meaning conferences, organized by the universities of Lódz and Maastricht (THELEN & LEWANDOWSKA-TOMASZCZYK 1990 and 1996; LEWANDOWSKA-TOMASZCZYK & THELEN 1990 and 1996), showed the wide variety of approaches being considered. The 1997 PALC – Practical Applications of Language Corpora conference (LEWANDOWSKA & MELIA, 1997) showed, amongst other things, the connection between translation and corpora, and several participants met again at the CULT – Corpora and Learning to Translate conferences organized by Guy Aston in 1997 (BERNARDINI & ZANETTIN 2000) and 2000 (ZANETTIN ET AL. 2003).

The move to build corpora in other languages soon gained pace and work in Portuguese was boosted by Diana Santos and the Linguateca project, which started in 1998 with a view to providing the resulting corpora and other tools for public use online through a distributed language centre dedicated to developing resources for the computational processing of Portuguese. Its activities allowed Portuguese to become 'computer literate' early, and it continues today, offering, amongst much else, corpora presently covering over two billion words, all of them automatically annotated morphologically by Eckhard Bick's PALAVRAS (see VISL project). The project also developed the parallel corpus COMPARA/DISPARA of literary texts, and was involved in the parallel corpus Cor-Trad belonging to the COMET project that was presented at PALC 2001 by Stella Tagnin (2003).

3. Comparable corpora for terminology research

It soon became clear that training translators in terminology work was not as straightforward as traditional terminology approaches suggested. The strict guidelines followed by the International Standards Organization (ISO) and other bodies interested in providing terminology that was standardized and legally binding were difficult to apply in the fast-moving world of science and technology of today. Scientists and technicians needed to create terminology, often 'on the fly', and translators had to follow suit. Besides, scientific projects often competed to produce the definitive terms, and commercial companies defended terms that referred to processes and objects that they had registered for copyright.

The terminology found in the translated part of parallel corpora needed to be properly verified by a domain expert to be reliable, and it was difficult to get access to such material. Comparable corpora, broadly understood here as texts by domain experts in both of the languages, and more readily available, came to be a major source of terminology. Starting in 2002 as a Linguateca node at the University of Porto, we were able to develop Corpógrafo, an on-line environment for the construction and analysis of corpora, and the creation of terminology databases (MAIA ET AL. 2006).

Although obviously now in need of renewal, Corpógrafo has served ever since to train thousands of translators to collect special domain texts and to extract and evaluate terminology. Researchers work in their individual space, but should they need to publish the results, they will need to take care of the copyright of texts and terminology themselves. However, experience has taught us that terminology is valuable and not everyone wants to share it.

Everything referred to in the above sections has been reflected in the work of Stella Tagnin and like-minded professors and instructors over the years. This is evidenced by the articles in the editions of the journals she edited, Cadernos de Tradução (2002/1), CROP (2004, No.10), TradTerm No. 10, 2004), the many articles she has written over the years, as well as by many articles and books by others that have been published on these themes.

While linguists were developing ways of studying language and translation through corpora, the more traditional humanities disciplines had discovered the theoretical importance of translation, and its relationship with literary theory, multiculturalism and plurilingualism. This often led to further divisions between the literature and linguistics areas in the humanities. However, it became clear as the 21st century progressed that technology was quickly changing the dynamics of professional translation and terminology studies.

4. Technology for translation

While teachers of translation were adapting to the world of professional translation, technology was developing ways to accelerate the translation process. Software developers needed to ‘localize’ their products for other languages, which meant training translators to not only translate menus and instructions, but also to do it consistently by standardizing the language used. This need for standardization, together with the realization that previous translations of such language could be re-used to accelerate the process, were factors in the development of translation memories with integrated terminology databases for reference. Although several companies produce this type of software, TRADOS, now SDL-Trados, became the major player when it was adopted by the Directorate-General of Translation (DGT) at the European Commission.

For teaching purposes, there seemed, at first, and annotation apart, to be little difference between parallel corpora and translation memories (TMs), but it soon became clear that the translation companies that offered internships to translation students would refuse requests for academic use of the TMs and the related terminology. Apart from client confidentiality, TMs and terminology were valuable commercial assets.

The DGT, once it realized how translation technology could contribute to the acceleration of the ever growing mountain of translation required by European ideals, not only adapted accordingly, but also allowed access online to databases like Eurodicautom and others, now available through IATE. It came as somewhat of a surprise to find that these databases, despite all the effort that had gone into them, graded the entries according to their reliability and sometimes disappointed users. For example, translators into European Portuguese were not too happy to find that many 'Portuguese' entries had been made by Brazilians, and provided terms that were not accepted in Europe.

During the first decade of the 21st century, many European universities were encouraged to train their translators in both translation technology and a variety of computer related skills, and in 2008 the first EMT - European Master's in Translation Network was approved through the efforts of the DGT, and continues to flourish today. Every effort was also made to help the universities involved coordinate with translation companies all over Europe and what was fast becoming the Language Industry, now officially represented on the EC site as LIND. The site describes the language industry as comprising the activities of translation, interpreting, subtitling and dubbing, localization, language technology tools development, international conference organization, language teaching, and linguistic consultancy.

Some would argue that the DGT is overstepping its mandate by including the LIND page on its website and providing such a long list of activities. However, when the EMT network board consulted the companies providing translation and language services it became clear that skills in all these areas were increasingly required. Many such companies also require project management, but this is too wide an area to be included. Others would argue that the list ignores the 'elephant in the room' – machine translation (MT) – and the fact that many professional translators find themselves increasingly expected to post-edit MT.

5.Natural Language Processing for Human Language Technologies

The ultimate aim of much NLP is to provide artificial intelligence that communicates with its human in the same way as another human. Other researchers would settle for good machine translation (MT) and one must accept that it has made great strides in the last decade. However, very few members of the general public understand the NLP effort that goes into tools they take for granted, like spelling and grammar checkers, predictive writing, programs that read books for the blind or phone our friends in response to our speaking a name. Even such apparently simple tools are based on large quantities of language data.

By the time the world began to worry about the use being made of all the material on the Internet, and privacy became important, NLP had been quietly gathering vast quantities of material to build a variety of language resources. Although corpora builders carefully obtain copyright permission for every text, it should be clear that Google and others developed means of fuelling a variety of language tools from all the plentiful amounts of language material online.

Now, we cannot access the media or most websites without formally agreeing to ‘accept cookies’, so that they can supply us with advertising and, more sinisterly, find out more about us. However, for years the software we use every day for ‘free’ has also taken advantage of the way we use it to promote language technology. For instance, emails and tweets can give insights into new developments in communication and language use; the users of Skype, WhatsApp, and similar software are no doubt helping with speech recognition; ‘discussion groups’ like Quora are contributing to Q&A (Question and Answer) technology; and Google Translate uses monolingual, parallel and comparable corpora, together with all the morphological and syntactic information that is attached to them, as well as all the information available to improve its Statistical and now Neural MT. It would be impossible to close Pandora’s box now, even if there were a real interest in doing so.

The very large LREC - International Conference on Language Resources and Evaluation conferences - have moved increasingly towards the computer side of the spectrum. At the first conferences in 1998 and 2001, the linguistics professors with their young computer ‘geeks’ were in evidence; today the ‘geeks’ have become professors, linguists are far fewer, and attendees are largely interested in creating resources to produce the type of tools just described.

The Corpus Linguistics conferences, which have existed since the 1980s, are clearly flourishing. However, the focus is more on the applications of corpus linguistics research to the humanities than on technical development. The publishers Elsevier have just announced a new journal on Applied Corpus Linguistics, so Stella Tagnin and her followers can look forward to more developments in this area for some time to come.

6. Corpus linguistics – a multi-faceted area

If one is looking for conferences that are either wholly or partly interested in corpus linguistics, one will find that the main problem will be to decide which area of the application of corpus linguistics to choose. Apart from the already mentioned lexicology, translation, language teaching, natural language processing, and computational linguistics, and more general titles like theoretical and applied linguistics, or systemic functional linguistics, one can attend conferences that associate corpus linguistics with sociolinguistics, psycholinguistics, cognition/cognitive linguistics, semantic prosody, discourse prosody, pragmatics, contrastive linguistics and literature.

No doubt the new Elsevier journal on Applied Corpus Linguistics hopes to encourage and take advantage of an approach with an emphasis on providing quantitative analysis to support opinions or theses. Too often does one read an otherwise interesting article or dissertation that, after a thorough presentation of the theoretical background, puts forward a possibly valid opinion, but fails to support it with more than a small number of examples. Conversely, of course, there are corpus linguists who produce numbers and graphs and then leave it to the reader to draw conclusions from them. However, the emphasis of corpus linguistics is usually, and should be, on providing qualitative data to support a thesis or opinion, or draw attention to some interesting phenomenon.

Large corpora – never large enough for some – can offer a broad analysis of language, usually related to lexical usage, the objective of lexicographers, whether the emphasis is historical, like the Oxford English Dictionary, or contemporary, as described above. Older and even recent changes in grammatical usage can be traced using corpora, as in (Leech et al, 2009). There are corpora of different varieties of English and of Portuguese that have existed for some time and have been made available by Mark Davies at English-corpora.org and https://www.corpusdoportugues.org/. Similar large corpora are also available in other languages.

Researchers in translation love to highlight linguistic and cultural differences when the original and translation are compared, but one needs large comparable corpora to advance beyond anecdotal descriptions. The translation of words, whether from general language or associated to special domains – or terms -, are the basis of discussion in many books and articles. But the situation is more complex when we are comparing areas considered universal to human experience, but which different languages express differently.

One of the popular areas of NLP at present is sentiment analysis, an area that the humanities may consider their specialty, but which is more often funded by those seeking to discover consumer tastes or political opinions, and even by those searching to police the more sinister users of the Net. Subjectivity is inherent to most human communication – even in the order in which 'facts' are presented in different news channels – and, although the sentiment lexicons available will help, they cannot, at least computationally, and thus quantitatively, deal easily with irony, sarcasm, and the cultural norms of the group being studied. Besides, many of the subjective elements are also expressed through structures usually classified as syntactic, like the difference between the use of the subjunctive in Portuguese and other languages, and the use of the auxiliary verbs in English. And, of course, there are the cultural differences to be found in various styles of discourse.

Global culture is affecting many forms of discourse. English is not only becoming the *lingua franca* of scientific and technical information; Anglo-American norms are also governing the structure of scientific discourse. An area of research that combines much of what has already been said on translation and terminology research is applicable to legal language, and we can see the effects every day in legal documents translated from Anglo-American versions. Legal terminology, however, is only part of the problem; legal discourse is probably even more difficult to analyze. The Anglo-American legal tradition is based on case law, or developed over the years from specific cases, but most of the rest of the world follow the traditions of civil law, based on formal statutes and legislation. This fact has important implications for EU law and partly explains the Brexiteers' rejection of it.

Deborah Cao (1996: 662) wrote that one of the socio-cultural difficulties in writing contracts between English and Chinese companies was that while English common law expects parties 'to commit themselves to what is relevant to the business transaction and what can actually be achieved …', the 'Chinese often regard contracts as statements of good intention and believe that the parties to a contract can work out the details … as needs arise’. Such cultural misunderstanding can have far-reaching consequences. However, in order to study such phenomena in any depth, one needs expert knowledge of both legal systems and sizeable corpora of comparable and parallel/translated texts from which to draw examples.

If cultures differ in their language use and habits, so do individuals. Professor Higgins in ‘My Fair Lady’ claimed he could tell where someone came from in England just by listening to them speak; today there is technology that serves to identify individuals by their speech, as well as helping us to dictate to our phones and computers. Each of us would seem to have an idiolect, which is the result of the language input we have received over the years, as Coulthard & Johnson (2007, Chapter 8) explain. The features of each person’s idiolect can help experts to prove the identity of the writer or speaker of texts. Grant (2010: 508-522), for example, describes a case when corpora of text messages were used to identify a murderer, by comparing a corpus of the suspect’s messages to a corpus of the victim’s messages, which were then compared to a control corpus of messages by others.

Authorship attribution and plagiarism detection have preoccupied researchers of literature and teachers and university professors for many years. Forensic linguists propose a variety of techniques from corpus linguistics to make such research more reliable (COULTHARD ET AL. 2010: 523-538), and Woolls (2010: 576-590) describes several techniques used by computational and corpus linguists.

I mention the areas of research above because of my personal contact with and interest in them, but no doubt others would point to examples in other areas where corpora and corpus linguistics methodology have proved equally illuminating.

7. Language resources for lesser-known languages

My reflections so far are partly the result of hindsight and the ability to see the bigger picture as one retires from mainstream teaching and research. However, it was also prompted by a project in which I am involved and which obliged me to reassess the experience of many years in these areas.

In October 2017 the project CLASS: Interdisciplinary Master Program on Computational Linguistics at Central Asian Universities began, and, for reasons best known to the organizers, I found myself involved as a member of the team from the University of Porto. The other European partners are from the Universities of Santiago de Compostela (also responsible for administration) and A Coruna in Spain, Poznán in Poland, and West Attica in Greece. Our role was to advise several universities in Kazakhstan and Uzbekistan in Central Asia (CA) on the creation of a new Masters' degree in Computational Linguistics.

It soon became clear that the CA team consisted almost entirely of computer scientists, although one university included people from the humanities. The divergence of the computational and linguistic interests reflected in the different developments of the LREC and Corpus Linguistics conferences described above no doubt played a part in the choice of the team.

Although most academics in the ex-Soviet countries communicate in Russian, there is clearly a movement to give more prominence to the languages in these countries, most of which belong to the Turkic language group and share similar linguistic features. The evidence for this are the annual TurkLang conferences that have taken place since 2013, and at which I had the honor to present a paper in 2018. An analysis of the online proceedings (with the help of Google translate!) shows that while the computational ambitions are considerable, the realization that considerable language resources are needed to further them has only recently begun to produce results.

Creating resources in Western European languages may have appeared difficult to those of us who tried to establish rules for collecting, annotating, and analyzing our corpora: let us remember, for example, how Portuguese verb forms require far more attention than English ones, not to mention the fact that all Portuguese nouns require a gender, and English auxiliary verbs do not correspond to Portuguese usage.

The Turkic languages are agglutinative, which requires a different approach that separates the main lexical item from the various affixes that can be added before and after it, and then writing syntactic rules as to how everything is combined into words and sentences. This would seem to affect attitudes to lexicography, and the response to requests for information on dictionaries and thesauri surprised us: for 'dictionaries' we received a list of bilingual Russian – Kazakh/Uzbek special domain dictionaries, where apparently the influence of Russian is considerable; 'thesauri' appeared to be more similar to, if not the same as, our own monolingual, alphabetically ordered dictionaries (but beware – the word 'thesaurus' seems to have a complex history).

If we add to this the fact that different scripts and alphabets have been used in these languages over the years – Persian, Arabic, Latin, Cyrillic, and that now the objective is to change to Latin again, one can see the situation is complex. For example, Cyrillic allows for at least ten more letters than the Latin alphabet and, as the written form of these languages reflects pronunciation, this may present problems. Although our CA colleagues assure us that converting Cyrillic to Latin is easily computerized, it would seem to be a far from trivial task and appears to have led to political and academic arguments that are easy to find online.

The EU team members have considerable experience of producing language resources, even if their theoretical approaches vary. But perhaps these differences actually help us all to work from different view-points to provide advice that may be adapted to suit the specific problems of the CA universities and the languages they propose to bring out of computational obscurity. The CA universities are in many ways luckier than the researchers working with Western European languages over the last thirty or forty years. For a start they can count on much more powerful hardware and software, and they can learn from the hard won experience of those who have gone before them. Although academic rivalry exists everywhere, there is much to be gained by working cooperatively with colleagues working with other Turkic languages to establish the most appropriate linguistic theory and methodology. They can also count on machine learning to accelerate the compilation of language resources. We can only wish them the best of luck.

Conclusions

The challenge to write this article allowed the re-evaluation of the work Stella Tagnin and others, like myself, carried out over the years using corpora in language, translation, and terminology teaching. It also allowed a reflection on how academic areas work together, as when NLP researchers worked with general linguists to produce corpora, and how this work gradually developed and diverged to create different areas and sub-areas, sometimes working together, and at other times in parallel. These developments show that any area of study dealing with language or languages is attracting a growing number of researchers.

Translation is no longer restricted to language teaching or a perceived choice between interesting literary texts and boring technical ones. It is done for a wide variety of reasons and deals with the many varieties of 'text' that are used today for communication. Not everyone accepts the role of technology, but few would argue against the statement that it helps to accelerate the (re)translating of repetitive (boring?) texts. Terminology research has always provided the opportunity to learn about new areas of knowledge and can be truly rewarding, as many translation students have discovered over the years.

There are several large corpora projects and each has a different aim and approach. Mark Davies of Brigham Young University offers billions of words in English and in Portuguese for language study; Linguateca, as mentioned above, supplies large quantities of Portuguese; Sketch Engine, aimed at students and researchers of language and linguistics, offers corpora and other tools for many languages; Jörg Tiedemann's OPUS corpora offer enormous quantities of parallel corpora in many languages, largely for machine translation purposes; and one could add several other sites of interest. The fact that there are so many, and the interest in them is so varied, testifies to the richness and variety of the area. It would appear that corpus linguistics methodology has much to offer other areas, especially in interdisciplinary projects that include language and languages.

The opportunity to work on a project that will need to produce language resources in languages outside my own experience encouraged this evaluation and opened up new perspectives. It has been a privilege to be involved. However, it leads me to focus one further point that needs to be made: as the only native speaker of English on a project in which everyone is meant to speak and write English to communicate, I have been yet again made aware of the dominance of the 'killer' language. A *lingua franca* has many uses, but translation between all languages, and work to provide understandable terminology in all fields in those languages, are of paramount importance if we value our languages and cultures.

References

BERNARDINI, S.; ZANETTIN, F. *I Corpora nella didattica della traduzione - Corpus Use and Learning to Translate*. University of Bologna: 2000.

CAO, D. Consideration in Translating English/Chinese Contracts. *Meta*, v. 42, n. 4, 1997, p. 661-669. https://doi.org/10.7202/002199ar

CHOMSKY, N. *Syntactic Structures*. Paris: Mouton; Co. 1957.

COULTHARD, M.; JOHNSON, A. *An Introduction to Forensic Linguistics - Language as Evidence*. London: Routledge. 2007.

COULTHARD, M.; JOHNSON, A. (Ed.) *The Routledge Handbook of Forensic Linguistics*. London: Routledge. 2010.

HALLIDAY, M. A. K. *Explorations in the Functions of Language*. London: Edward Arnold. 1973.

HALLIDAY, M. A. K. *An Introduction to Functional Grammar. 2nd Edition.* London: Edward Arnold. 1985.

LEECH, G.; HUNDT, M.; MAIR, C.; SMITH, N. *Change in Contemporary English – a Grammatical Study.* Cambridge University Press. 2009.

LEWANDOWSKA-TOMASZCZYK, B. (Ed.) *PALC 2001: Practical Applications in Language Corpora.* Frankfurt: Peter Lang. 2003.

LEWANDOWSKA-TOMASZCZYK, B.; THELEN, M. (Ed.). *Translation and Meaning, Part 2. Proceedings of the Lódz Session of the 1990 Duo Colloquium on 'Translation and Meaning, held in Lódz, Poland, 20-22 September, 1990.* Maastricht: Euroterm. 1990.

LEWANDOWSKA-TOMASZCZYK, B.; MELIA, P. J. (Ed.) *Proceedings of Practical Applications of Language Corpora.* University of Lodz Press. 1997.

MAIA, B. Do-it-yourself corpora… with a little bit of help from your friends! In: LEWANDOWSKA-TOMASZCZYK, B.; MELIA, P. J. pp. 403-410. 1997.

MAIA, B. Training Translators in Terminology and Information Retrieval using Comparable and Parallel Corpora. In: ZANETTIN ET AL. pp. 43-54. 2003.

MAIA, B.; SARMENTO, L.; SANTOS, D.; CABRAL, L.; PINTO, A. S. The Corpógrafo - a Web-based environment for corpus research. Proceedings from the Corpus Linguistics 2005 *Conference Series; Corpus Linguistics Conference* (Birmingham, UK, 14-17 July 2005), s/pp. ISSN: 1747-9398

SANTOS, D. Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties. *OSLa: Oslo Studies in Language*, v. 3, n. 2, 2011, pp. 113-128. At https://www.linguateca.pt/Diana/download/SantosOSLa2010.pdf

SIMÕES, A.; BARREIRO, A.; SANTOS, D.; SOUSA-SILVA, R.; TAGNIN, S. E. O. *Linguística, Informática e Tradução – mundos que se cruzam.* Oslo Studies in Language, v. 7, n. 1, 2015. https://journals.uio.no/osla/issue/view/100

TAGNIN, S. COMET – a multilingual corpus for teaching and translation. In: LEWANDOWSKA-TOMASZCZYK, B. (Ed.) pp. 535-540. 2003.

TAGNIN, S. Ed. *Cadernos de Tradução – Tradução e Corpora.* v. 1, n. 9. Universidade de Santa Catarina, 2002.

TAGNIN, S. (Guest editor). *CROP – vol. 10.* São Paulo: FFLCH-USP, 2010.

TAGNIN, S. (Ed.) *Tradterm*, n. 10, 2004. http://www.revistas.usp.br/tradterm/issue/view/3912

TAGNIN, S.; TEIXEIRA, E. Lingüística de Corpus e Tradução Técnica - Relato da montagem de um corpus multivarietal de culinária. In: *Tradterm*, n. 10, dec. 2004, p. 313-358. https://doi.org/10.11606/issn.2317-9511.tradterm.2004.47184

TAGNIN, S. Corpus driven glossaries in translator training courses. In: SIMÕES ET AL. pp. 359-377. 2015

THELEN, M.; LEWANDOWSKA-TOMASZCZYK, B. (Ed.). *Translation and Meaning, Part 1. Proceedings of the Maastricht Session of the 1990 Duo Colloquium on 'Translation and Meaning, held in Maastrict, The Netherlands, 4-6 January 1990.* Maastricht: Euroterm. 1990.

THELEN, M.; LEWANDOWSKA-TOMASZCZYK, B. (Ed.). *Translation and Meaning, Part 3. Proceedings of the Maastricht Session of the 2nd International Maastricht~Lódz Duo Colloquium on 'Translation and Meaning, held in Maastrict, The Netherlands, §9-22 April, 1995.* Maastricht: University of Maastrict. 1995.

VARANTOLA, K. Translators and Disposable Corpora. In: ZANETTIN ET AL. pp. 55-70. 2003

ZANETTIN, F.; BERNARDINI, S.; STEWART, D. *Corpora in Translator Education.* (Ed.) Manchester: St. Jerome Pub. Co. 2003.

Internet references – all sites last accessed May 2020.

British National Corpus (BNC) Official site - http://www.natcorp.ox.ac.uk Also consultable at: https://www.english-corpora.org/bnc/ & http://corpora.lancs.ac.uk/bnc2014/

COBUILD project - https://www.collinsdictionary.com/cobuild/

CLASS: Interdisciplinary Master Program on Computational Linguistics at Central Asian Universities – http://erasmus-class.eu

CoMET – Corpus Multilingue para Ensino e Tradução – http://comet.fflch.usp.br/corporamultilingue

COMPARA/DISPARA – online parallel corpus of Portuguese/English literary texts. Part of the Linguateca project. https://www.linguateca.pt/COMPARA/dispara.php?language=en

CORPÓGRAFO – a set of online tools for creating corpora and terminology databases. Part of the Linguateca project. https://www.linguateca.pt/corpografo/

Directorate General of Translation of the European Commission http://cdt.europa.eu/en/partners/european-commission-directorate-general-translation

Eckhard Bick –VISL project – s research and development project at the Institute of Language and Communication at the University of Southern Denmark. https://visl.sdu.dk

Elsevier Journals – Applied Corpus Linguistics - https://www.journals.elsevier.com/applied-corpus-linguistics

European Language Industry platform – LIND https://ec.europa.eu/info/departments/translation/language-industry-platform-lind_pt

Google Translate - https://translate.google.com

IATE - Interactive Terminology for Europe https://iate.europa.eu/home

OPUS – open-source parallel corpus – compiled and organized by Jorg Tiedemann http://opus.nlpl.eu

Linguateca – a distributed language resource centre for Portuguese - https://www.linguateca.pt

LREC - International Conference on Language Resources and Evaluation - http://www.lrec-conf.org

Mark Davies' corpora project, Brigham Young University - https://corpus.byu.edu/overview.asp

Mark Davies' English corpora at https://www.english-corpora.org

Mark Davies' Portuguese corpora at https://www.corpusdoportugues.org/

Quora – a Question and Answer platform that invites one to participate in debates https://pt.quora.com

SDL-Trados – well-know translation technology software https://www.sdltrados.com

Skype - https://www.skype.com/en/

TurkLang conferences – conferences dedicated to the computational study of the Turkic languages – http://www.turklang.net/en

WhatsApp - https://www.whatsapp.com

Recebido em: 06/08/2020

Publicado em janeiro de 2021