Aproximação à linguística de corpus como metodologia de base empírica. Compilação e anotação do Corpus Paralelo PALOP (português-espanhol) de Narrativa Pós-colonial.

Paulo Malvar Fenández
paulomal@usc.es

Resumo:

A ideia de levar a cabo, desde a Linguística, um trabalho de investigação debruçado sobre o campo das Literaturas Africanas de Língua Portuguesa nasceu durante a assistência ao congresso *Cinco Povos Cinco Nações* celebrado na Universidade de Coimbra entre os dias 8 e 11 de Outubro de 2003 (actas sem publicar). Neste congresso, em concreto na mesa de debate à volta das relações entre as "Literaturas africanas e outras literaturas", na qual intervieram o professor de Tradução da Universidade de Valhadolide Joaquín Garcia-Medall com a comunicação intitulada "Breve panorama de la narrativa lusoafricana en sus traducciones al español," e o tradutor profissional José Gómez, com a comunicação "Traduzir em castelhano: dificuldade e gratificação," foram expostas as dificuldades actualmente existentes no campo da tradução para castelhano de obras provenientes das referidas literaturas. Deste jeito, além da inexistência de suficientes profissionais da tradução especialmente treinados no campo destas literaturas, na citada mesa de debate foram também indicadas, por parte de ambos os profissionais da tradução, importantes dificuldades referentes à inexistência de materiais de ajuda à tradução especialmente desenhados para se poderem confrontar com as particularidades –linguísticas e culturais– específicas dos textos deste tipo de literaturas.

Tendo em conta, pois, a realidade desenhada por estes dois profissionais da tradução, a ideia inicial foi a de levar construir um corpus paralelo (português-espanhol) de textos narrativos (romances e livros de contos) provenientes das Literaturas Africanas de Língua Portuguesa.

Neste sentido, no primeiro apartado deste trabalho far-se-á uma revisão histórica das metodologias de estudo empregadas no seio da Linguística. Esta revisão histórica servirá de base para a elaboração nos dois seguintes apartados de uma revisão do campo da chamada Linguística de Corpus, do conceito de corpus e da tipologia de corpora. No quinto apartado será feita, após uma revisão teórica do *status*, da história e da finalidade dos corpora paralelos, uma descrição em detalhe das diferentes questões tidas em conta e das diversas tarefas realizadas para a compilação e anotação do corpus paralelo criado no desenvolvimento deste projecto de investigação, assim como uma aproximação à sua actual aplicabilidade nos campos da investigação, didáctica e exercício profissional da tradução humana.

Palavras-chave: Empirismo, Linguística de Corpus, Literaturas Africanas de Língua Portuguesa, Tradução Português-Espanhol

Abstract:

The idea of carrying out a research from a Linguistics point of view in the field of Portuguese Language African Literatures was born after having attended the conference *Cinco Povos Cinco Nações*, celebrated on October 8th-11th, 2003 at Universidade de Coimbra (non-published porceedings). At this conference, in particular at the round table called "Literaturas africanas e outras literaturas", in which Professor Joaquín Garcia-Medall (Universidad de Valladolid) gave a paper called "Breve panorama de la narrativa lusoafricana en sus traducciones al español," and the professional translator José Gómez gave another paper called "Traduzir em castelhano: dificuldade e gratificação," were exposed the most important difficulties for translating those literatures novels into Spanish. It was argued that there are not enough profissional translators specialized in these literatures, as well as there are not translation aid materials designed to allow the confrontation of the cultural and linguistic specificities of texts from these literatures with the same Spanish specifities.

Drawing from the frame designed by these two professional translators, the initial idea was to buid a parallel corpus (Portuguese-Spanish) of Portuguese Language African Literatures texts of narrative (novels and books of tales)

Thus, in the first section of this paper I will carry out a historical revision of the methodologies of reseach used in Linguistics. This will allow me to elaborate in the next two sections a revision of the field known as Corpus Linguistics, the concept of corpus itself and the tipology of corpora. In the fifth section, after a theoretical revision of the *status*, history and *raison d'être* of parallel corpora, I will

develope a detailed description of the different issues and tasks that were taken into account and/or performed during the compilation and tagging processes. Finally, I will show, using some examples, how this corpus is useful for human translators not matter they want to do research on, to teach translation strategies or to professionally translate texts from those literatures into Spanish.

Key words: Empirism, Corpus Linguistics, Portuguese Language African Literatures, Translation Portuguese-Spanish

1. Introdução

Como primeiro passo na elaboração deste trabalho e para que possa ser entendido e valorizado de acordo com o alcance do estudo que nele se pretende realizar, têm de ser, antes de mais, feitas certas considerações a respeito da história da Linguística e das diversas opções metodológicas empregadas na investigação linguística, desde os seus inícios até à actualidade. O que será feito, pois, nesta breve introdução histórica é ressaltar tanto aqueles momentos históricos que, tanto pelo tipo de investigações realizadas quanto pelas mudanças contextuais acontecidas, resultaram ser definitivos para que trabalhos como o presente possam ser hoje em dia elaborados.

1.1. Um bocado de História

Antes do surgimento da Linguística no século XIX como campo de estudo diferenciado, a quase totalidade dos estudos realizados em Ocidente que se debruçaram sobre o estudo das línguas foram estudos gramaticais de carácter prescritivo, que empregaram, mercê à sua filiação com a tradição gramatical greco-latina, uma metodologia baseada na obtenção de dados tirados do próprio conhecimento dos estudiosos como falantes das línguas estudadas. Não foi, desta maneira, até ao século XVIII que, graças ao importante papel desempenhado por multidão de investigadores inseridos dentro das conhecidas como ciências naturais, nas quais o modelo filosófico maioritário era e é o empirismo, começaram a aparecer timidamente alguns trabalhos no âmbito da Lexicografia que seguiram, embora só parcialmente, uma das premissas formuladas dentro das ciências duras para ser seguidas em todo trabalho que se pretendesse científico: a observação da realidade.

Durante o século XIX, século em que surgiram diferentes escolas como a Histórico-comparativista, primeiro, e a Neo-gramática, com posterioridade, que dominaram o campo das investigações linguísticas, a quase totalidade das suas investigações debruçaram-se sobre o estudo dos estádios pretéritos das línguas europeias. Desta forma, embora sob perspectivas distintas e com finalidades bem diferentes, ambas as escolas se encontraram com o problema da impossibilidade de acesso aos dados que pretendiam investigar através de processo nenhum de introspecção. Deste jeito, foram os próprios testemunhos antigos em forma de textos escritos que tiveram que ser empregados. Além destas escolas, a finais do século XIX foi também tomando corpo uma disciplina linguística, a Dialectologia, que desde os seus inícios com os primeiros trabalhos levados a cabo por Áscoli, teve, igualmente, de empregar dados empíricos, fruto nalguns casos da observação directa e, noutros, do desenho e aplicação de formulários por parte dos/as investigadores/as. Em qualquer caso, o desconhecimento por parte dos/as investigadores/as a respeito do objecto de estudo escolhido em todas estas investigações determinava, de facto, que a metodologia escolhida não pudesse ser outra que a observação de usos linguísticos reais.

Apesar destes inícios metodologicamente empiristas da Linguística no século XIX, a publicação em 1916 do *Course de Linguistique Générale* de Ferdinand de Saussure, que sentou umas bases epistemológicas firmes para consideração da

Linguística como uma disciplina científica, pois foi o primeiro em explicitar o seu objecto de estudo, a metodologia escolhida para estudá-lo e a finalidade da própria investigação linguística[1], trouxe consigo uma mudança no que à natureza dos dados empregados nas investigações linguísticas se refere, reflectida na distinção por ele estabelecida entre *Langue* e *Parole*. Esta distinção, que tinha por base a consideração das línguas como sistemas homogéneos, socialmente compartilhados, em que cada uma das suas unidades se define pelas oposições que estas estabelecem entre si, implicava, assim mesmo, uma priorização do trabalho linguístico para o estudo daquelas unidades pertencentes ao nível sistémico e abstracto da *Langue*, rejeitando, deste jeito, qualquer atenção às manifestações concretas das línguas (*Parole*), por se considerarem heterogéneas e individualmente condicionadas, isto é, totalmente assistémicas, e, portanto, inúteis e secundárias para a consecução de um conhecimento objectivo da linguagem. Em palavras de Saussure (1916: 29-30): "Al separar la lengua del habla (*langue et parole*), se separa a la vez: 1º, lo que es social de lo que es individual; 2º, lo que es esencial de lo que es accesorio y más o menos accidental."

Desde esta concepção racionalista, a rejeição a respeito da valide dos dados empíricos como base para a observação e análise da linguagem, entendida esta como um fenómeno humano de natureza complexa, fez, em primeiro lugar, com que, quase de maneira generalizada, o trabalho do/a investigador/a, na maior parte dos casos falante das línguas objecto de estudo, se convertesse num trabalho de introspecção e reflexão desde o qual elaborar, mediante uma metodologia dedutiva, qualquer tipo de hipótese e/ou teoria; e, em segundo lugar, com que o/a próprio/a investigador/a se tornasse, assim, em fornecedor dos exemplos necessários para fundamentar ditas hipóteses e/ou teorias.

Deste jeito, na Europa, sendo a concepção racionalista-dedutiva aquela que principalmente dirigia os estudos linguísticos levados a cabos desde o Estruturalismo europeu com o objecto de estudo da função das diferentes unidades de *Langue*, o recurso ao emprego de dados empíricos, embora constante, passou a ser marginal na maioria disciplinas linguísticas cultivadas, sobretudo, no Círculo Linguístico de Copenhaga, mas também no Círculo Linguístico de Praga, desde os quais se chegou, desde as posições mais ortodoxas, a negar, incluso, a validade da *Parole* para a verificação da Teoria Linguística: "linguistic theory cannot be verified (confirmed or invalidated) by reference to any texts and languages" (Hjelmslev, 1969 *apud* Beaugrande, 1999: 244).

Cabe, porém, matizar que, embora baixo esta concepção sistémica das línguas, na Europa, por um lado, existiram durante as décadas de 20, 30, 40 e 50 certas disciplinas como a Dialectologia ou como a História da Língua que continuaram, no desenvolvimento das suas investigações, com a tradição metodológica empirista dos seus inícios; e existiram, por outro, desde os anos 30 alguns linguistas filiados ao Círculo Linguístico de Praga que

> undertook quantitative studies (mainly of Czech, English and Russian) of frequency of certain grammatical processes, the relatives frequencies of different parts of speech, the location and distribution of information in the sentence, and the statistical distribution of syllable types and structures. (Kennedy, 1998: 10)

Nos EE.UU., porém, onde a pegada de Saussure não tinha sido tão funda, desenvolveu-se durante a década de 30 uma escola estruturalista diferente das europeias, chamada Distribucionalismo, desde a qual, baixo a liderança, sobretudo, de Bloomfield, as investigações realizadas viraram-se para o estudo da "repartição dos elementos […] e a sua capacidade de associação e de substituição" (Kristeva, 1988: 241). Tal e como na

[1] Premissas estabelecidas por Fernández Pérez (1986: cap. 1) para o seguimento e verificação da validade de qualquer corpo teórico, aspectos essenciais para o reconhecimento de toda a disciplina ou estudo como científico.

altura puderam comprovar os/as investigadores/as filiados/as a esta escola, um estudo deste tipo precisava do emprego de uma metodologia estatística, que obrigava ao manejo de grandes quantidades de dados, necessariamente empíricos, já que a sua recompilação, assim como o seu processamento resultavam totalmente impossíveis para os/as investigadores/as, mercê à sua incapacidade fisiológica e cognitiva para realizarem este tipo de buscas e operações na sua própria memória. Esta viragem no objecto de estudo da linguística estruturalista americana permitiu que, a partir dos inícios da década de 1930, a vertente quantitativo-estatística da Linguística Matemática –nascida "por razões técnicas: a construção de computadores destinados a ler e a escrever ou de máquinas destinadas à tradução automática" (Kristeva, 1988: 253)–, se tornasse autónoma com a importante achega feita, sobretudo, pelo filólogo americano George Zipf, cujo trabalho

> was concerned with such quantitative analyses as the relation between the frequency of words in text length, the frequency of words and their antiquity, and the relation between the rank order of an item in a word frequency list and the number of occurrences or tokens of that item in a text. (Kennedy, 1998: 10)

Além disso, desde esta escola estruturalista foi, também, desenvolvido, sobre todo nas décadas de 30 e 40, um importante trabalho no que ao ensino de línguas se refere, que ficou plasmado na teoria pedagógica hoje em dia conhecida como Condutismo. Esta teoria, baseada no emprego de colecções de textos para a elaboração tanto do seu material didáctico como para o desenvolvimento efectivo das tarefas docentes, pretendia o aperfeiçoamento das habilidades linguísticas dos discentes mediante o estudo repetitivo das estruturas das línguas estudadas nos diferentes textos empregados.

Foi, assim, nas décadas de 40 e 50 que se deram os passos definitivos para a formação e consolidação da, actualmente, conhecida como Linguística de Corpus. Por um lado, nos EE.UU. e baixo a influência da tradição linguística pós-Bloomfieldiana, foram realizados diversos trabalhos e investigações, como os de Fries (1952) em sintaxe ou Lorge (1949) em semântica, que, na elaboração dos seus trabalhos, partiram do tratamento quantitativo de grandes quantidades de dados linguísticos como única base para o estudo das diferentes unidades das línguas e a elaboração das suas gramáticas. Por outro lado, surgiram nesta altura os computadores (*computers* em inglês), cuja função primária era 'contar', tal e como a etimologia do seu nome nos sugere ("[radical do particípio] *computado* (do v. *computar*) + *-or*; calcado no ing. c*omputer* […]" (Houaiss, 2001)), e que começaram a ser aplicados como ferramentas para o estudo quantitativo das línguas, ainda que de maneira minoritária, devido tanto ao seu quase inacessível preço como às suas consideráveis dimensões, não acordes, por outro lado, com a sua baixa velocidade e limitada capacidade de armazenagem

Porém, uma nova tendência surgida também nos EE.UU. sob a influência pós-Bloomfieldeana que, em origem, igualmente tinha botado mão dos corpora como recurso para a análise formal das línguas, passou nos finais dos anos 50 a ser liderada por um jovem linguista, Noam Chomsky, quem readaptando a dicotomia saussureana *Langue* e *Parole*, formulou como princípio do estudo generativo a dicotomia entre *Competence*, "capacidade do sujeito falante de formar e de reconhecer frases gramaticais" (Kristeva, 1988: 258) e *Performance*, "realização concreta dessa capacidade" (íbid.). Em base a esta dicotomia, Chomsky, delimitou, desde uma perspectiva "mentalista" com raízes cartesianas idealistas, que

> Lo que concierne primariamente a la teoría Lingüística es un hablante-oyente ideal, en una comunidad lingüística del todo homogénea, que sabe su lengua perfectamente y al que no afectan condiciones sin valor gramatical, como son limitaciones de memoria, distracciones, cambios del

> centro de atención e interés, y errores (…) al aplicar su conocimiento de la lengua al uso real. (Chomsky, 1970: 5)

Desde esta concepção do trabalho linguístico, a rejeição pelos dados empíricos é total, pois a própria concepção mentalista da língua, "realidad mental subyacente en la conducta concreta" (*íbid*.: 6), traz consigo a consideração da *Performance* como "a poor mirror of competence" (McEnery & Wilson, 2001: 6):

> Like most facts of interest and importance (…) information about the speaker's-hearer's competence (…) is neither presented for direct observation nor extractable from data by inductive procedures of any known sort. (Chomsky, 1962 *apud* Tognini-Bonelli, 2001: 50)

Desde este tipo de posicionamento ortodoxamente racionalista, Chomsky atacou frontalmente todos aqueles trabalhos que baseavam as suas investigações em colecções de textos, isto é, em corpus, já que, sendo "the number of sentences in a natural language […] potencially infinite" (McEnery & Wilson, 2001: 8), "[…] some sentences won't occur because they are obvious, others because they are false, still others because they are impolite" (*íbid*.). Deste jeito, Chomsky conclui que "any natural corpus will be skewed. […]"(Chomsky, 1962 *apud* Tognini-Bonelli, 2001: 50), já que não poderia ser construído, nem com a ajuda dos então incipientes computadores, um corpus suficientemente grande como para conter todas as possíveis frases de uma língua.

Explicitada, desta maneira, a parcialidade dos corpora, a introspecção é defendida por Chomsky como a metodologia da investigação linguística: "if you sit and think for a few minutes, you're just flooded with relevant data" (Chomsky, 1984 *apud* McEnery & Wilson, 2001: 11).

Assim, a validade desta metodologia é, por outro lado, explicitamente certificada mercê aos próprios conhecimentos dos investigadores como falantes das línguas objecto de estudo, como pode ser verificado na seguinte conversa:

> Chomsky: The verb *perform* cannot be used with mass word objects: one can *perform a task* but one cannot *perform labour*.
> Hatcher: How do you know, if you don't use a corpus and have not studied the verb *perform*?
> Chomsky: How do I know? Because I am a native speaker of the English language. (Hill, 1962 *apud* McEnery & Wilson, 2001: 11)

De qualquer jeito, apesar de ser a parcialidade dos corpora a principal falha que nesta altura lhe era achacada às investigações de base empírica, outros problemas de processamento dos corpora foram, igualmente, ressaltados. Assim, Abercrombie (1965) numa discussão sobre metodologia linguística desenvolveu o conceito de 'pseudo-procedimento', que aplicou à maioria das aproximações linguísticas baseadas em corpus do seu tempo (*cfr*. McEnery & Wilson, 2001, 12). Este qualificativo de 'pseudo-procedimento' era-lhe assignado, principalmente, à maioria das aproximações baseadas em corpus, já que, não estando na altura em absoluto vulgarizado o uso de computadores entre os linguistas, o processamento de corpora só podia ser levado a cabo "using nothing more than [researcher's] eyes" (*íbid*.), o qual, inevitalvente, implicava empregar um imenso tempo nas buscas, de maneira que as investigações levadas a cabo, além de resultarem estar cheias de erros, resultavam também caras e, portanto, não rentáveis.

Apesar destas críticas, durante esta década de 60, de pleno domínio do generativismo chomskyano na Linguística dos EE.UU. e do estruturalismo saussuriano na Linguística da Europa, foram levadas a cabo diversas investigações e projectos, cuja contribuição é, desde o ponto de vista actual, considerada fundamental para a consolidação da Linguística de Corpus. Assim, por um lado, começou a ser desenvolvida toda uma nova linha de estudos que pretendia automatizar a análise

quantitativa dos corpora, como a de investigadores como Roberto Busa, desde as Humanidades, ou Alphonse Juilland, desde a Linguística. Roberto Busa, religioso jesuíta e estudioso do filósofo Tomás de Aquino, foi o primeiro em produzir em colaboração com IBM um corpus legível por computador, chamado *Index Thomisticus* – que após dezoito anos de trabalhos, acabou por estar constituído em 1967 por 10.600.000 palavras–, e em realizar buscas num corpus digitalizado mediante a ajuda de ferramentas computerizadas (*cfr*. McEnery & Wilson, 2001: 20). Alphonse Juilland que começou em 1956 um projecto de investigação baseado em corpora que ele baptizou como "Mecano-linguística", desenvolveu um primitivo sistema de anotação que lhe servia para a catalogação por géneros dos conjuntos de textos empregados na sua análise quantitativa. Por outro lado, nos inícios dos anos 60 começaram a ser construídos desde a Linguística diversos corpora que, com a passagem do tempo, se converteram em referências hoje me dia obrigadas da Linguística de Corpus, como por exemplo o *Survey of English Usage* ou o *Brown* C*orpus*, desenhados, planificados e construídos por Quirk e Francis & Kucera, respectivamente, em 1960.

Assim mesmo, ocorreram no mundo nesta década de 60 toda uma série de mudanças de ordem política que obrigaram a reformular a metodologia das ciências sociais e que tiveram, por extensão, uma profunda repercussão na Linguística. Assim, por um lado, fruto das guerras anti-coloniais acontecidas em vastos territórios do planeta, mercê à separação política em dois blocos, um comunista e outro capitalista, que se seguiu à II Guerra Mundial; começaram a surgir toda uma série de incipientes nacionalidades e Estados que obrigaram os estados ocidentais a repensarem o papel do "outro" nos novos relacionamentos internacionais desenhados. Por outro lado, dentro dos próprios estados ocidentais o crescimento e afiançamento de "*minorias*" como, por exemplo, os latino-americanos/as e afro-americanos/as nos EE.UU., os gays e as lesbianas, as minorias religiosas como o islamismo, o judaísmo, o hinduísmo, etc., assim como minorias nacionais como as várias nações sem estado europeias ou os índios americanos, fizeram com que os estados tivessem que começar a repensarem o papel do "outro" nos novos relacionamentos intra-estatais desenhados. Ciências como a Sociologia ou a Antropologia viram-se, então, forçadas a ultrapassarem o marco teórico estruturalista dominante, para deste jeito poderem compreender os processos de mudança sócio-política que na altura estavam a acontecer. As manifestações concretas das culturas passaram, assim, a ser o foco de atenção dentro das ciências sociais. Desta maneira, com base na sua observação, tornou-se essencial para o seu estudo o desenvolvimento de um novo objecto de análise e interpretação.

Na Linguística, que na altura se pretendia configurar desde as escolas europeias e americanas tradicionalistas numa ciência eminentemente mentalista e puramente teórica separada das ciências sociais e sem qualquer tipo de projecção aplicada, as mudanças vieram através do surgimento de novas disciplinas, fruto da superação dos marcos teóricos dominantes, da intersecção da Linguística com outras ciências e do surgimento de disciplinas linguísticas aplicadas, fruto da exigência, por parte dos estados industrializados, do desenvolvimento de aplicações que ajudassem a resolver os muito diversos problemas e/ou dificuldades detectados no emprego concreto das línguas em situações de uso reais. Deste jeito, quanto à superação acontecida na Linguística dos marcos teóricos dominantes, cabe ressaltar a importantíssima contribuição de Eugénio Coseriu, após à publicação de *Sincronía, Diacronía e Historia. El Problema del Cambio Lingüístico* na década de 50, decisiva para a superação da antinomia saussureana Sincronia~Diacronia que impedira durante décadas entender o problema da mudança linguística desde uma óptica estrutural. Assim, para Coseriu as línguas, que só funcionavam em Sincronia, isto é, num dado momento, não eram essencialmente estáticas, senão dinâmicas, de jeito que a sua construção só tinha lugar em Diacronia. Assim, as línguas passaram de serem consideradas homogéneas e estáticas a serem

vistas como sistemas tremendamente heterogéneos e em constante reconstrução. O renovado interesse pela mudança operado, sobretudo, nos estudos de História da Língua suscitou ao mesmo tempo que surgisse um, também, renovado interesse pelos estudos de teor variacionista como os da Dialectologia. O clima de optimismo dominante neste tipo de estudos propiciou, por sua vez, o surgimento de novas disciplinas como a Sociolinguística, intersecção da Linguística e da Sociologia, viradas para o estudo da variação inter e intralinguísticas. Os estudos desenvolvidos nos EE.UU. por Labov, Weinreich e Herzog nos anos 60 no térreo da Sociolinguística acabaram por ter, assim, uma grande importância para o desenvolvimento da Linguística de Corpus. A razão é que, mercê aos seus bons resultados conseguidos no conhecimento da variabilidade das línguas em relação com parâmetros de ordem social e à sua vocação eminentemente quantitativa, consistente no processamento estatístico de grandes quantidades de dados empíricos, este tipo de estudos começaram a ser tomados a sério no seio da Linguística.

Ao mesmo tempo que era, então, desenvolvida desde a Sociolinguística uma metodologia quantitativa para o processamento dos dados, na Antropologia foram, por sua vez, operdas mudanças e melhoras a respeito da metodologia empírica de recolha de dados, que resultaram essenciais para a consolidação da Linguística de Corpus. Neste sentido, Malinovski tornou-se, nesta altura, especialmente relevante, já que foi ele quem nos anos 30 começou a desenvolver diferentes trabalhos de campo desenhados para, através de um método chamado em inglês *participant-observation*, poder chegar obter um conhecimento objectivo dos feitos humanos, inseparáveis do seu contexto cultural:

> The neglect of the obvious has been fatal to the development of scientific thought. The false conception of language as a means of transfusing ideas from the head of the speaker to that of the listener has, in my opinion largely vitiated the philological approach to language. The view set forth here is not merely academic: it compels us, as we shall see, to correlate others activities, to interpret the meaning – text; and this means a new departure in the handing of linguistics evidence. It will also force us to define meaning in terms of experience and situation. (Malinovski, 1935 *apud* Kennedy, 1998: 8)

Para a compreensão das raízes e do porquê do surgimento de toda esta nova linha de investigação baseada numa nova conceptualização do trabalho de campo não pode ser, além disso, esquecida a contribuição de um linguista, Sapir, quem, por ter vivido académica e cientificamente à sombra da Linguística Bloomfieldiana, a qual não reconhecia a importância da linha antropológica em que ele se encontrava instalado na última etapa da sua vida académica, isto é, a partir dos anos 30, só começou a ter uma certa influência neste novo contexto desenhado nos anos 60. Sapir foi, deste jeito, junto com Malinovski, um dos primeiros em reconhecer a dimensão social da Língua: "The primary function of language is generally said to be communication […]. Language is a great force of socialization, probably the greatest that exists" (Sapir, 1933 *apud* Fernández Casas, 2004: 194)

Mas Sapir foi muito além. Para ele, "Language is primarily a cultural or social product and must be understood as such" (Sapir 1929 *apud* Fernández Casas, 2004: 194). A consideração da Língua como um produto social levou-o, assim, a dar mais um passo no seu pensamento e a pensar em termos quarenta anos depois retomados pela moderna Pragmática: "meaning is given by the context in which or to which it fits. Implication bears ninety percent of the work of language" (Sapir, 1933 *apud* Fernández Casas, 2004: 215)

Por outro lado, os trabalhos iniciados por Sapir durante os anos 30 referentes ao estudo de comunidades humanas desde uma linha antropológica inspiraram, durante a época de domínio generativista chomskyano, multidão de trabalhos de estudo das línguas dos índios da América. Estas línguas, que começaram a desaparecer sob a pressão dos colonizadores europeus e que, baixo a repressão dos estados norte-

americano e canadiano, continuaram a ser maciçamente abandonadas, tinham, assim, na altura começado a serem consideradas em perigo de extinção. Por parte dos linguistas, aos quais lhes urgia realizarem este tipo de investigações pela razão acima mencionada, a impossibilidade de acederem a estas línguas através de um trabalho de introspecção, mercê, fundamentalmente, a que eram línguas ágrafas, totalmente desconhecidas para aqueles, fez com que tivesse de ser reconhecida a necessidade do emprego, apesar da sua rejeição explícita, de corpora recolhidos mediante uma metodologia de trabalho de campo.

De igual maneira, tinha, também, nesta altura começado a ter influência o trabalho feito durante as décadas de 30, 40 e 50 por J.R. Firth, quem, ao igual que Malinovski e Sapir, pôs especial ênfase na importância de estudar as línguas em função do seu contexto de uso:"The central concept… is the context of situation in that context are the human participant or participants, what they say, what is going on[…]" (Firth, 1957 *apud* McEnery & Wilson, 2001: 23)

Porém, a especial menção que merecem os contributos de Firth em relação com a Linguística de Corpus é devida à sua decisiva influência sobre os chamados "neo-Firthian linguists, such as Halliday, Hoey and Sinclair" (McEnery & Wilson, 2001: 23), que durante as décadas de 80 e 90, sobretudo, começaram a trabalhar a partir da sua Teoria contextual do significado (originalmente em inglês *contextual theory of meaning*) e a empregar conceitos por ele criados, mas hoje em dia amplamente difundidos, como o de *collocation* ou o de *colligation*, que mais adiante serão explicados em pormenor.

A mudança que, deste jeito, foi tendo lugar no seio de, pelo menos, uma parte importante da Linguística que, nesta altura, se tinha virado para os dados de natureza empírica, trouxe consigo que o objecto de estudo da Linguística se fosse ampliando, tendo-se, assim, em conta também os contextos situacionais comunicativos, a experiência e conhecimento do mundo por parte dos/as interlocutores/as, assim como as suas intenções ou desejos. Desta maneira, durante as décadas de 60 e 70 foi tomando corpo uma nova disciplina, hoje conhecida como Pragmática, que enfraqueceu o domínio da Linguística Estruturalista e Generativa mediante o questionamento da sua rejeição da *Parole* (usos linguísticos concretos) para a correcta e completa compreensão da dimensão semântico-pragmática das línguas:

> One could say that, in general, the "pragmatic perspective" centers around the adaptability of language, the fundamental property of language which enable us to engage in the activity of talking which consists in the constant making of choices, at every level of linguistic structure, in harmony with the requirements of people, desires and intentions, and the real world circumstances in which they interact. (Verschueren, 1987 *apud* Fernández Casas, 2004: 214)

Além de enfraquecer o domínio das acima referidas tendências dominantes de corte racionalista, questão da maior importância para o afiançamento da Linguística de Corpus, o surgimento da Pragmática elevou aquela, em datas mais recentes, para uma nova dimensão nas disciplinas linguísticas aplicadas, tal e como mais à frente será pormenorizado.

Outro dos factores determinantes para o desenvolvimento da referida Linguística de Corpus foi o surgimento de disciplinas linguísticas aplicadas, mercê às exigências feitas pelos estados industrializados, ocidentais e orientais, dado o novo *status quo* sócio-político desenhado após a divisão do mundo, depois da II Guerra Mundial, em dois blocos e a independização de vastos territórios africanos e asiáticos. Foi, assim, neste contexto que surgiram disciplinas como a Planificação Linguística e a Tradução Automática –inserida esta última hoje em dia dentro da chamada Linguística Tecnológica. Apesar da grande influência que teve o informe ALPAC (*Automatic Language Processing Advisory Commitee*) da *National Academy of Science* dos EE.UU.

que, em 1966, travou o forte interesse por parte do governo norte-americano em dispor de sistemas de tradução automática do par russo-inglês e, portanto, os importantes investimentos destinados a essa área de investigação nos inícios da década de 60, o desenvolvimento da Tradução Automática nas décadas de 70, 80 e 90 foi imparável. Foi precisamente nesta última década em que os trabalhos realizados nesta disciplina aplicada cobraram especial relevância para a Linguística de Corpus, pois com o ascenso dos modelos de tradução probabilísticos os corpora se converteram, como veremos mais adiante, em fonte primária de dados.

De qualquer jeito, uma outra importante mudança acontecida na década de 70 acabou por dar um pulo definitivo à Linguística de Corpus. Esta mudança foi o rápido desenvolvimento das novas tecnologias ocorrido no âmbito dos computadores. Estes, que na sua primeira geração não estavam em absoluto popularizados nem entre os investigadores nem entre o público em geral, dado o seu enorme tamanho e desmedido preço em relação com as suas relativamente baixas velocidade de processamento e capacidade de armazenamento, sofreram uma surpreendente popularização após o surgimento dos chamados PCs. Os PCs revolucionaram o mercado das novas tecnologias porque conseguiram, graças ao progresso registado com a invenção de chips de ínfimas dimensões e ao desenvolvimento de novos suportes de armazenagem, reduzir o tamanho das máquinas e, ao mesmo tempo, aumentar exponencialmente tanto a sua velocidade de processamento quanto a sua capacidade de armazenagem. No âmbito da Linguística de Corpus, isto supôs a possibilidade de construir corpora de milhões de palavras nos quais levar a cabo complexas buscas de informação num tempo de processamento muito inferior. Neste sentido, as novas potencialidades derivadas do emprego de computadores para a realização das tarefas de busca, que conseguiam, assim, reduzir o preço e aperfeiçoar a precisão destas tarefas, começaram, então, a pôr em causa as críticas vertidas desde a Linguística Generativa que diziam respeito, por um lado, à impossibilidade de construir corpora suficientemente grandes como para serem representativos de qualquer variedade linguística e, por outro lado, à condição de 'pseudo-procedimento' da própria Linguística de Corpus.

Desta maneira, graças a estas mudanças, nas décadas de 70, 80 e 90 foram desenvolvidos outros corpora, tanto de língua escrita quanto oral, hoje internacionalmente conhecidos e afamados como o *London-Lund corpus*, iniciado por Jan Svartvik em 1975, o projecto COBUILD (*Collins Birmingham University Language Database*), iniciado por John Sinclair em 1980, ou o *Lancaster-Oslo-Bergen corpus* (LOB) e o *British National Corpus* (BNC), construídos baixo a direcção de Geoffrey Leech nos anos 90.

Por último, para finalizarmos esta breve introdução histórica, cabe ainda mencionar que, nos anos 90, tiveram lugar aquelas que podem ser consideradas as últimas grandes mudanças acontecidas durante o século XX que propiciaram o afiançamento da Linguística de Corpus: a estandardização dos modelos de anotação dos corpora, que permitiram a introdução de informação adicional nos textos dos corpora com o intuito de a aproveitar na sua exploração. Foi, pois, após o encontro realizado em 1987 no Vassar College de Poughkeepsie (New York) que se formou a chamada Text Encoding Initiative (TEI), constituída pela Association for Computers and the Humanities (ACH), a Association for Literary and Linguistic Computing (ALLC) e a Association for Computational Linguistics (ACL). Desta maneira, em 1990, a TEI publicou, depois dalguns anos de investigação, as TEI Guidelines P1 para a anotação de corpora, que não eram senão umas directrizes desde as quais se pretendia, tal e como mais à frente será explicado em pormenor, unificar o heterogéneo campo da anotação. A elaboração destas Guidelines teve, assim mesmo, outro efeito positivo para o desenvolvimento da Linguística de Corpus, já que para a elaboração das diferentes etiquetas empregadas para a codificação dos diferentes campos de anotação de corpora

recomendados nas Guidelines a TEI adoptou como linguagem de anotação o SGML (Standard Generalized Markup Language). Com a passagem dos anos, o emprego do SGML como linguagem estândar de anotação de corpora trouxe consigo, além disso, a sua especialização e o subsequente desenvolvimento de uma nova linguagem de anotação herdeira, conhecida como XML (*eXtensible Markup Language*), mais potente e versátil.

2. O que é a Linguística de Corpus?

Feita, pois, esta introdução à história da evolução e progressão da Linguística de Corpus ao longo de vários séculos mas, sobretudo, ao longo do século XX, encontramo-nos plenamente imersos, assim, no que é à nossa actualidade científica diz respeito. Antes de introduzirmo-nos, porém, na descrição do trabalho que nos ocupa, caberia, ainda, perguntar-se acerca do que nós entendemos aqui por Linguística de Corpus.

Após uma leitura demorada da breve introdução realizada no capítulo anterior, pode-se observar que várias são as denominações que ao longo da história da Linguística apareceram para denominar similares aproximações baseadas no uso de corpora, por exemplo: ramo quantitativo-estatístico da Linguística Matemática, aplicada na década de 30; Mecano-Linguística, aplicada na década e 50, e Linguística de Corpus, denominação maioritária hoje em dia mas que, de facto, só apareceu após as críticas realizadas na década de 60 por Chomsky, a quem realmente tem de ser atribuída a sua rápida expansão e consolidação nas décadas de 60, 70, 80 e 90.

Apesar dos seus 40 anos de história, na actualidade "among linguists at large, corpus linguistics is not universally acclaimed for its contributions to linguistic theory" (Svartvik, 1996: 11). Parece, pois, claro que a Linguística de Corpus se encontra imersa num processo de consolidação no seio da Linguística, que explica a existência, ainda na actualidade, de uma controvérsia científica a respeito do seu *status*.

Num extremo, encontram-se linguistas como Tom McEnery e Andrew Wilson que consideram a Linguística de Corpus como uma mera metodologia de trabalho aplicável a muito diferentes investigações com, também, muito diferentes finalidades:

> Corpus linguistics is not a branch of linguistics in the same sense as syntax, semantics, sociolinguistics and so on. All of these disciplines concentrate on describing/explaining some aspect of language. Corpus linguistics in contrast is a methodology rather than an aspect of language requiring explanation or description. A corpus-based approach can be taken to many aspects of linguistics enquiry. […] Corpus linguistics is a methodology that may be used in almost any area of linguistics, but it does not truly delimit an area of linguistics itself. (McEnery & Wilson, 2001: 2)

Nesta mesma linha encontra-se também o linguista Geoffrey Leech:

> But is corpus linguistics really comparable with these other hyphenated branches of linguistics? [se refiere a la sociolingüística, la psicolingüística, la lingüística del texto]. No, because 'corpus linguistics' refers not to a domain of study, but rather to a methodological basis for pursuing linguistic research. In principle (and often in practice) corpus linguistics combines easily with other branches of linguistics: we can study phonetics, syntax, sociolinguistics, and any other aspect of linguistics by means of corpora, and when we are doing this we can be said to be combining techniques of corpus linguistics with the subject matter of phonetics, syntax, sociolinguistics, and so on. (Leech, 1992: 105-106 *apud* Caravedo, 1999: 19)

Noutro extremo, encontram-se linguistas como Elena Tognini-Bonelli, quem considera que

> although corpus linguistics belongs to the sphere of applied linguistics, it differs from the other partner disciplines under the same umbrella in that it can be seen as a *pre-application*

> *methodology*. […] by pre-application we mean that, unlike other applications that start by accepting certain facts as given, corpus linguistics is in a position to define its own sets of rules and pieces of knowledge before they are applied […] corpus linguistics has, therefore, a theoretical status and because of this it is in a position to contribute specifically to other applications. (Tognini-Bonelli, 2001: 1)

A propósito desta consideração da Linguística de Corpus como um estádio prévio a outro tipo de estudos –teóricos e aplicados– a mesma autora, umas páginas mais à frente expressa-se nos seguintes termos:

> Saussure's famous words 'c'est le point de vue qui crée l'object', can be reinterpreted in this turn of events […]. We could say that is our chosen methodological standpoint which determines both the object and the aim of our enquiry. (*ibid*.: 49)

Diversas apreciações podem ser feitas, porém, a respeito destas duas posições encontradas. Por um lado, dizer, para contestar o primeiro dos posicionamentos acima apresentados, que, tal e como afirma Rocío Caravedo (1999: 19),

> si por metodología se entiende el conjunto de técnicas y estrategias de aproximación y manejo de la realidad que se persigue estudiar, también la *lingüística de corpus* puede agrupar distintas metodologías […], y no se identifica con sólo una de ellas, sino más bien con un modo de relación entre teoría y realidad.

Por outro lado, para contestar a posição de Tognini-Bonelli, dizer que o facto de que se possa elaborar um construto teórico à volta de uma determinada metodologia de estudo, não implica que dito construto se tenha, ineludivelmente, de erigir numa disciplina autónoma da área disciplinar desde a qual aquela teorização tinha sido elaborada. Precisamente se se aceitar que o uso de corpora foi durante décadas uma mera metodologia empregada desde diferentes áreas da linguística, o facto de que actualmente se tenha formulado toda uma colecção terminológica nova, surgida, aliás, após o desenvolvimento dos computadores e da aplicação de técnicas estatístico-quantitativas no processamento dos corpora, não é senão um indicador dos progressos feitos no seio daquelas áreas disciplinares, que se adaptaram às potencialidades oferecidas por novas ferramentas de estudo. Por isso, de igual maneira, cabe também dizer, para matizar a interpretação das palavras de Saussure acima apresentadas, que uma reformulação do objecto de estudo, motivada por uma teorização prévia à aplicação de uma certa metodologia, modifica, em todo o caso, o ponto de vista epistemológico das disciplinas nas quais se pretende aplicar dita metodologia. De qualquer maneira, em toda disciplina científica, independentemente de que a sua posição acerca dos dados seja mais ou menos empírica, a teorização acerca da metodologia a empregar para atingir uns determinados objectivos é, antes de mais, uma necessidade epistemológica. Além disso, resulta imprescindível também indicar que a popularização de uma metodologia baseada em corpus não foi senão uma consequência de uma mudança generalizada no terreno filosófico no seio das ciências sociais, após as mudanças experimentadas no mundo nas décadas de 40, 50, 60 e 70 e as subsequentes necessidades surgidas nas diversas sociedades industrializadas, que alentaram e exigiram das ciências sociais o desenvolvimento de soluções aplicadas. Foi, deste jeito, que apareceram as disciplinas aplicadas da Linguística, entre as quais sem dúvida devemos incluir a chamada Linguística Tecnológica, macro-etiqueta dentro da qual se reúnem a Linguística Informática, a Linguística Computacional, as Tecnologias da Fala e as Industrias da Língua, mas não a Linguística de Corpus.

Desta maneira, a Linguística de Corpus neste trabalho é entendida como um conjunto de metodologias de base empírica, aperfeiçoadas mercê às necessidades surgidas após a reformulação filosófica do objecto de estudo da Linguística no último

terço do século XX, pelo menos no que a uma parte da Linguística se refere, que passou a considerar os dados empíricos como fonte legítima para a consecução de um conhecimento objectivo das línguas. A Linguística de Corpus não se cinge, portanto, a uma disciplina linguística particular, senão que pode ser aplicada em muito diferentes disciplinas da linguística e com muito diferentes finalidades. Assim, "[it] defines not just a newly emerging methodology for studying language, but a new research enterprise, and in fact a new philosophical approach to the subject" (Leech 1992a *apud* Svartvik, 1996: 12).

Apesar, por outro lado, das enormes vantagens de poupança em tempo, esforço e dinheiro que supõe o uso dos computadores para a armazenagem e processamento dos corpora, facto pelo qual hoje em dia a quase totalidade dos corpus em uso foram ou são realizados e/ou transvasados a formato electrónico, o emprego de corpora implica toda uma orientação de pensamento não indentificável com um recurso tecnológico, já que "reducir una orientación de pensamiento a [isto] no solo desmerece la propia propuesta sino que lleva a desconocer la larga práctica de la lingüística precomputerizada que se valía de la construcción de corpus para el estudio de las lenguas." (Caravedo, 1999: 17).

3. O que é um corpus?

Muitas são as diferentes definições de corpus que podem ser encontradas nas publicações académicas especializadas em Linguística de Corpus:

> a corpus is a collection of texts assumed to be representative of a given language, dialect or other subset of a language to be used for linguistic analysis. (N. Fancis, 1992: 17 quoting Fancis 1982: 7)
>
> a corpus is a collection of naturally-occurring language text, chosen to characterize a state or variety of a language. (Sinclair, 1991: 171)
>
> [a corpus is] a subset of an ETL (Electronic Text Library) built according to explicit design criteria for a specific purpose (Atkins, Clear & Osler, 1992: 1)
>
> a corpus is understood to be a collection of samples of running text. The text may be spoken, written or intermediate forms, and the samples may be of any length. (Jan Aarts, 1991: 45)
>
> A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language. (EAGLES 1994: 2.1) (Tognini-Bonelli, 2001: 53)

Como pode ser verificado, embora exista um maioritário acordo a respeito da consideração dos corpora como uma colecção de textos, cada uma destas definições incide sobre diferentes características que tanto os textos que formam parte das colecções como as próprias colecções de textos devem cumprir para poder ser consideradas corpora.

3.1. Representatividade

A primeira destas características diz respeito a um dos conceitos sobre o qual, por ter resultado ser uma das questões mais determinantes para a certificação da validade do uso de corpora tanto desde uma perspectiva académica quanto industrial, mais se tem versado na Linguística de Corpus: a *representatividade* das colecções de textos, isto é, dos próprios corpora em relação com a língua ou variedade com a qual se pretende, mediante eles, trabalhar. Deste modo, "according to Leech […] a corpus is representative when *the findings* based on its contents can be generalized to a larger hypothetical corpus" (Leech, 1991 *apud* Tognini-Bonelli, 2001: 57). Tendo, porém, em

conta que o próprio autor afirma que "the assumption of representativeness *must be regarded largely as an act of faith"(íbid*.), como fazer frente, então, às críticas vertidas por Chomsky na década de 60 que iam no sentido de considerar que em um corpus "some utterances would be excluded because they are rare, other […] utterances might be excluded simply by chance, and chance might also act so that some rare utterances were actually included" (McEnery & Wilson, 2001: 30).

Neste sentido, só uma nova concepção do funcionamento das línguas e um novo quadro de estudo poderiam permitir superar a inevitável parcialidade inerente de todo corpus. Como foi referido na anterior introdução histórica, o desenvolvimento da Pragmática a partir da década de 60 resultou de extraordinária importância para a Linguística de Corpus, não só pelo seu questionamento a respeito da posição epistemológica e da supremacia das teorias linguísticas racionalistas, senão, sobretudo, pelo novo quadro teórico que se desenhou. Desde um ponto de vista pragmático, as línguas naturais são códigos de comunicação e socialização entre humanos que só funcionam num determinado contexto, seja situacional, histórico, etc. Isto implica, pois, que, para uma correcta interpretação do significado, isto é, para uma efectiva comunicação, os/as destinatários/as de uma mensagem precisam realizar toda uma série de descodificações e inferências semânticas acerca da intenção comunicativa dos/as emissores/as, nas quais tanto o contexto situacional como a informação pragmática, isto é, a informação co(n)textual e o conhecimento do mundo e do interlocutor, jogam um papel determinante. Sob esta perspectiva, cada uma das intervenções, isto é, cada um dos *enunciados* dos/as interlocutores/as constitui uma unidade de interacção, cuja natureza não tem necessariamente de coincidir com a noção tradicional de oração. Um enunciado é, pois, entendido como uma unidade do discurso enquadrada entre duas pausas, e delimitada pela mudança de emissor (*cfr*. Escandell, 2002: 27-28) e cuja extensão pode, desta maneira, ultrapassar a da oração tradicional. Desde este ponto de vista, um texto inteiro, seja um livro, um poema, um artigo, etc., pode constituir, dentro desta teoria da comunicação, uma unidade de interacção de um emissor, isto é, um enunciado. Em palavras de Tognini-Bonelli (2001: 2): "[…] we can say that text exists in a unique communicative context as a single, unified language event mediated between two (sets of) participants". Isto, que poderia parecer meramente anedótico, supôs uma revolução para a Linguística de Corpus, pois preparou o terreno para a sua consolidação ao abrir o caminho para o uso de colecções de textos, tanto escritos quanto orais, para o estudo e processamento das línguas naturais.

Houve, porém, uma teoria, que poderia ser considerada hoje em dia precursora dos estudos pragmáticos gerais e que, apesar da sua marginalização durante décadas, teve uma especial importância para a Linguística de Corpus actual. Referimo-nos, em concreto, à Teoria Contextual do Significado de Firth (em inglês *Contextual Theory of Meaning*),

> where a text is seen as an integral part of its context and the formalisation of a contextual patterning of a given word or expression is assumed to be relevant to the identification of the meaning of that word or expression. (Tognini-Bonelli, 2001: 4)

Partindo de que cada emissão ocorre num contexto situacional determinado culturalmente, Firth afirma que "the relation of utterance to context is a special case of Item and Environment, and it persists when applied to smaller units of language than sentence" (*íbid*.). Desde este ponto de vista, assume-se, então, que cada elemento linguístico ocorre num determinado contexto e que este é altamente relevante para a determinação do significado do elemento, seja este uma palavra ou uma oração (*cfr*. *íbid*.). Dentro desta teoria, dois são os conceitos importantes a respeito do tipo de relação co-textual estabelecido entre os diferentes elementos de um texto: as 'colocações' (em inglês *collocations*): "the mere word accompaniment, the other word-

material in which they are most commonly or most characteristically embedded" (Palmer, 1968 *apud* Tognini-Bonelli, 2001: 5); e as 'coligações' (em inglês *colligations*): "the interrelation of grammatical categories in syntactical structure" (*íbid*.).

Em definitiva, dentro desta nova concepção as línguas deixam, assim, de ser entendidas não só como uma possibilidade abstracta, tal e como podia ser extraído da formulação do conceito chomskyano de *Competence*, e passam, portanto, a ser consideradas também como uma probabilidade em co(n)texto.

De qualquer jeito, é dentro do quadro teórico pragmático delimitado por Firth que pode e deve ser entendido o conceito de representatividade de um corpus. A razão é que, concebendo as línguas como uma probabilidade em co(n)texto, o facto de que um corpus possa ser considerado suficientemente representativo como para poderem ser generalizadas as descobertas nele feitas a corpora de maiores dimensões, só pode ser sustido pela noção Firthiana de "eventos repetidos", ligada com a noção de *behavior* da linguística estruturalista bloomfieldiana. Segundo esta noção as actualizações concretas das línguas –embora só parcialmente, conforme deve ser matizado– respondem, em relação com o contexto, a parâmetros de conduta, em muitos casos, repetitiva. Desta maneira, existe, desde uma perspectiva probabilístico-quantitativa, a possibilidade de analisar ou processar colecções finitas de textos e de realizar, a partir delas, generalizações a entidades como as línguas naturais, por definição, não só maiores, mas infinitas. Assim, como a representatividade é um conceito matizável em si mesmo, para a consecução de um grau de representatividade suficientemente elevado num corpus duas são as considerações que hão-de ser tidas em conta: o seu volume de palavras e a sua amostragem (*sampling* em inglês), que, dependendo do tipo de entidade linguística que se pretenda representar, serão, assim mesmo, variáveis.

Tendo em conta quais eram as críticas iniciais à moderna Linguística de Corpus não surpreende, pois, que a questão do volume de palavras se tenha erigido numa questão da máxima centralidade e, ao mesmo tempo, de enorme controvérsia na Linguística de Corpus. Tal e como aponta Sinclair (2000): "There are of course some kinds of corpora which are by their nature small […] [isto é] […] Some corpora are inevitably small and we cannot do anything about them […]" (Sinclair, 2000: 32-33). Este é o caso, por exemplo, dos corpora de línguas mortas, como o latim, ou dos corpora de determinadas variedades linguísticas, como os literários, construídos em função de critérios extralinguísticos de representatividade e que, por natureza, são limitados. Contudo, nos últimos anos verifica-se a tendência a construir corpora de extraordinárias dimensões, como os chamados corpora de *referência* que, como o CREA (Corpus de Referencia del Español Actual), podem chegar a contar com milhões de textos representantes de multidão de variedades linguísticas; ou como os chamados corpora *monitor*, como o projecto COBUILD, criado baixo a direcção de John Sinclair, cuja característica principal é que se encontram num contínuo processo de construção, já que são objecto de um constante acréscimo de textos.

Assumido, pois, que um corpus tem de ser representativo, as decisões mais importantes que terão de ser tomadas para a sua construção são aquelas que dizem respeito à sua amostragem. Dependendo do tipo de corpora que se deseje construir, necessariamente haverá que decidir, pois, sobre "issues such as what kind of texts are to be selected, the number of texts, the selection of particular texts, the selection of texts samples from within texts and the length of the samples." (Tognini-Bonelli, 2001, 59). Deste jeito, todo corpus que se pretenda representativo necessitará guardar um certo equilíbrio entre o tipo e o número de textos de cada uma das variedades incluídas e o tamanho das diferentes amostras representantes destas variedades.

De qualquer forma, se houve algo que, finalmente, acabou por tornar inválidas ou, quanto menos, matizáveis aquelas críticas vertidas desde o Generativismo e que, do

meu ponto de vista, supôs, nas últimas décadas, o definitivo pulo para a Linguística de Corpus no nível da potencial representatividade que os corpora podem alcançar, foi o emprego dos computadores. Como antes foi referido, os computadores permitiram, por um lado, a construção de corpora de grandes dimensões e facilitaram, por outro, o emprego de técnicas estatísticas para o seu processamento que, além do mais, permitiram, desde uma óptica quantitativa, realizar testes de significância, como o *chi-squared test* (para uma descrição mais detalhada ver McEnery e Wilson, 2001, 81-101), através dos quais é possível verificar a validade dos próprios resultados.

3.2. Autenticidade

A segunda característica dos corpora na qual se incide com especial ênfase nalgumas das definições acima referidas é a pressuposição de autenticidade dos dados incluídos num corpus, ou seja, "all the material included in a corpus, […], is assumed to be taken from genuine communications of people going about their normal business." (Tognini-Bonelli, 2001: 55). Do meu ponto de vista, esta não é senão uma característica directamente derivada do planejamento filosófico subjacente a todo estudo baseado em corpora, isto é, desde uma perspectiva empirista. Como pode ser induzido após uma leitura demorada da introdução histórica antes realizada, o emprego de corpora surgiu nas primeiras investigações de carácter linguístico do século XIX, mais do que como uma alternativa metodológica, como uma necessidade dado o objecto de estudo e os objectivos que se perseguiam. Assim, a negação feita por Saussure do recurso aos dados obtidos através da observação da *parole*, durante décadas, significou para os linguistas formados nesta escola a impossibilidade não só de estudar, senão, sobretudo, de entender a natureza heterogénea das línguas. Desta maneira, quando nas décadas de 30, 40 e 50 se pretendeu dar um novo giro às investigações linguísticas, passando mais uma vez a empregar dados de natureza empírica, os corpora voltaram surgir como a metodologia mais adequada para o estudo deste tipo de dados. É por tudo isto que se pode afirmar que o trabalho com corpora implica uma "*empirical approach* in that […] the start point is […] authentic data." (*ibid*.: 2)

3.3. Natureza escrita ou oral

A terceira característica dos corpora diz respeito à natureza escrita ou oral dos dados linguísticos que podem estar incluídos num corpus. Embora na actualidade existam alguns corpora onde se incluem textos linguísticos procedentes da oralidade, esta não foi nem muito menos uma constante nas diversas aproximações baseadas em corpora. Por um lado, nos primeiros trabalhos do século XIX não existia a tecnologia necessária para registar este tipo de manifestações linguísticas. Por outro lado, se o desenvolvimento dos diferentes dispositivos analógicos de gravação de som que foram desenvolvidos no século XX possibilitou por fim o recurso a este tipo de dados, a sua própria natureza analógica não admitiu, durante décadas, um processamento computerizado e automatizado. O surgimento, porém, na actualidade de ferramentas de gravação e suportes de armazenagem e reprodução digitais fez não só com que este tipo de corpora orais pudessem ser mais rápido e melhor recompilados, senão com que sofressem um extraordinariamente veloz avanço que os colocou ao mesmo nível de concorrência que os corpora escritos.

3.4. Formato de armazenagem

A quarta característica apontada nas anteriores definições diz respeito à natureza do formato em que são armazenados os textos. Como pode ser verificado, na terceira das

definições um corpus é definido como um conjunto de bibliotecas de textos electrónicos (em inglês *Electronic Text Library*, ETL). Do meu ponto de vista, caberia matizar que, em consonância com o já exposto, isto não é senão uma redução, em relação com um só recurso tecnológico, dos tipos de formato de armazenagem dos corpora. Durante décadas, apesar da inexistência de computadores, foram realizadas diversas investigações baseadas no uso de corpus, empregando para a sua armazenagem o único recurso disponível, ou seja, o papel. Tal e como foi antes foi sublinhado, isto supunha uma enorme eiva para este tipo de investigações, pois além do tempo, dinheiro e esforço necessários para o processamento dos corpora, a própria metodologia empregada para o seu processamento erigia-se num 'pseudo-procedemento', mercê à impossibilidade da sua automatização. É certo que os computadores conseguiram resolver grande parte destes problemas e que tornaram, assim, inválidas aquelas críticas. Também é certo que hoje em dia ninguém recompila e armazena nenhum corpus se não é empregando um formato electrónico digital. De qualquer forma, o emprego unânime de uma única ferramenta tecnológica não pode determinar de por si uma natureza unitária para o formato de armazenagem dos corpora. Não seria justo, mas não seria, sobretudo, lícito desde um ponto de vista historiográfico. Em qualquer caso, o facto de que hoje em dia só se empreguem e construam corpus de textos em formato electrónico obriga-nos, pois, a ter de definir o que o formato electrónico representa em relação com a natureza dos textos. Um texto electrónico "non é máis que unha sucesión de caracteres ASCII [embora também UNICODE], é dicir, unha sucesión de bytes almacenada nun soporte determinado (en CD-ROM, en disquete ou, normalmente nos discos duros de potentes estacións de traballo situadas nos centros de textos electrónicos). O texto electrónico, polo tanto, non ten entidade física, non é un obxecto material: non o podemos ver nin tocar." (Álvarez, 1996: 88).

3.5. Função

A quinta e última característica dos corpora citada na terceira das anteriores definições é a necessidade de delimitação da função específica que se lhes quiser dar aos próprios corpora. Função específica que, por sua vez, virá determinada pelos objectivos que, desde uma óptica académica e/ou industrial, se persigam e para os quais resulte necessária a sua compilação e construção. Neste sentido, pode-se afirmar, pois, que os objectivos da empresa académica e/ou industrial determinam *a priori* certos aspectos que têm a ver com o tipo de colecções de textos necessários para os atingir. Assim, dependendo de quais sejam os objectivos desejados, terão de ser tomadas, em relação com os diferentes graus de representatividade atingíveis pelos corpora, algumas decisões que dizem respeito à amostragem de textos que serão neles serão incluídos, isto é, tanto ao tipo de textos quanto à sua extensão. Se bem na última das definições acima apresentadas são referidos os critérios de tipo linguístico como os principais critérios aos quais se tem de submeter a compilação de uma colecção de textos para que esta possa ser considerada um corpus, do meu ponto de vista, toda a complexa série de critérios que, na sua construção, hão-de ser tomados em consideração para a delimitação de uma amostragem de textos suficientemente representativa para lograr uns determinados objectivos não pode responder a uma classificação tão simplista. Deste modo, os textos podem ser escolhidos em função de

> **criterios externos** (que hacen referencia a los tipos de texto que establecen las clasificaciones más usuales, o las características del contexto social en que ocurren los textos) y **criterios internos** (que hacen referencia a las características diferenciadoras del lenguaje del texto). (Rafel i Fontanals & Soler i Bou, 2003: 48)

Como os critérios internos "por el hecho de que no están avalados por ningún tipo de evidencia científica, ya que, en la práctica, conviven diferentes clasificaciones [con niveles variables de detalle] sin criterios objetivos que hagan preferir una u otra" (*cfr*. *íbid*.), neste trabalho, em virtude da sua maior objectivabilidade, atender-se-á a critérios externos baseados de maneira pormenorizada nas clasificações feitas por Atkins, Clear & Ostler (1992) e por Rafel i Fontanals & Soler i Bou (2003). Deste jeito, entre este tipo de critérios podemos encontrar, entre outros, os seguintes:

- Modo: Segundo este critério os textos podem ser escritos, escritos para ser lidos, escritos para ser falados, falados ou falados para ser escritos.
- Origem: Neste critério têm-se em conta aspectos que têm a ver com o local de origem do texto, a nacionalidade do/a autor/a ou do/a editor/a, etc.
- Participação: Este critério diz respeito ao número de pessoas originadoras do texto.
- Preparação: Sob esta denominação são classificados os textos em função do nível de preparação, isto é, se são mais ou menos espontâneos.
- Meio: Este critério refere-se ao meio de publicação dos textos: em livro, revista, televisão, rádio, etc.
- Estilo: Com o intuito de oferecer soluções práticas que não acabem por complicar muito a classificação dos textos, os estilos actualmente distinguidos segundo este critério são a prosa e o verso.
- Género: Os géneros em que, na escrita, podem ser classificados os textos são: romance, conto, obra de teatro, poema, ensaio, carta, anúncio, horóscopo, etc. Na oralidade, por sua vez, são leitura, debate, conversa, etc.
- Domínio: Este critério refere-se ao contexto social ao qual pertence o texto, isto é, escolar, laboral, familiar, etc.
- Função: Neste critério há-de se ter em conta a força ilocutiva dos textos, segundo a qual estes podem ser narrativos, informativos, exortativos, persuasivos, instrutivos, etc.
- Tema: Este critério refere-se ao domínio de conhecimento ao qual pertence o texto, p. ex.: biologia, química, literatura, música, lazer, política, etc.
- Tecnicidade: Segundo este critério os textos podem também ser classificados em função do grau de especialização técnica dos/as emissores/as ou dos/as receptores/as. Assim, os textos podem ser gerais, técnicos ou semi-técnicos.
- Data: Este critério está relacionado com a data em que os textos foram publicados –se escritos– ou emitidos –se orais–.
- Status do texto: Segundo este critério os textos podem ser originais, actualizações, revisões, etc.
- A(s) língua(s): Por muito evidente que pareça a selecção de textos procedentes de uma ou outra língua é uma escolha que, necessariamente, terá de ser realizada.
- Ligação entre línguas: Em relação com este critério os textos podem estar isolados ou alinhados em pares, tripletos, quartetos, etc.
- Status da(s) língua(s): Segundo este critério as línguas podem ser fonte ou tradução.
- Autoria: Este critério tem a ver com os autores/emissores dos textos.
- Sexo do/a autor/a: Tradicionalmente os/as autores/as são classificados em homens ou mulheres.
- Idade do/a autor/a: Este critério classifica os/as autores/as do texto em função da sua idade.

- Língua nativa do autor: Este é, como resulta evidente, um critério de classificação dos/as autores/as em função da sua língua inicial e não da língua do texto.
- Idade dos/as receptores/as: Segundo este critério o público receptor de um texto pode ser adulto, juvenil, infantil, etc.

3.6. Tipologia de corpora

Os tipos de corpora que deste jeito ficam, então, delimitados são os seguintes:

1. Corpora escritos: As colecções de textos escritos representam actualmente a maioria dos corpora disponíveis.

1.1. Sincrónicos ou diacrónicos: Os corpora escritos podem ser tanto sincrónicos, isto é, representativos dalguma variedade linguística ou de toda uma língua num determinado ponto do tempo, ou diacrónicos, ou seja, que contêm textos procedentes de uma série de diferentes momentos temporais.

1.1.1. Monolingues: Os corpora monolingues são um tipo de corpora empregues, sobretudo, na elaboração de gramáticas baseadas em usos concretos e reais, na elaboração de dicionários electrónicos e outro tipo de ferramentas de correcção gramatical.

1.1.1.1. Específicos: Os corpora específicos são aquele tipo de corpora nos quais estão incluídos variedades linguísticas restringidas, isto é, as chamadas *Languages for Special Pourposes* (LSP). Estes podem-se, por exemplo, empregar no ensino de segundas línguas ou na elaboração de bases de dados terminológicas restringidas a âmbitos muito específicos.

1.1.1.2. Gerais: Dada a sua natureza generalista, isto é, pretensamente representativa de uma língua, estes são o tipo de corpora monolingues empregues, como antes foi dito, na elaboração de gramáticas baseadas em usos concretos e reais, dicionários electrónicos, etc.

1.1.1.2.1. Corpora de referência: Os corpora de referência são corpora que incluem textos de uma ampla gama de variedades de uma língua, pelo qual "can be taken as representative of the language as a whole." (Toginini-Bonelli, 2001: 9)

1.1.1.2.2. Monitor: São um tipo de corpora que têm como característica diferenciadora a sua condição de "open-ended entit[ies]" (McEnery & Wilson, 2001: 30), ou seja, de corpora sem um limite de palavras definido, o qual implica um continuo crescimento do volume palavras.

1.1.2. Multilingues: Além dos corpora monolingues, existem na actualidade multidão de corpora multilingues que, como o seu próprio nome indica, incluem textos escritos em mais de uma língua.

1.1.2.1. De tradução: "Translation corpora are corpora of texts which stand in a translational relationship to each other, […]. The most common use of translational corpus […] remains the access to translations as products where the translated corpora reveal cross-linguistic correspondences and differences that are impossible to discover in a monolingual corpus." (Toginini-Bonelli, 2001: 6)

1.1.2.1.1. Paralelos: São corpora que contêm a mesma colecção de textos em mais de uma língua, ou seja, as versões originais acompanhadas pelas suas traduções (*cfr*. Abaitua, 2002: 65), que recebem o qualificativo de paralelos porque os diferentes pares ou tripletos de textos estão alinhados entre si. Assim mesmo, este tipo de corpora, dependendo da distribuição dos tipos de textos (fonte ou tradução) em relação com as línguas empregadas, podem ser classificados em:

1.1.2.1.1.1. Corpora unidireccionais: São corpora que só incluem textos fonte escritos na língua A e textos traduzidos escritos na língua B.

1.1.2.1.1.2. Corpora bidireccionais: São que corpora incluem textos fonte escritos na língua A, textos traduzidos escritos na língua B, textos fonte escritos na língua B e textos traduzidos escritos na língua A.

1.1.2.1.2. De traduções livres: Estes são um tipo de corpora nos quais estão incluídos mais de uma versão traduzida de um mesmo texto, mas cujo alinhamento não resulta possível, dada a condição de traduções livres das diferentes versões.

1.1.2.2. Comparáveis: "Baker 1995 introdujo este término para corpora monolingües compuestos por textos originales en una lengua y traducciones de otros textos semejantes en la misma lengua. Martínez 1999 amplía el término a corpora multilingües que contienen textos en distintos idiomas, que sin ser traducciones, comparten similar origen, temática, extensión y número." (*íbid*.). Como pode ser deduzido, este tipo de corpora baseiam, pois, a sua comparabilidade no facto de incluírem textos em diferentes línguas de similares características.

1.1.2.3. Com textos em diferentes idiomas: São colecções "de textos en varios idiomas recopiladas con la intención de servir para estúdios cuantitativos o estadísticos […] (pero sin llegar a ser comparables)" (*íbid*.)

2. Corpora orais: Apesar do seu tímido progresso em termos de volume de corpora existentes e disponíveis, estes são um tipo de corpora empregados já nas primeiras aproximações baseadas em corpus. Pela sua natureza oral, este tipo de corpora resultam de mais difícil tratamento automaticamente computerizado, mas muitos são os esforços hoje em dia neles invertidos que fazem deste tipo de corpora uma fonte de recursos muito atractiva para muito diferentes aproximações.

2.1. Transcritos/não-transcritos: Dada a sua natureza oral, para facilitar o seu processamento automaticamente computerizado muitos corpora são transcritos graficamente.

2.1.1. Monologados: Estes são um tipo especial de corpora que contêm intervenções orais de um único interveniente.

2.1.1.1. Formais: Dependendo do tipo de situações em que se produzam as intervenções, estes corpora podem incluir textos

orais formais como as leituras em alto de escritos ou as palestras preparadas mas não escritas.

2.1.1.2. Menos formais: Entre os textos orais monologados que têm um carácter informal estão as leituras académicas, os comentários de notícias, as demonstrações, etc.

2.1.2. Dialogados: Estes são um tipo de corpora com um valor importante em aproximações nas quais se pretende estudar, desde uma óptica conversacional, as dinâmicas linguísticas, argumentativas, etc. dos intervenientes.

2.1.2.1. Formal: Ao igual que os monologados os corpora dialogados podem incluir textos orais que, mercê às situações contextuais em que têm lugar, têm um carácter formal, como por exemplo as conversas telefónicas entre desconhecidos, as entrevistas, os debates, os comícios, as transacções comerciais, etc.

2.1.2.2. Menos formal: Do mesmo jeito, entre os textos orais dialogados que têm um carácter informal estão as conversas telefónicas entre conhecidos, as conversas entre familiares, amigos, cônjuges e/ou colegas de trabalho, etc.

Dada a complicação deste tipo de ordenação numérica, a classificação de todos estes tipos de corpus pode ser, para uma melhor compreensão, sintetizada na seguinte tabela:

Tabela 1

<table>
<tr><td rowspan="7">Escritos (Sincrónicos ou Diacrónicos)</td><td rowspan="2">Monolingues</td><td colspan="3">Especializados</td></tr>
<tr><td colspan="3">Gerais</td></tr>
<tr><td rowspan="5">Multilingues</td><td rowspan="3">De tradução</td><td rowspan="2">Paralelos</td><td>Unidireccionais</td></tr>
<tr><td>Bidireccionais</td></tr>
<tr><td colspan="2">De traduções livres</td></tr>
<tr><td colspan="3">Comparáveis</td></tr>
<tr><td colspan="3">Com textos em diferentes idiomas</td></tr>
<tr><td rowspan="4">Orais (Transcritos ou Não-transcritos)</td><td rowspan="2">Monologados</td><td colspan="3">Formais</td></tr>
<tr><td colspan="3">Menos formais</td></tr>
<tr><td rowspan="2">Dialogados</td><td colspan="3">Formais</td></tr>
<tr><td colspan="3">Menos formais</td></tr>
</table>

Para finalizarmos este apartado, a definição que, em relação com o conceito de corpus, poderia, desde o meu ponto de vista, ser formulada é a seguinte:

> *Um corpus é uma colecção de textos naturais (electrónicos ou não), assumida, mercê à auto-avaliabilidade dos resultados do seu processamento, como representativa de toda uma língua ou de qualquer das*

suas variedades e construída, atendendo a diversos critérios de classificação textual, para a consecução de uns determinados objectivos.

4. Introdução aos corpora paralelos

Antes de nos debruçar de cheio na descrição do corpus paralelo criado para o desenvolvimento deste projecto de investigação, conviria que fossem feitas algumas considerações teóricas a respeito do *status*, história e finalidade dos corpora paralelos para que, desta maneira, possam ser compreendidas tanto as razões que levaram à escolha e compilação deste tipo de corpus como a metodologia empregada para tal efeito.

4.1. O que é um corpus paralelo?

Se durante a década de 90, à volta dos diferentes tipos de corpora multilingues, existia uma certa indeterminação terminológica devida, na altura, sem dúvida à novidade da sua aplicação, na actualidade esta indeterminação já foi superada. Desta maneira, quando hoje em dia se emprega a denominação *corpus paralelo* está-se a fazer referência a uma colecção de textos originais e as suas traduções noutra(s) língua(s).

Apesar do recente acunhação desta denominação terminológica e do recente emprego sistemático deste tipo de corpora tanto em projectos investigação em linguística contrastiva como em projectos de PLN ou como no desenvolvimento de tarefas docentes, a compilação de corpora paralelos é uma prática que se remonta muitos séculos atrás na História. Neste sentido, o mais antigo corpus paralelo de que se tem conhecimento é a pedra Rosetta, que se estima foi talhada à volta do ano 196 a.C. Esta pedra, cujos textos fazem referências às honras apresentadas ao rei Ptolomeu V pelos templos de Egipto, contém textos escritos em duas línguas –grego e egípcio–, escritos em três diferentes sistemas gráficos –os textos em egípcio foram escritos em hieróglifos e demótico– (*cfr*. Véronis, 2000: 1). Deste jeito, se fizermos um rastejo ao longo da história, veremos que a compilação deste tipo de corpora é uma prática comummente desenvolvida em muito diferentes âmbitos e momentos históricos. Assim, por um lado, muitos corpora paralelos foram construídos no processo de redacção de textos legais como, por exemplo, contratos e/ou pactos. Outros foram construídos, por outro lado, com a edição de versões bilingues de textos sagrados como, por exemplo, as versões bilingues da bíblia (latim/línguas romances) editadas na idade média. Por último, outros corpora paralelos foram abundantemente desenvolvidos à volta das versões bilingues de textos literários de muito diversos géneros.

De qualquer jeito, apesar desta tradição compilatória e da multiplicidade de âmbitos em que são potencialmente aplicáveis, só recentemente é que o emprego de corpora paralelos se encontra em auge. De facto, foi, para sermos exactos, a finais da década de 80 e princípios da de 90 no âmbito do PLN que os corpora paralelos passaram a ocupar o centro de atenção. Nesta altura, atingido um ponto de estancamento na aplicação dos modelos de interpretação das línguas naturais puramente baseados em regras os corpora começaram a ser empregados no desenvolvimento dos chamados modelos probabilísticos e conexionistas. Desta maneira, como mais à frente será pormenorizado, os corpora começaram a ser empregados, por um lado, como recursos de treino para o desenvolvimento de ferramentas automáticas de metodologia probabilística de anotação e de tradução e, por outro, como bases de dados reaproveitáveis para o desenvolvimento de técnicas de tradução baseadas em exemplos.

Tal e como anteriormente foi assinalado, para garantir a aplicabilidade dos corpora paralelos actualmente costuma-se implementar um processo de codificação mediante o qual fica explicitada a relação de correspondência de cada unidade de

tradução –seja esta uma palavra, um sintagma, uma oração ou um parágrafo– dos textos fonte e cada unidade de tradução dos textos traduzidos. Desta maneira, mediante este processo, hoje em dia conhecido como *alinhamento*, estabelece-se, pois, que a relação de correspondência existente entre estes dois segmentos é uma relação de equivalência de tradução.

Os acima referidos conceitos de unidade de tradução e equivalência, desenvolvidos dentro dos Estudos de Tradução, são conceitos derivados da concepção que neste tipo de estudos se tem sobre tipo de processos que constituem actos de tradução. Por tradução entende-se, pois, toda

> operación de transferencia interlingüística que consiste en interpretar el sentido de un <texto de origen> y producir un <texto de llegada> buscando establecer una relación de <equivalencia> entre ambos, de acuerdo con los parámetros inherentes a la comunicación y dentro de los límites de las restricciones impuestas al <traductor> (Dedisle, Lee-Jahnke & Cormier, 1999: 295).

Deste jeito, por um lado, por unidade de tradução entende-se o conjunto de elementos do texto de origem que possuem rasgos semânticos em comum e que o tradutor interpreta (*cfr*. *íbid*.: 304) reactivando uma série de conhecimentos extralinguísticos que, à hora de procurar uma equivalência, contribuem para a configuração do sentido (*cfr*. *íbid*.: 232). Por outro lado, por equivalência de tradução entende-se a relação de identidade estabelecida no discurso entre duas unidades de tradução de línguas diferentes, cuja função discursiva é idêntica ou quase idêntica (*cfr*. *íbid*.: 245). Sendo a tradução um actividade de mediação interlinguística em que devem ser tidos em conta aspectos de natureza discursiva e extralinguística –ou seja, contextal-situacional–, não surpreende, pois, que, subsequentemente, o conceito de equivalência seja um conceito relativo determinado, por um lado, pelas condições histórico-culturais em que os textos fonte são produzidos e os traduzidos são recebidos e, por outro, por uma série de factores de natureza linguístico-textual, igualmente histórico-culturalmente condicionados (*cfr*. Koller, 1995: 196).

4.2. O alinhamento

O alinhamento dos corpora paralelos, isto é, a explicitação das relações de correspondência entre unidades de tradução, pode ser realizado a muito diferentes níveis:

a) Alinhamento de palavras (w*ords*): a explicitação da relação de correspondência entre palavras é uma técnica mediante a qual são criadas bases de dados lexicográficas multilingues. Esta técnica foi já empregada nos começos da tradução automática no desenvolvimento de sistemas de tradução directa, isto é, palavra-por-palavra.

b) Alinhamento de segmentos (*segments*): esta técnica, de mais recente implementação, é actualmente empregada para superar, por um lado, as limitações das bases de dados lexicográficas multilingues, e, por outro, a falta de flexibilidade inerente ao alinhamento de orações.

c) Alinhamento de orações (*sentences*): este tipo de alinhamento, embora muito na moda nos estudos de tradução baseados em corpora, é um tipo de alinhamento ao qual, dada a extensão do tipo de unidades alinhadas, está inerentemente associada uma certa falta de flexibilidade.

d) Alinhamento estrutural (*structural*): se bem este tipo de alinhamento apresenta uma falta de flexibilidade muito mais agudizada do que o anterior, este é um tipo que, em certos casos, convém realizar para evitar, mercê ao ruído estrutural dos textos –notas a pé de página, títulos, capítulos, etc.– discrepâncias intertextuais.

Qualquer destes tipos de alinhamento pode ser levado a cabo bem de maneira manual, bem de maneira automática. Dado que, tal e como já foi anteriormente indicado no apartado 4.4., dedicado à anotação dos corpora, a cada uma destas metodologias está associada uma determinada margem de erro e inconsistência, a escolha entre elas dependerá em cada caso do tipo de textos que se pretendam alinhar, assim como do volume de unidades que se deseje processar. Se, por um lado, o alinhamento automático de textos muito rigidamente estruturados, independentemente do seu volume, oferece, em virtude da abundância de repetições de expressões formulaicas, garantias para a consecução de um alto grau de precisão, por outro, o alinhamento automático de textos muito criativos e expressivos não estará isento de problemas já que, em muitos casos, a correspondência não é 1:1. Desta maneira, o alinhamento manual, ou quando menos um certo grau de intervenção humana, tornar-se-ão neste caso necessários, tendo-se de inverter, pois, –se se tratar de grandes quantidades de volume– muito tempo no seu processamento. De qualquer forma, foram nos últimos anos desenvolvidas diversas metodologias que, bem individualmente bem combinadamente, são empregadas para a efectivação do alinhamento automático de textos fonte e traduzidos:

- Metodologia estatística: num modelo probabilístico combina-se a relação entre o número de caracteres da frase numa e na outra língua com o número de frases, dando *penalties* para a falta de probabilidade (*cfr*. Hallebeek, 1999: 8). Isto é, as frases mais longas num língua costumam, em síntese, corresponder-se com frases mais longas noutra língua, e viceversa.
- Metodologia de base linguística: um modelo de base linguística está baseado no reconhecimento e marcação de certos signos comuns aos textos das línguas envolvidas, chamados 'pontos de ancoragem' (em inglês *anchor points*), que servem para o reconhecimento de quais são os segmentos que numa e noutra(s) língua(s) são correspondentes. Assim, entre os tipos de signos que servem de pontos de ancoragem podemos encontrar, por exemplo, certas unidades léxicas, os sinais de pontuação, interrogação e exclamação, os nomes próprios, os parênteses, as aspas ou as expressões numéricas.
- Metodologias híbridas: para a superação, num certo sentido, das limitações dos modelos probabilísticos que, frente ao 98% de precisão que alcançam em textos estrutural e expressivamente rígidos, só logram em textos de expressividade e criatividade elevadas percentagens do 54%[2], e as dos modelos puramente baseados no emprego de pontos de ancoragem que, com certas garantias, só podem ser individualmente empregados para o alinhamento de corpora paralelos de reduzidas dimensões, a alternativa actualmente desenhada é a combinação dos dois modelos.

4.3. Porque os corpora paralelos?

Para que a definição de tradução anteriormente introduzida possa ser entendida fará falta necessariamente que se pense em termos de evolução de pensamento no âmbito da Linguística e da Teoria da Literatura. Em primeiro lugar, porque a consolidação dos Estudos de Tradução como um campo de trabalho autónomo da Linguística e da Teoria da Literatura só se tem começado a verificar desde há umas quatro décadas. Em segundo lugar, porque, estando este tipo de estudos fortemente influenciados pela Linguística, a posição marginal em que se encontravam dentro da Teoria da Literatura antes da sua actual consolidação era devida a existência da muito estendida ideia de que os textos traduzidos para uma língua constituíam uma deviação –isto é, não constituíam eventos

[2] Dados tirados de Hallebeek (1999: 8).

comunicativos genuínos– a respeito dos textos não-traduzidos existentes nessa mesma língua.

Neste sentido, no que diz respeito à Linguística, a evolução de pensamento, produzida só a partir das décadas de 60 e 70 após a emergência da Pragmática como disciplina autónoma dentro da Linguística, consistiu numa mudança epistemológica que, através de uma reformulação do objecto de estudo, permitiu a existência de perspectivas linguísticas, como a anteriormente exposta, que, ao lado dos factores mais puramente linguísticos, conseguem integrar nas suas teorizações semânticas toda uma série de factores de ordem extralinguística como, por exemplo, o contexto (situacional e discursivo) de actualização dos usos linguísticos, o conhecimento do mundo (sócio-histórico-culturalmente condicionado) ou a intencionalidade e estado de ânimo dos/as interlocutores/as. Desde esta perspectiva, a tradução pode, desta maneira, ser entendida como uma actividade de intermediação baseada no conhecimento (linguístico e extralinguístico) e, portanto, sujeita a condicionamentos de ordem sócio-histórico-cultural.

No que à Teoria da Literatura diz respeito, a evolução de pensamento veio da mão da conhecida como Teoria dos Polissistemas (Even-Zohar, 1990). Nesta teoria, desenvolvida desde uma focagem com uma forte componente sócio-cultural, a Literatura, entendida como um fenómeno semiótico, isto é, como um fenómeno de comunicação humana regido por signos, é concebida como um polissistema –incluído dentro do macro-sistema cultural– dentro do qual podem ser distinguidos um número variável de sistemas, igualmente compostos de um número variável de sub-sistemas, que poderíamos, *grosso modo*, fazer respectivamente corresponder com cada uma das tradicionalmente chamadas literaturas nacionais e com cada um dos géneros literários em vigor dentro destas. Dentro do macro-sistema literário, assim como dentro de cada sistema ou sub-sistema literário particular, a natureza dinâmica das relações (tensões) existentes entre os seus elementos determina que as posições (mais ou menos centrais ou mais ou menos periféricas) que cada elemento é susceptível de ocupar estejam sujeitas a uma contínua redefinição. Desta maneira, dentro desta teoria a literatura traduzida, entendida como um dos mais activos sub-sistemas existentes em cada sistema literário, não está desconectada da sua literatura original. Deste jeito, a posição da literatura traduzida dentro de cada sistema literário dependerá em cada momento histórico tanto da sua própria relação e posição a respeito doutros sistemas literários como do seu próprio ordenamento interno.

Em resumo, tendo em conta que a tradução é hoje em dia concebida como uma actividade baseada no conhecimento, não só de transferência interlinguística, mas também intercultural, através da qual são produzidos eventos comunicativos, embora de natureza distinta, igualmente genuínos aos eventos comunicativos de qualquer língua, pode-se afirmar que os corpora paralelos, sendo produtos desta actividade, são, em consequência, colecções de textos (fonte e traduzidos) escritos em diferentes línguas, nos quais ficam condensados os conhecimentos linguísticos e extralinguísticos necessários para a produção de eventos comunicativos equitativamente legítimos aos eventos originais da língua desde a qual tinha sido realizada a tradução.

4.4. O que é uma Memória de Tradução (MT)?

Durante a última década do s. XX, como aconteceu, em certo modo, a meados da década de 60 com a publicação do informe ALPAC, começou a ser sentido, incluso dentro do próprio campo da Tradução Automática, que a tradução automática não ia, pelo menos em várias décadas, conseguir solucionar de um jeito totalmente automático as necessidades de comunicação entre as diferentes línguas. Desta maneira, nessa mesma década de 90, imersa de cheio na lógica de mercado, a Engenharia Linguística, na qual

se tinha principiado a pensar que para a consecução de traduções que roçassem a aceitabilidade a intervenção humana resultava indispensável, começou a derivar para a filosofia de Tradução Assistida por Computador (TAC) e, em consequência, para a formação, dentro das empresas de software, de grupos de trabalho que priorizaram o desenvolvimento de aplicações, conhecidas como *estações de trabalho* (em inglês *workstations*) ou ambientes *integrados de tradução* (em inglês *translator's workbench*), que, no âmbito da Tradução Humana Assistida por Computador (THAC), estão baseadas no emprego das chamadas Memórias de Tradução. Deste jeito, as memórias de tradução, consistentes no alinhamento e anotação de corpora paralelos de textos originais e traduções realizadas manualmente e validadas pelo/a tradutor/a, permitem, na realização de novos projectos de tradução, a reutilização e aproveitamento, mercê à sua armazenagem em formato electrónico, de traduções humanas já efectuadas anteriormente.

4.5. O que é o TMX?

A proliferação nos últimos anos da década de 90 de aplicações implementadas desde diferentes empresas de software (IBM, Trados-Microsoft, Star e Atril) e dirigidas à criação, gestão e reutilização de memórias de tradução, levou associada uma importante incompatibilidade entre sistemas que trouxe consigo, inevitavelmente, a impossibilidade de partilha e reutilização destas ferramentas entre plataformas. Desta maneira em 2001, desde LISA[3], com o intuito de superar esta fase de incompatibilidade, foi desenvolvido o estândar conhecido como TMX (*Translation Memory eXchange*) que, mediante o emprego de etiquetas baseadas na metalinguagem XML, habilita uma anotação e armazenagem estandardizada das memórias de tradução.

4.6. Porque TMX?

Já que na actualidade o TMX é assumido pela maioria dos fabricantes e consumidores deste tipo de software como o formato de intercâmbio e migração entre sistemas de memórias de tradução, sem nenhuma ou quase nenhuma perda de informação, o seu emprego, dada a sua função estandardizadora, está justificado, pois o intercâmbio e reusabilidade efectivos das memórias de tradução, entendidas como corpora paralelos alinhados, anotados e validados, permite, mediante a sua aplicação no âmbito da tradução automática, a superação de fases pretéritas em que, nas operações de transferência tradutológica automatizada ou semi-automatizada, só se tinham em conta factores de ordem linguística.

5. Corpus Paralelo PALOP (português-espanhol) de Narrativa Pós-colonial (PALOP-PENP)

5.1. Desenho do corpus

Quando este trabalho de investigação foi iniciado, para a consecução do primeiro dos objectivos específicos nele pretendidos, isto é, a criação, tal e como foi indicado no Prefácio deste estudo, de um corpus paralelo de textos narrativos procedentes das Literaturas Africanas de Língua Portuguesa e as suas traduções para espanhol, a primeira das tarefas que teve de ser levada a cabo foi o desenho do próprio corpus.

[3] Organização sem ânimo de lucro que, desde uma filosofia baseada na cooperação global, promove, no âmbito da sociedade da informação e da comunicação, o respeito pela idiossincrasia linguística na tecnologia.

5.1.1. Critérios de selecção dos textos

Na delimitação, pois, do desenho do corpus, os critérios seguidos para a selecção dos textos nele incluídos foram, em consonância com as recomendações que fazem finca-pé no seu maior grau de objectivabilidade (*vid*. apartado 3), unicamente de tipo externo. Deste jeito, os critérios tidos em conta foram os seguintes:
- Modo: em relação a este critério, os textos seleccionados foram todos eles textos escritos.
- Línguas: o critério linguístico aqui manejado foi o da selecção, por um lado, de textos originalmente escritos em português e, por outro, de traduções destes mesmos textos escritas em espanhol.
- Origem: neste caso o critério empregado foi o da selecção de textos escritos por autores/as nascidos/as nalguma das ex-colónias portuguesas em África –Cabo Verde, Moçambique, Angola, Guiné-Bissau ou São Tomé e Príncipe– e que, em consequência, são considerados/as como autores/as pertencentes ao campo literário da Literaturas Africanas de Língua Portuguesa.
- Participação: em relação com este critério, ao serem seleccionados só textos escritos por um/a único/a autor/a, ficaram excluídas todo o tipo de antologias ou compilações de textos escritos por mais de uma pessoa.
- Estilo: seguindo a classificação tradicionalmente aceite no âmbito da Linguística de Corpus, os textos seleccionados foram unicamente textos escritos em prosa.
- Género: tendo-se escolhido unicamente textos escritos em prosa, foram, dentro deste critério, seleccionados unicamente romances e livros de contos escritos.
- Status do texto: dois tipos de textos foram seleccionados segundo este critério: textos originais e traduções.
- Ligação entre línguas: ao escolhermos para a compilação deste corpus originais numa língua e traduções destes noutra língua, o formato de ligação das línguas, desta maneira, atingido é o alinhamento em pares.
- Status da(s) língua(s): as línguas escolhidas, segundo este critério, foram associadas a um único tipo de texto. Desta maneira, enquanto os textos escritos em português são sempre originais, os textos escritos em espanhol são sempre traduções.

Contudo, tendo em conta a experiência acumulada durante o processo de construção deste corpus, caberia ainda indicar que na prática, devido às possibilidades reais de acesso aos textos, teve de ser aplicado, em última instância, um outro critério de selecção: o de disponibilidade. Este critério, relevante na compilação de corpora monolingues, torna-se no caso da compilação dos corpora paralelos duplamente importante, porque os/as compiladores/as hão-de dispor ao mesmo tempo de, no mínimo, duas versões de um mesmo texto escritas em diferentes línguas.

5.1.2. Corpus desenhado

O corpus desenhado segundo estes critérios, conhecido como Corpus Paralelo PALOP (português-espanhol) de Narrativa Pós-colonial (PALOP-PENP), é, pois, um corpus paralelo, unidireccionalmente alinhado, de textos narrativos (romances e livros de contos) originalmente escritos em português por um/a único/a autor/a filiado/a ao âmbito das Literaturas Africanas de Língua Portuguesa e as suas traduções para espanhol.

5.2. Composição e tamanho do corpus

Partindo, pois, do desenho realizado mediante a aplicação dos acima referidos critérios de selecção, dos vinte e um pares de textos susceptíveis de serem incluídos no corpus (*vid*. Apêndice 1), este conta na actualidade com os sete seguintes:

Autor	Português	Espanhol
Almeida, G.	(2001). *O Testamento do Sr. Napumoceno da Silva Araújo*. 4.ª edição. Lisboa: Editorial Caminho.	(2000). *El testamento del Señor Napumoceno da Silva Araújo*. Barcelona: Ediciones del Bronce. Tradução de Jordi Cerdà.
Lopes, B.	(1961). *Chiquinho*. 2ª edição. Lisboa: PRELO, Sociedade Gráfica Editorial, Lda.	(2003). *Chiquinho*. Barcelona: ElCobre Ediciones. Tradução de Lluís Agustí y Pere Comellas.
Couto, M.	(2000). *O último voo do flamingo*. 2.ª edição. Lisboa: Editorial Caminho.	(2002). *El último vuelo del flamenco*. Madrid: Santillana Ediciones Generales. Tradução de Mario Merlino.
	(1999). *Terra Sonâmbula*. 5.ª Edição. Lisboa: Editorial Caminho.	(2002). *Tierra sonámbula*. Madrid: Suma de Letras. Tradução de Eduardo Naval.
	(1995). *Vozes Anoitecidas. Contos*. 3.ª edição. Lisboa: Editorial Caminho.	(2001). *Voces anochecidas*. Tafalla: Txalaparta. Tradução de Andrés Salter Iglesias.
Pepetela	(1997). *Parábola do cágado velho.* 2.ª Edição. Lisboa: Publicações Dom Quixote.	(1999). *Parábola de la vieja tortuga*. Madrid: Alianza editorial. Tradução de Basilio Losada
Rui, M.	(1999). *Quem me dera ser onda*. 5ª Edição. Lisboa: Edições Cotovia.	(2000). *Si pudiera ser una ola*. Barcelona: Seix Barral. Tradução de Isabel Soler.

O número total de palavras contidas nestes sete alinhamentos é de 566.590, das quais 274.341 pertencem aos textos escritos em português e 292.249 aos textos escritos em espanhol.

5.3. Representatividade e amostragem do corpus

Tendo-se por objectivo a construção de um corpus paralelo de textos narrativos originalmente procedentes das Literaturas Africanas de Língua Portuguesa e as suas traduções para espanhol, a representatividade actualmente alcançada pelo corpus PALOP-PENP, ao não se pretender, mediante ele, representar por completo nenhum dialecto ou língua concreta, senão um tipo determinado de textos, terá necessariamente de ser medida em relação com o número total de textos que cumpram os critérios de selecção previamente delimitados. Desta maneira, se o número total de alinhamentos susceptíveis de serem incluídos no corpus é de 21 e se o número total de alinhamentos actualmente incluídos é de 7, pode-se afirmar que o grau de representatividade alcançado é do 30%.

No que diz respeito à amostragem indicar que, sendo, precisamente, a pretensão a representação de um tipo textos narrativos procedentes das Literaturas Africanas de Língua Portuguesa e as suas traduções para espanhol, cada um dos textos inseridos é uma unidade inteira, isto é, corresponde-se com o corpo íntegro do texto das obras publicadas em formato livro de papel, o qual não significa que tenham sido incluídos todos e cada um dos elementos presentes em cada livro. Neste sentido, enquanto, por um lado, foram incluídos os glossários, as dedicatórias, as notas a pé de página dos originais e as notas de tradução, ficaram, por outro, excluídos as notas biográficas, os direitos de copyright, os prefácios, as introduções, os índices, assim como outro tipo de

elementos cuja função era anunciar outros títulos já publicados ou de próxima publicação.

5.4. Direitos de copyright

Tanto internacionalmente como nacionalmente quase qualquer tipo de texto está protegido por toda uma série de direitos de *copyright* que dizem respeito à sua aquisição, edição, manipulação e exploração. Se, por um lado, na aquisição de textos orais procedentes de pessoas individuais a lei não costuma obrigar mais do que a petição directa de permissão às pessoas a partir das quais se pretendam obter as colecções de dados necessários para uma determinada investigação, para a compilação de textos escritos para os quais o seu *copyright* não tenha expirado e de textos orais procedentes doutro tipo de fontes, "compilers must seek and gain the permission of authors and publishers who hold copyright for a work" (Olohan, 2004: 76). Desta maneira, nas petições de permissões deverão, necessariamente, de estar indicados os seguintes aspectos: o propósito da compilação e uso dos textos do corpus, o tipo de pessoas que terão acesso aos textos do corpus e, por último, o seu grau de acessibilidade. (*cfr*. *ibid*.: 51)

Apesar, porém, da apriorística imposição legal a que, em princípio, se devem ater tanto os investigadores por conta própria como os integrados em projectos de investigação, a verdade é que a legislação vigente em diversos países do mundo não está completamente clara. No Reino Unido, por exemplo, a lei diz que "copyright is infringed where either the whole or a 'substancial part' of a work is used without permission, unless the copyright falls within the scope of one of the copyright exceptions" (website da Copyright Licensing Agency *apud ibid*.: 50). Quais são, então, estas excepções? Em princípio, os usos privados e/ou educativos de corpora, mercê ao seu *status* legal, são actividades que não hão de se ater aos direitos de copyright vigentes para os diferentes tipos de textos. Apesar desta excepcionabilidade teórica, na prática estes usos não estão totalmente isentos de limitações, pois, de facto, dados os importantes interesses económicos em causa nesta questão, a este tipo de usos é-lhe aplicado o internacionalmente conhecido como 'fair dealing'. "So it is probably within the scope of the above fair dealing exception to make single photocopies of short extracts of a copyright work for the purposes of research or private study" (*ibid*.).

Como pode ser induzido, a limitação está, pois, no facto de que só podem ser empregadas partes substanciais ou pequenos extractos dos textos e, portanto, o uso de textos completos precisa do beneplácito do(s) possuidor(es) dos direitos de copyright. Do meu ponto de vista, estas limitações resultam muito pouco precisas e levantam, quando menos, várias questões: quando constitui um extracto uma parte substancial de um texto? Estar-se-ia usando um texto completo se se omitissem, num corpus, certas partes de um livro como, por exemplo, o índice, o prólogo, etc.? Por outro lado, se a questão do copyright se torna escura e complicada no nível legal, à hora de realmente solicitar as permissões aos possuidores de um texto a situação torna-se labiríntica:

> The process is sometimes slow and tedious. It is not always clear who holds copyright; publishers are sometimes taken over by other companies; authors of articles sometimes prove difficult to trace; some works which are no longer subject to copyright in a particular edition are not free to be published in another. (Kennedy, 1998: 77)

Se se tratar de corpora paralelos de textos fonte e textos traduzidos, além dos direitos de copyright dos autores e dos editores dos textos originais haverão também de se ter em conta os dos tradutores e os dos editores dos textos traduzidos. Desta maneira, a questão torna-se, então, potencialmente mais problemática. Assim, por um lado, "Baker predicted a defensiveness on the part of translators, who, she believed, will be

reluctant to provide translations, suspecting that their translations, and they themselves, might be subject to criticism." (Olohan, 2004: 51). Por outro lado, da parte editorial "it may well be possible to obtain permission to use the source text but not the translation, or vice versa" (*íbid*.).

Tendo, finalmente, em conta que as petições de permissões a multinacionais e grandes editoras são frequentemente ignoradas (*cfr. íbid.*) e que, habitualmente, para o uso de muitos textos, dentro do mundo editorial mundial, há de se pagar um cânon para obter uma versão electrónica legal, embora não se tenha por objectivo a sua exploração comercial, a questão do copyright faz com que muitos investigadores individuais e alguns projectos de investigação que não podem contar com a infraestrutura necessária para, por um lado, levar a cabo este tipo de tarefas nem com o dinheiro suficiente para, por outro, pagar os direitos de uso de diferentes textos, se vejam obrigados a se arriscarem a construir corpora sem a garantia de que, com isso, não estejam a incorrer num delito.

No que a este corpus diz respeito, os textos nele incluídos são, como já foi indicado, textos literários para os quais existem direitos de copyright em vigor que proíbem a sua reprodução, registo ou transmissão mediante nenhum tipo de meio. Vejamos, como exemplo, os avisos legais incluídos na versão traduzida para espanhol de *O último voo do flamingo* de Mia Couto (2002: 6):

> Todos los derechos reservados. Esta publicación no puede ser reproducida, ni en todo ni en parte, ni registrada en o transmitida por, un sistema de recuperación de información, en ninguna forma ni por ningún medio, sea mecánico, fotoquímico, electrónico, magnético, electroóptico, por fotocopia, o cualquier otro, sin el permiso previo por escrito de la editorial.

Contudo, embora estes direitos estejam na actualidade vigentes para todos e cada um dos textos incluídos neste corpus, não foi, dentro do âmbito deste trabalho de investigação, solicitada nenhuma permissão nem aos/às autores/as, nem às editoriais nem, no caso dos textos traduzidos, aos/às tradutores/as. A razão é que, tal e como já foi referido, os direitos de copyright são uma questão legal cheia de lacunas de cujo cumprimento ficam, em princípio, isentos usos privados ou com fins educativos sem nenhum tipo, portanto, de ânimo de lucro. Desde este ponto de vista, o emprego que neste trabalho de investigação se faz dos textos incluídos no corpus pode ser considerado um uso privado, neste caso para investigação, sem ânimo de lucro.

Por outro lado, apesar da sua excepcionabilidade, este tipo de usos está, mesmo assim, sujeito a limitações impostas pelo já citado *fair dealing*. Deste jeito, incluso nestes casos, legalmente só está permitido o emprego de pequenos trechos ou extractos que em nenhum caso constituam uma parte substancial do texto. Sem uma definição mais pormenorizada, a interpretação a respeito do que constitui uma parte substancial é, quando menos, complicada. De qualquer jeito, como acima foi referido, os textos incluídos neste corpus não incluem as notas biográficas, os direitos de copyright, os prefácios, as introduções nem os índices.

Contudo, uma última e importante aclaração tem de ser feita já que, estando o corpus publicado na Internet (*vid*. apartado 7.10.) e sendo, pois, acessível via web para a realização de consultas, a gravidade dos prejuízos que poderiam ser causados e, portanto, dos delitos que, mediante este tipo de práticas, poderiam ser cometidos é considerável. Deste jeito, com o intuito de evitar potenciais acções legais, o corpus está de tal forma publicado que o número de frases devolvidas após uma consulta não constituam nunca a obra completa, senão que o sistema está limitado na sua versão pública a 150 frases.

5.5. Hardware e software

Por um lado, o hardware empregado para a captura, armazenagem e processamento dos textos seleccionados para serem incluídos neste corpus foi seguinte:

- Modelo de computador: Dell Inspiron 1150
- Processador: Intel ® Celeron ® 2.4 GHz
- Memória RAM: 256 MB
- Scanner: BENQ 5250C (1200 dpi/ 48-bit) com conexão USB 2.0
- Espaço empregado no disco rígido: 8.46 MB (os textos em formato documento de Word (.doc) sem nenhum tipo de anotação) e 10,7 MB (os textos em formato TMX com a informação anotada acerca do seu alinhamento e dos fenómenos tradutológicos presentes).

Por outro lado, para podermos realizar uma exposição detalhada dos programas necessários para a construção de um corpus das características do PALOP-PENP, convirá que distingamos dentro do próprio software empregado diferentes níveis:

- Sistema operativo: Microsoft Windows XP e Ubuntu-LINUX 5.04
- Captura dos textos: MiraScan 6 (5250C) e ABBY FineReader 5.0 Sprint (ORC software)
- Armazenagem e correcção ortográfica: Microsoft Office Word 2003 (SP1) e OpenOffice.org 1.1.3
- Alinhamento: TRANS Suite 2000 Align 1.4.2
- Edição de ficheiros TMX: XMLSpy 5.3 (Enterprise Edition)

5.6. Captura dos textos

Ao não ter sido solicitado nenhum tipo de permissão de emprego dos textos incluídos neste corpus, o acesso a versões já digitalizadas dos textos foi totalmente impossível e, por isso, resultou imprescindível levar a cabo um processo de captura dos textos em formato papel, para o qual a metodologia seleccionada, dados, por um lado, o elevado número de obras escolhidas para serem incluídas neste corpus e, por outro, as dimensões da sua amostragem, foi a digitalização mediante scanner.

O programa ABBY FineReader 5.0 Sprint é o software de reconhecimento óptico de caracteres proporcionado pela empresa BENQ com o seu scanner BENQ 5250C. Este programa, versão *share*, ou seja, versão gratuita e reduzida da versão comercial, permite, porém, o reconhecimento através de uma prévia digitalização mediante o software de escaneamento MiraScan 6 (5250C), dos caracteres de textos de muito diversas línguas, entre elas castelhano e português. A vantagem deste software é que, além de páginas individuais, possibilita, naqueles livros cujo tamanho se ajusta às dimensões do scanner, a digitalização de duas páginas ao mesmo tempo. Desta maneira, aqueles livros que ultrapassavam as dimensões do scanner tiveram de ser digitalizados página a página, o qual reduziu à metade a velocidade de escaneamento dos textos.

O processo de captura dos 14 textos incluídos neste corpus não esteve isento de problemas, como já foi indicado no apartado 4.3., dadas as limitações que esta metodologia tem associadas. Assim, além dos erros de reconhecimento já indicados naquele apartado (‘d’ por *cl*, ‘m’ por *in*, *ni* ou *ir* ou ‘c’ por *e*, entre outros), foram identificados outros erros próprios das línguas envolvidas neste processo de captura. No caso do português foram sistematicamente cometidos erros de reconhecimento relacionados, por um lado, com a acentuação gráfica como, por exemplo, ‘ã’ por *á* ou ‘ê’ por *é*, e, por outro, com os sinais de pontuação como, por exemplo, ‘.’ por ‘,’ ou ‘„’ por ‘...’. No caso do espanhol os erros sistematicamente cometidos tinham a ver, por um lado, com a manutenção, na versão traduzida, da acentuação do português e, por outro,

com os sinais de pontuação das orações interrogativas e exclamativas, cujas marcas iniciais eram sistematicamente omitidas pelo programa de OCR.

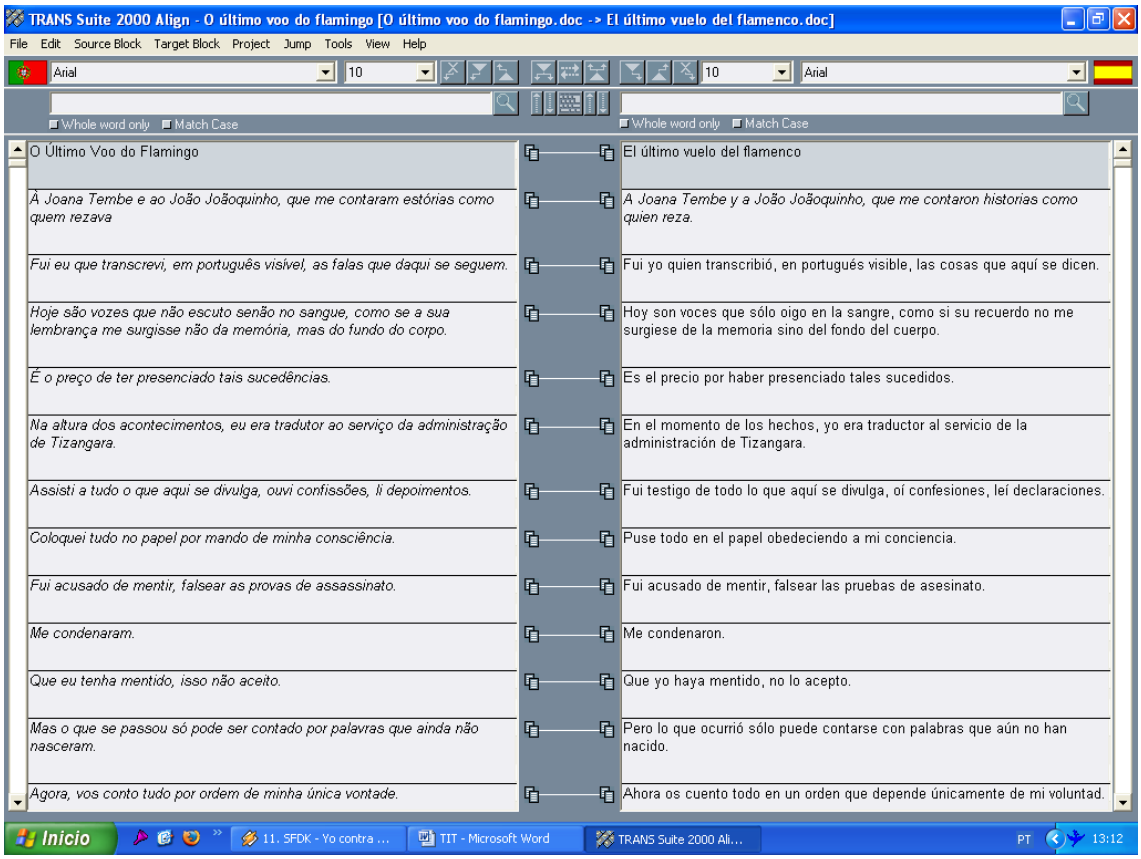
5.7. Armazenagem e correcção ortográfica dos textos

Uma vez digitalizados e reconhecidos opticamente, os textos foram subsequentemente armazenados em formato documento de texto (.doc) mediante o software específico de processamento de textos: Microsoft Office Word 2003 (SP1) e OpenOffice.org 1.1.3. Como, além de erros sistemáticos, outros erros circunstancialmente cometidos eram contidos pelos textos escritos em qualquer das duas línguas, estes tiveram de ser submetidos a um processo de correcção para o qual foram empregadas técnicas de busca e substituição e de correcção ortográfica automatizada. Apesar das vantagens que este tipo de técnicas trazem consigo, o processo de correcção, levado a cabo mediante os módulos de verificação ortográfica incluídos nos programas de processamento de textos, tem, de igual maneira, associada uma certa margem de erro determinada pela impossibilidade de identificação de erros de reconhecimento em palavras que os sistemas reconhecem como correctas. Deste jeito, uma palavra como *lá* em português que o software de OCR reconheceu como *lã*, não pode ser identificada pelos módulos de verificação ortográfica já que ambas são palavras existentes no seu dicionário. Assim, na prática a correcção ortográfica teve de ser também realizada através da leitura dos textos durante o processo de alinhamento dos mesmos.

5.8. Alinhamento dos textos

O processo de alinhamento dos textos, desenvolvido entre os meses de Março e Maio de 2005, foi levado a cabo mediante o programa TRANS Suite 2000 Align 1.4.2, versão *share* da versão comercial da falida empresa Cypresoft. Este programa, que emprega para a segmentação em frases pontos de ancoragem baseados nos sinais de pontuação, permite o alinhamento automático ou manual dos pares de textos escolhidos. Vejamos um exemplo:

Figura 1



Tal e como anteriormente foi indicado, o alinhamento automático dos textos leva associada uma certa margem de erro porque em muitos casos a correspondência entre unidades de tradução não é 1:1, senão n:1 ou 1:n. Desta maneira, para a criação dos alinhamentos a metodologia escolhida neste trabalho esteve baseada na combinação da técnica de segmentação em frases mediante pontos de ancoragem com a técnica de edição e alinhamento manual dos segmentos.

Além disso, frente à metodologia puramente automatizada, o alinhamento manual dos pares de textos foi preferido porque, dada a sua natureza mediadora, o processo de tradução leva implícitos, sobretudo no caso de textos literários, certos fenómenos tradutológicos considerados necessários pelos/as tradutores/as humanos/as para a representação numa língua e cultura diferente do sentido dos textos originais. Neste trabalho, seguindo as directrizes marcadas nos corpus paralelos do CLUVI construídos pelo Seminario de Linguística Informática da Universidade de Vigo (*cfr*. Gómez & Sacau, 2004: 2-5), os fenómenos tradutológicos considerados de interesse para a sua marcação manual foram os seguintes:

(i) Omissão: na omissão há uma parte do texto de partida que não tem correspondência no texto de chegada, isto é, uma frase ou uma parte de uma frase não é traduzida.
(ii) Adição: a adição implica a inserção de fragmentos no texto de chegada que não têm correspondência no texto de partida.
(iii) Reordenamento: o reordenamento implica deslocamentos de frases inteiras, ou movimentos de fragmentos de frases do original a outras frases na tradução.

(iv) Glossários e notas a pé de página dos originais e das traduções[4]: os glossários e as notas a pé de página introduzem, em paralelo ao texto, informação complementar necessária para a compreensão do próprio texto.

5.9. Edição e anotação dos ficheiros TMX

Uma vez completamente alinhado e fenomenologicamente marcado cada par de textos, as memórias de tradução, exportadas mediante o programa TRANS Suite 2000 Align 1.4.2 para o formato TMX, foram editadas para a sua optimização com o programa XMLSpy 5.3 Enterprise Edition (versão completa), especificamente destinado ao processamento de ficheiros XML, que permite tanto a edição do texto como das etiquetas automaticamente criadas mediante a exportação.

Desta maneira, os ficheiros TMX, em que cada uma das memórias de tradução está armazenada, precisaram da anotação dos fenómenos tradutológicos previamente identificados e marcados na anterior etapa de alinhamento manual dos pares de textos.

a) Omissão:

A omissão, fazendo uso da etiqueta **<hi>**, fechada com **</hi>**, cuja função é ressaltar porções de texto para algum fim específico, delimitado mediante o atributo **type**, foi codificada da seguinte maneira: **<hi type="supr">** … **</hi>**, onde "supr" significa 'omissão'. Vejamos um exemplo:

```
<tuv lang="PT-PT">
    <seg>
        <hi type="supr">Mamãe velha brigava sempre:</hi>
    </seg>
</tuv>
<tuv lang="ES">
    <seg/>
     </tuv>
```

b) Adição:

Fazendo uso igualmente da etiqueta **<hi>**, a adição foi codificada do seguinte jeito: **<hi type="incl">** … **</hi>** onde "incl" significa 'adição'. Vejamos um exemplo:

```
<tuv lang="PT-PT">
                <seg>Comeram, se deitaram com umas raparigas, dormiram, levaram um cabrito de
Mande como multa de qualquer falta incompreensível, e arrastaram também o pobre Ufolo nas
cordas.</seg>
</tuv>
<tuv lang="ES">
                <seg><hi type="incl">Mientras tanto,</hi> comieron, se acostaron con unas chicas,
        durmieron, se llevaron un cabrito de Mande como multa por alguna falta incomprensible, y
        arrastraron también al pobre Ufolo tirando de las cuerdas. </seg>
</tuv>
```

c) Reordenamento:

O reordenamento é anotado mediante a combinação da etiqueta **<hi>**, à qual se lhe acrescenta o atributo x, com o elemento **<ph>**, que significa "Place Holder" e cuja função é indicar o ponto do texto ao qual originalmente pertence um elemento

[4] Não aparecendo estes fenómenos originalmente delimitados no artigo que serve de base para esta classificação, neste trabalho considerou-se de interesse a conservação tanto dos glossários e notas a pé de páginas dos originais (embora não constituam fenómenos tradutológicos em si próprios) como dos glossários e as notas de tradução dos textos traduzidos.

reordenado. Desta maneira, esta combinação resulta na seguinte anotação: **<hi type="reord" x="n">…</hi>** …. **<ph x="n"/>**, onde n é um número associado para a marcação do ponto de origem do segmento reordenado. Vejamos um exemplo:

```
<tuv lang="PT-PT">
        <seg>-- A Muari tem sempre razão -- escarneceu Munakazi.-- </seg>
    </tuv>
    <tuv lang="ES">
        <seg>--Muari tiene siempre razón. <hi type="reord" x="2">--dijo rabiosa Munakazi.</hi>
        </seg>
    </tuv>
</tu>
<tu creationdate="20050316T155130Z" creationid="TS2!ALIGN" changedate="20050316T155130Z">
    <tuv lang="PT-PT">
        <seg>Fica então com a Muari.</seg>
    </tuv>
    <tuv lang="ES">
        <seg>Pues quédate con Muari <ph x="2"/>
        </seg>
    </tuv>
```

d) Glossários e notas a pé de página:

Tanto os glossários como as notas a pé de página receberam, no que ao seu *status* diz respeito, o mesmo tratamento ao se considerarem versões diferentes do mesmo tipo de fenómeno tradutológico.

Desta maneira, os glossários e notas a pé de página dos textos originais, fazendo uso da etiqueta **<hi>**, foram codificados da seguinte maneira: **<hi type="ndo">…</hi>**, onde "ndo" significa 'nota do original'. Vejamos alguns exemplos:

Exemplo 1 (glossário):

```
<tu creationdate="20050417T171552Z" creationid="TS2!ALIGN" changedate="20050417T171552Z">
    <tuv lang="PT-PT">
        <seg>
            <hi type="ndo">GLOSSÁRIO</hi>
        </seg>
    </tuv>
    <tuv lang="ES">
        <seg/>
    </tuv>
</tu>
<tu creationdate="20050417T171552Z" creationid="TS2!ALIGN" changedate="20050417T171552Z">
    <tuv lang="PT-PT">
        <seg>
                    <hi type="ndo">ALEMBAMENTO — Pagamento feito aos pais da noiva (em língua
Kimbundu).
</hi>
        </seg>
    </tuv>
    <tuv lang="ES">
        <seg/>
    </tuv>
</tu>
```

Exemplo 2 (nota a pé de página):

```
    <tu creationdate="20051004T133021Z" creationid="TS2!ALIGN" changedate="20051004T133021Z">
        <tuv lang="PT-PT">
            <seg>
            <hi type="ndo">(1) O Machimbombo: autocarro.</hi>
            </seg>
        </tuv>
        <tuv lang="ES">
            <seg/>
        </tuv>
    </tu>
```

Seguindo o mesmo modelo, os glossários e notas de tradução dos textos traduzidos foram codificados da seguinte maneira: **<hi type="ndt">…</hi>**, onde "ndt" significa 'nota da tradução'. Vejamos alguns exemplos:

Exemplo 1 (glossário):

```
<tu creationdate="20050417T171552Z" creationid="TS2!ALIGN" changedate="20050417T171552Z">
    <tuv lang="PT-PT">
        <seg/>
    </tuv>
    <tuv lang="ES">
        <seg>
            <hi type="ndt">Glosario</hi>
        </seg>
    </tuv>
</tu>
<tu creationdate="20050417T171552Z" creationid="TS2!ALIGN" changedate="20050417T171552Z">
    <tuv lang="PT-PT">
        <seg/>
    </tuv>
    <tuv lang="ES">
        <seg>
                    <hi type="ndt">Alembamento: Pago hecho a los padres de la novia (en lengua
      kimbundu).</hi>
        </seg>
    </tuv>
</tu>
```

Exemplo 2 (notas a pé de página da tradução):

```
<tu creationdate="20051004T133021Z" creationid="TS2!ALIGN" changedate="20051004T133021Z">
    <tuv lang="PT-PT">
        <seg/>
    </tuv>
    <tuv lang="ES">
        <seg>
            <hi type="ndt">* Bandos: designación popular de bandidos armados.</hi>
        </seg>
    </tuv>
</tu>
```

5.10. Implementação via web: crescimento e retroalimentação

A última das etapas de construção e processamento de qualquer corpus para o qual se pretenda, por um lado, evitar o seu desaproveitamento e, por outro, possibilitar a optimização da sua aplicabilidade, consiste na sua implementação via web. Deste jeito, se bem é certo que, tal e como assinala Diana Santos (*cfr*. 1999: 326), a implementação via web de um corpus não garante que este vá começar a ser imediatamente empregado por parte de potenciais investigadores/as, já que, além da sua publicação, é necessária uma reeducação que leve aqueles/as a uma mudança de hábitos de investigação, a mais iminente vantagem que este tipo de implementação oferece, ao contrário do que acontece na linguística de introspecção, é a possibilidade de avaliação, em relação a um padrão comum, de este tipo de ferramentas de análise, que, em definitiva, permita tornar mais objectivo o progresso na área das ciências humanas.

Além deste importante benefício, a implementação via web de corpora oferece diferentes vantagens em relação ao resto de elementos em causa no particular contexto comunicativo desenhado por este tipo de implementação. Assim, por um lado, no que diz respeito ao meio de comunicação, isto é, à Internet, resulta claro que qualquer corpus obtém, mediante a sua divulgação via web, um valor acrescentado, já que, dado o impressionante ritmo de crescimento e vulgarização nas sociedades ocidentais, a Internet constitui na actualidade uma das mais poderosas ferramentas de divulgação do conhecimento.

Por outro lado, no que aos emissores se refere, isto é, às instituições possuidoras de um corpus, a sua implementação via web, dada a sua natureza centralizadora, brinda

a importante vantagem de permitir a obtenção centralizada de *feedback*, mediante a qual, ao serem detectados os seus pontos fortes, fracos e/ou deficitários, podem ser efectuados de maneira unitária o processo de correcção das suas falhas, fazendo-o crescer, se fosse necessário, para equilibrá-lo, e, ao mesmo tempo, o processo de implementação de melhores ou mais novas funcionalidades, sem possibilitar, porém, o uso, por parte de vários usuários, de versões diferentes do corpus.

Por último, no que aos receptores diz respeito, ou seja, aos/às usuários/as do corpus, a implementação via web deste tipo de recursos traz, também, consigo importantes benefícios, já que, além de permitir desde o local de trabalho o acesso remoto a recursos localizados fisicamente longe, possibilita a minimização, em primeiro lugar, dos conhecimentos técnicos e, em segundo lugar, dos recursos tecnológicos (espaço de memória e requisitos do sistema operativo) necessários para a armazenagem, tratamento e exploração deste tipo de materiais (*cfr*. *íbid*.: 331-332).

De qualquer jeito, a implementação via web de ferramentas de análise como, por exemplo, os corpora não está isenta de problemas. Assim, por um lado, no que diz respeito às questões legais derivadas da sujeição dos textos a direitos de copyright, acontece com frequência que, mercê ao exagerado medo à facilitação de um potencial aproveitamento lucrativo, os donos dos textos (as editoras, os jornais ou os/as autores/as, entre outros) "fazem questão em restringir o acesso, quer por sobrevalorização do produto (textos), quer por ignorância em relação ao que está, de facto, em causa" (íbid.: 329). Com o intuito de superar esta dificuldade, merece ser destacado, porém, que o tipo de implementação via web actualmente existente, baseado na filosofia cliente-servidor (client-server), está perfeitamente capacitado para a restrição, por parte das instituições possuidoras de uma determinada ferramenta, do seu acesso através da Internet. Deste jeito, os/as usuários/as de uma versão publicada na Internet podem ser facilmente impedidos de ter acesso ao corpus inteiro. Por outro lado, no que à partilha de recursos se refere, não resulta estranha, desde uma mentalidade um tanto egoísta, a afirmação, por parte das instituições possuidoras, de que a possessão de um determinado recurso, com que outras instituições não contam, redunda num benefício "para a renovação (ou obtenção) de financiamento" (*íbid*.: 330). Se isto poderia ser possível desde um ponto de vista empresarial, a partilha de recursos, no que à pura investigação diz respeito, é uma condição, desde o ponto de vista científico, indispensável para garantir um equitativo progresso da Humanidade. Neste sentido, para afiançar, no caso da investigação académica, hábitos colaborativos que conduzam à partilha deste tipo de recursos, as instituições públicas concessoras de partidas orçamentais deveriam, antes do que a criação de recursos para uso exclusivo, premiar todos aqueles esforços que visem a sua implementação via web, já que, tal e como foi indicado, a Internet é na actualidade uma das mais poderosas ferramentas de divulgação do conhecimento.

No que a este corpus diz respeito, indicar, por um lado, que tendo-se até agora alcançado uma representatividade do 30% nesta primeira versão do corpus PALOP-PENP, a intenção dentro deste projecto é que o corpus continue a crescer com a adição de novos alinhamentos de textos que cumpram os critérios acima referidos, até se conseguir uma representatividade de mais de 50%. Por outro lado, cabe também destacar que, sendo a actual versão do corpus consultável via web na secção do CLUVI do Seminario de Lingüística Informática da Universidade de Vigo (http://sli.uvigo.es/CLUVI/)[5], neste projecto previu-se que a retroalimentação fosse realizada mediante o sistema de buscas implementado no citado endereço web e o sistema de correio electrónico do Seminário de Lingüística Informática, no qual se centralizam os serviços da secção do CLUVI.

[5] Consultado por última vez a dia 30 de Dezembro de 2006.

Alem disto, no que ao próprio sistema de buscas implementado via web diz respeito, vários aspectos têm ainda de ser ressaltados. Assim, por um lado, cabe indicar que o sistema implementado na secção do CLUVI do SLI, possibilitando a realização de buscas complexas mediante o emprego de diferentes símbolos de apoio –que podem ser consultados na sua página web–, está, contudo, especialmente desenhado para a obtenção unicamente de concordâncias entre pares de segmentos. Por outro lado, cumpre também apontar que este sistema de buscas é facilitado, em base à filosofia cliente-servidor, através de um servidor Apache montado num PC com sistema operativo Linux, que emprega, para as buscas, um formulário programado em PHP e que, finalmente, devolve um número do ocorrências inferior a 150 para não incorrer, como anteriormente foi indicado, na violação dos direitos de copyright dos textos.

5.11. Aplicabilidade do corpus

A aplicabilidade do corpus PALOP-PENP encontra, dada a natureza da sua actual versão, um campo no qual o seu emprego pode, do meu ponto de vista, ser considerado prioritário. Estou-me a referir ao da investigação, da didáctica e do exercício profissional da tradução humana.

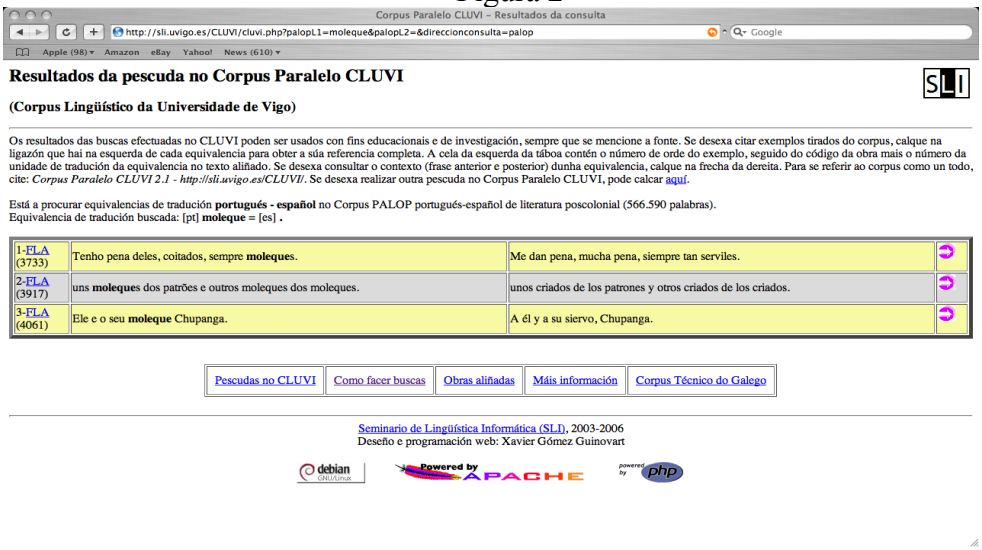
A aplicabilidade deste tipo de corpus no campo da investigação, didáctica e exercício profissional da tradução humana está, do meu ponto de vista, totalmente justificada porque tanto a anotação fenomenológica codificada como a implementação do anteriormente citado sistema de buscas via web garantem a obtenção de resultados, independentemente da sua finalidade, úteis para os/as tradutores/as interessados/as neste tipo de textos específicos.

As principais dificuldades que podem ser encontradas à hora de traduzir textos das Literaturas Africanas em Língua Portuguesa a espanhol as principais dificuldades que poderão ser encontradas são: a presença da componente léxica das línguas autóctones de cada uma das regiões das ex-colónias portuguesas da África, a remissão constante para a oralidade e, especialmente no caso de Mia Couto, a invenção de palavras com o intuito de lhe dar mais expressividade aos textos.

Vejamos, então, algum exemplo de busca para podermos entender o tipo de resultados que, desta maneira, podem ser obtidos:

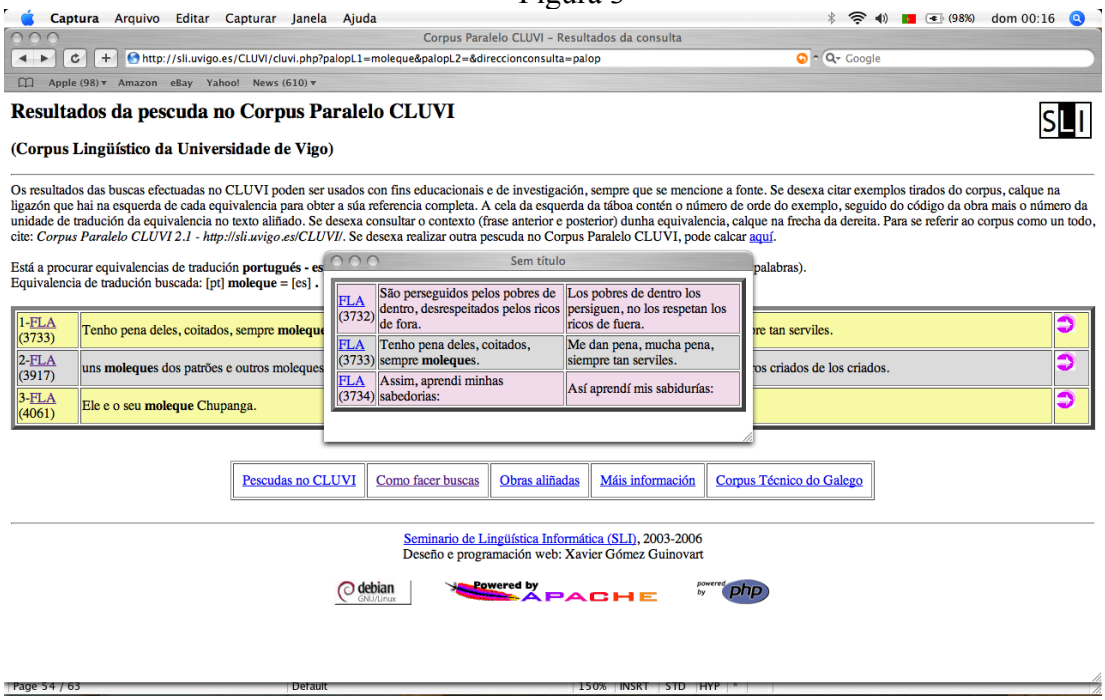
- Componente léxica das línguas autóctones:
  i) Texto da busca: moleque

Figura 2



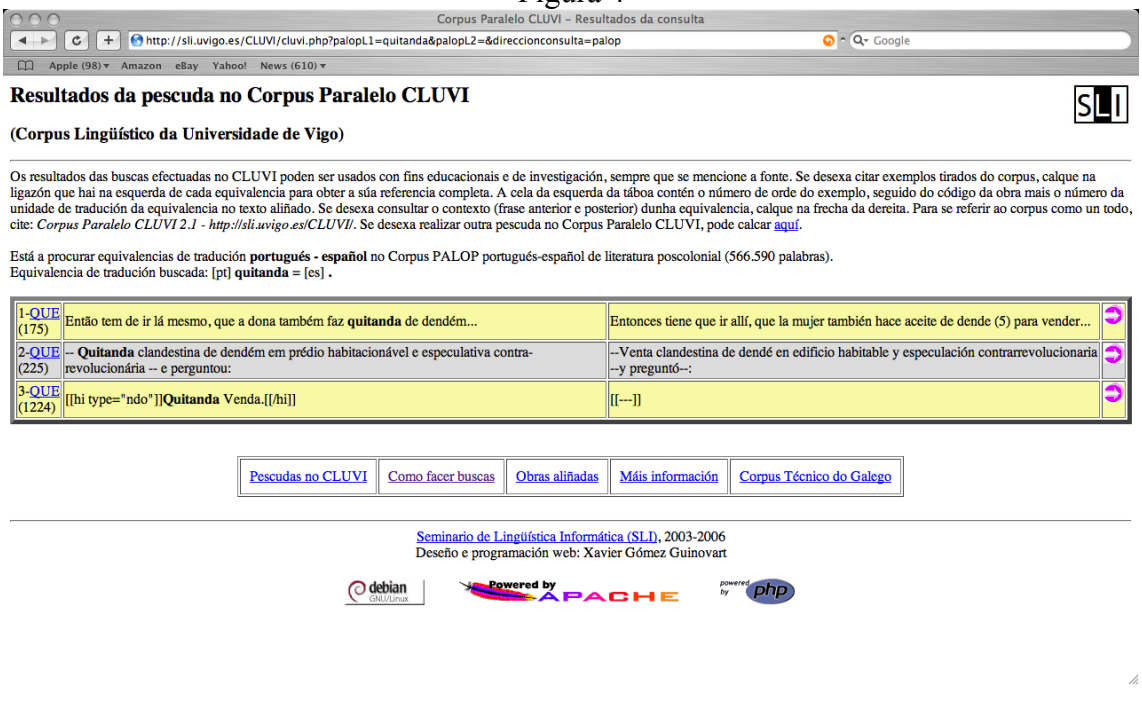
Como pode ser observado, este tipo de busca, que permite localizar uma sequência de palavras nalgum dos textos, devolve os segmentos em que a sequência foi localizada acompanhada da sua unidade de tradução. Além disso, se fizermos um clique nas setas localizadas à direita dos alinhamentos, podem ser também consultados, junto com o alinhamento em questão, os segmentos anterior e posterior:

Figura 3



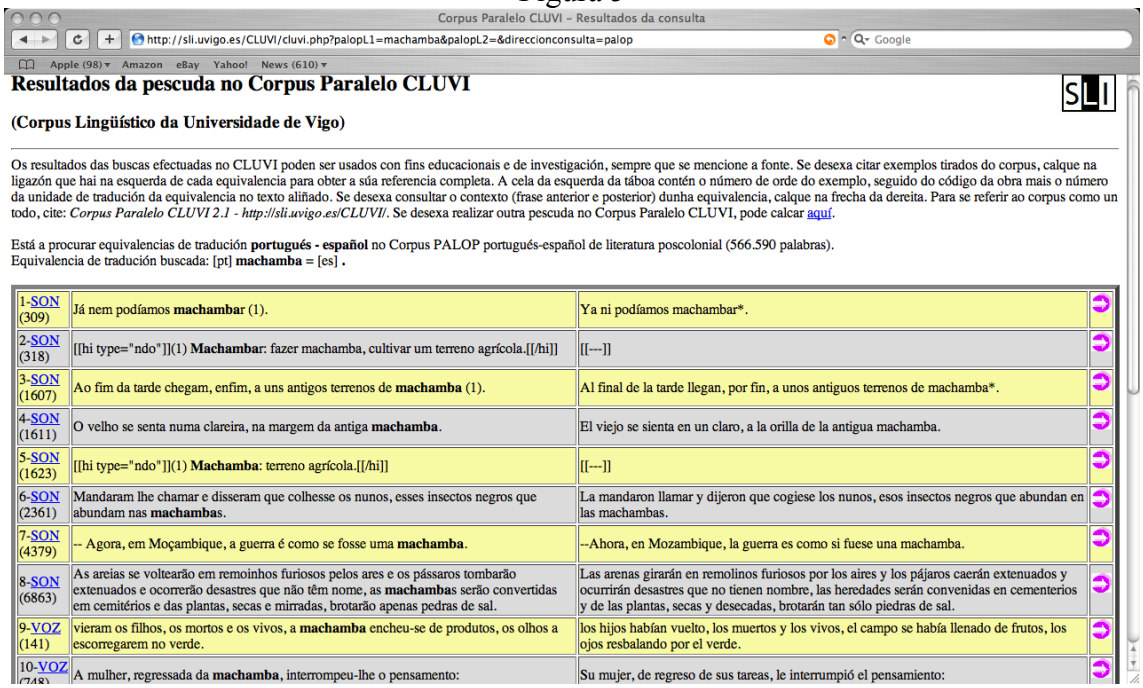
ii) Texto de busca: quitanda

Figura 4



iii) Texto de busca: machamba

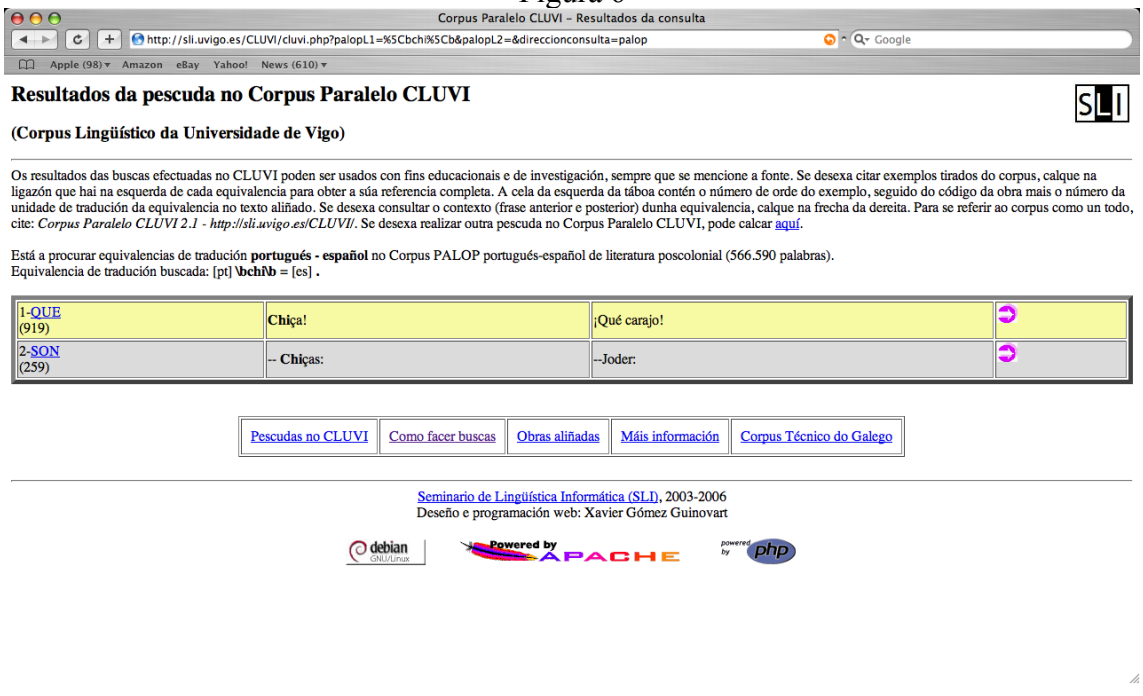
Figura 5



– Remissão para a oralidade:

i) Texto de busca: chiça

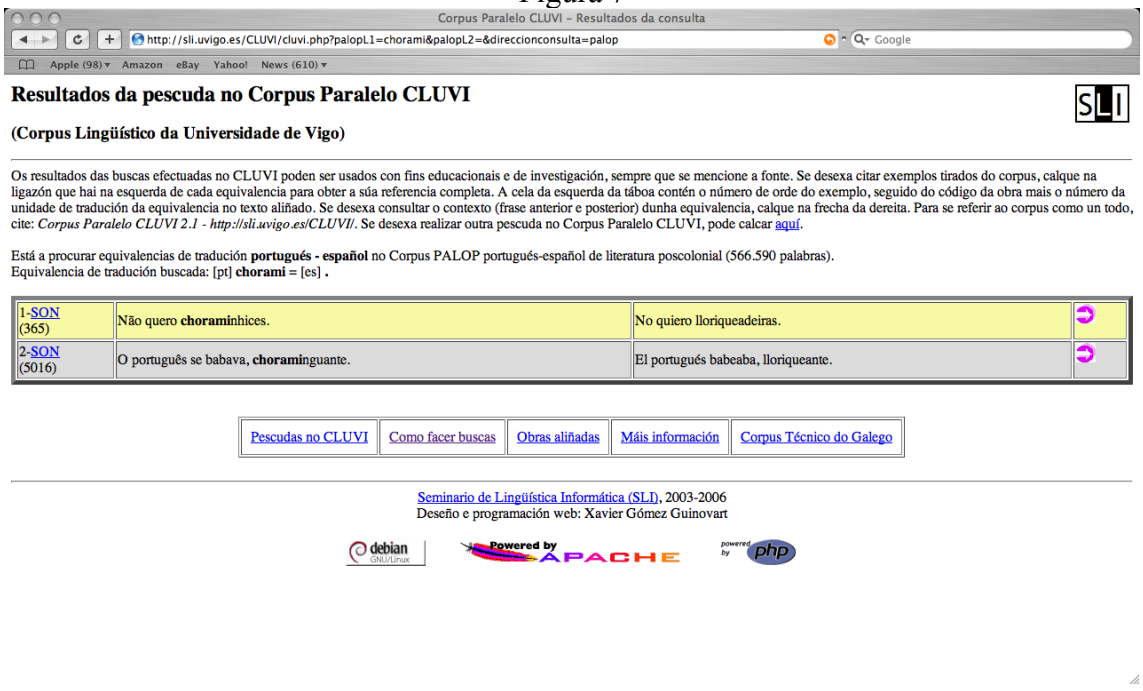
Figura 6



– Invenção de palavras:

i) Texto de busca: chorami*

Figura 7



Para finalizar este apartado referido à aplicabilidade gostaria de acabar dizendo que, além da aplicabilidade que hoje em dia pode ser atingida neste campo, espera-se no futuro poder chegar a implementar nos ficheiros TMX, em que está armazenada cada memória de tradução, um tipo de cabeceira TEI que permita a realização de buscas por nome e nacionalidade do/a autor/a e/ou por obra.

6. Conclusões

Uma vez finalizada a exposição pormenorizada do processo de construção do corpus PALOP-PENP e das suas aplicabilidades, conscientemente delimitadas, proceder-se-á, para finalizarmos este trabalho, à realização, a modo de síntese, de um repasso das mais importantes questões que, para chegarmos a este ponto, necessariamente tiveram de ser discutidas.

Desta maneira, tal e como pôde ser comprovado ao longo deste trabalho, o papel central que na actualidade jogam, ou quando menos podem potencialmente jogar, os corpora no desenvolvimento da actividade investigadora, industrial ou laboral de muito diferentes tipos profissionais, contrasta, ao mesmo tempo, com a ainda vigente diversidade de perspectivas acerca do *status* da chamada Linguística de Corpus. Por isso, para uma correcta interpretação do trabalho desenvolvido dentro deste projecto foi necessário, antes de mais, levar a cabo um trabalho prévio de aclaração terminológica baseado na realização de um repasso das diversas metodologias que ao longo da história da Linguística foram empregadas, assim como dos mais importantes acontecimentos que tiveram lugar no seio das ciências e sociedades ocidentais e que, do meu ponto de vista, levaram de maneira unívoca para a consideração da actual Linguística de Corpus como a moderna e aperfeiçoada versão das diversas metodologias de base empírica que foram ensaiadas nalgum momento da história da Linguística. Os corpora, concebidos, desde este ponto de vista, como colecções representativas de dados linguísticos naturais que, pela sua natureza empírica, servem, independentemente do seu formato de armazenamento, para o estudo de fenómenos e usos concretos e reais das línguas, tiveram de superar, no seu processo de consolidação como ferramentas de análise linguística legítimas, uma longa fase de ostracismo, mercê às duras críticas vertidas por

Chomsky desde a linguística generativa no que diz respeito à sua natureza empírica e à sua potencial representatividade. Neste sentido, se a consolidação de novas disciplinas dentro da Linguística –como a Pragmática ou a Sociolinguística– permitiu o assentamento, no seio da própria Linguística, da ideia, frente às concepções estruturais ortodoxas derivadas da linguística saussureana, de que o estudo da *Parole* –isto é, dos usos linguísticos concretos– é necessário para a consecução de um completo conhecimento do funcionamento complexo das línguas naturais, a invenção e o rápido desenvolvimento dos computadores –que permitiram não só um mais sistemático e rápido processamento dos dados contidos nas colecções textos, senão a construção de imensos corpora– acabaram definitivamente por tornar inválidas às críticas relacionadas com a impossibilidade de chegar, mediante o emprego deste tipo de colecções, a qualquer tipo de grau de representatividade. Deste jeito, os corpora foram, nos últimos anos, ocupando posições cada vez mais centrais nas metodologias de investigação linguística, até se converterem na actualidade em ferramentas indispensáveis para o desempenho de muitas tarefas, anteriormente impensáveis, em muitas disciplinas linguísticas.

De qualquer jeito, tal e como já foi discutido, a construção e implementação de corpora de dados linguísticos –escritos ou orais– não estão, mesmo de um ponto de vista de pura investigação, isentos de complicações, já que a vigência de direitos de copyright, tanto para a maioria dos textos como para muito do software necessário para o seu tratamento, obriga à petição de permissões, em muitas ocasiões sujeitas a algum tipo de pagamento. De facto, dada a insuficiente dotação económica de muitos projectos de investigação, assim como as limitações aquisitivas de muitos/as investigadores/as por conta própria, as complicações derivadas dos direitos de copyright vêm, na maioria dos casos, da mão da impossibilidade de fazer frente às importantes somas de dinheiro que teriam de ser pagadas para não incorrer em nenhum tipo de delito. Contudo, a ambiguidade da legislação vigente em matéria de copyright, assim como o espectacularmente rápido desenvolvimento experimentado nos últimos anos no campo do software livre, conseguiram manter ainda abertas, na actualidade, multidão de portas para o emprego de corpora que com certeza não podem ser desaproveitadas.

Apesar de que, durante o desenvolvimento do processo de descrição da construção do corpus PALOP-PENP, a ampla e diversa aplicabilidade do conjunto de corpora descritos neste trabalho, limitou-se ao campo da Tradução Humana –nas suas modalidades investigadora, didáctica e profissional–, a aplicabilidade que o corpus PALOP-PENP pode encontrar noutros campos não fica, por isso, nem muito menos reduzida. Deste jeito, o emprego, dentro da Linguística, de corpora multilingues pode perfeitamente encontrar, desde uma perspectiva teórica, aplicada ou didáctica, um lugar em campos tão diversos como a Sociolinguística, a Pragmática, a Análise do Discurso, a Lexicografia, a Dialectologia, a Linguística Contrastiva, a Linguística Geral ou, incluso, em âmbitos do PLN, como a Tradução Automática. Por outro lado, dado que o corpus está composto de textos literários pertencentes as Literaturas Africanas de Língua Portuguesa, pode-se, assim mesmo, afirmar que fora da Linguística existem, igualmente, outros campos –como, por exemplo, a Teoria da Literatura ou os próprios estudos literários das Literaturas Africanas de Língua Portuguesa– em que este corpus pode ser aplicado.

Portanto, tendo em conta que anteriormente não existia nenhum corpus paralelo (português-espanhol) especialmente dedicado ao campo das Literaturas Africanas de Língua Portuguesa, pode-se, a modo de conclusão, afirmar que o corpus PALOP-PENP representa, dado o seu carácter novidoso, uma interessante achega que enriquecerá não só o campo dos Estudos de Tradução, senão que, ao vir a ocupar um lugar anteriormente vazio dentro do mundo dos corpora, poderá resultar atractivo para muitos/as outros/as investigadores/as, entre os quais possivelmente se encontrem pessoas que nunca tinham

empregado este tipo de recursos. Desde este ponto de vista, o corpus PALOP-PENP representa, pois, mais uma contribuição para a consolidação de metodologias baseadas na exploração de colecções de dados linguísticos empíricos.

7. Bibliografía

Abaitua Odriozola, J.K. (2001). "Memorias de tradución en TMX compartidas por Internet". em *Revista Tradumàtica*. 0 (http://www.bib.uab.es/pub/tradumatica/15787559n0a9abaitua.pdf).

Abaitua Odriozola, J.K. (1997). "Traducción automática: Presente y futuro (1ª versión)". em www.foreignword.com (http://www.foreignword.com/es/Technology/art/Abaitua/Abaitua_3.htm).

Abaitua Odriozola, J.K. (2002). "Tratamiento de Corpora Bilingües". em Mª.A. Martí & J. Llisterri (eds.). *Tratamiento del lenguaje natural*. Barcelona: Edicions Universitat de Barcelona. pp. 61-90.

Abaitua Odriozola. J.K.; A. Casillas Rubio & R. Martínez Unanue (1997). "Segmentación de Corpus Paralelos para Memorias de Traducción". em *Procesamiento del Lenguaje Natural*. 21. pp. 17-30 (http://www.sepln.org/revistaSEPLN/revista/21/21-Pag17.pdf).

Agencia Española del ISBN http://www.mcu.es/bases/spa/isbn/ISBN.html (1-10-2004).

Aguirre Moreno, J.L.; N. Andino Rodríguez & X. Gómez Guinovart (2001). "Aspectos ortográficos, léxicos y morfosintácticos del etiquetado lingüístico de un corpus de informática en lengua gallega". em *Procesamiento del Lenguaje Natural*. 27. pp. 1-7. (http://www.sepln.org/revistaSEPLN/revista/27/27-articulo1.pdf).

Aguirre Moreno, J.L.; A., Álvarez Lugrís; I. Bragado Trigo; L. Castro Pena; X. Gómez Guinovart; S. González Lopo; A. López López; J.R. Pichel Campos; E. Sacau Fontenla & L. Santos Suárez (2003). "Alinhamento e etiquetagem de corpora paralelos no CLUVI (Corpus Lingüístico da Universidade de Vigo)". em J.J. Almeida (ed.). *Actas do Worksohp CP3A 2003, Corpora Paralelos: Aplicações e Algoritmos Associados*. Braga: Universidade do Minho. (http://webs.uvigo.es/sli).

Aijmer, K. & B. Altemberg (eds.) (1991). *English Corpus Linguistics: studies in honour of Jan Svartvik*. London/New York: Longman.

Allen, J. (1995). *Natural Language Understanding*. Redwood City: Benjamin/Cummings.

Almeida, J.J. (ed.) (2003). *Actas do Worksohp CP3A 2003, Corpora Paralelos: Aplicações e Algoritmos Associados*. Braga: Universidade do Minho.

Alonso Martín, J.A. (2003). "La traducción automática". em Mª.A. Martí Antonín (cord.). *Las tecnologías del lenguaje*. Barcelona: Editorial UOC. cap. 4.

Álvarez Lugrís, A. (1996). "Informática e análise textual". em X. Gómez Guinovart & A.M. Lorenzo Suárez (eds.). *Lingüística e informática*. Santiago de Compostela: Tórculo Edicións. pp. 87-152.

Álvarez Lugrís, A. (2001). "Creación e Explotación de Corpus Bilingües." em A.L. Soto Vázquez (ed.), B. Crespo & P. Cancelo López (eds. associados). *Insights into Translation*.Vol. III. A Coruña: Tórculo. pp. 185-201.

Aarts, J. (1991). "Intuition-based and Observation-based Grammar". em K. Aimer & B. Altemberg (eds.). *English Corpus Linguistics: studies in honour of Jan Svartvik*. London/New York: Longman.

Atkins, S.; J. Clear & N. Ostler (1992). "Corpus Design Criteria". em *Journal of the Association for Literary and Lingistic Computing*. 7:1. pp. 1-16.

Baker, M. (1993). "Corpus Linguistics and Translation Studies. Implications and Applications" em M. Baker; G. Francis & E. Tognnini-Bonelli. *Text and Technology*. *In Honour of John Sinclair*. Philadelphia/Amsterdam: John Benjamins. pp. 232-250.

Baker, M. (1995). "Corpora in Translation Studies: An Overview and Some Suggestions for Future Research". em *Target*. 7:2. pp. 223-243.

Baker, M. (1999). "The Role of Corpora in Investigating the Linguistic Behaviour of Professional Translators". em *International Journal of Corpus Linguistics*. 4:2. pp. 281-298.

Baker, M.; G. Francis & E. Tognnini-Bonelli (1993). *Text and Technology*. *In Honour of John Sinclair*. Philadelphia/Amsterdam: John Benjamins.

Beaugrande, R. De. (1999). "Reconecting Real Language with Real Texts: Text Linguistics and Corpus Linguistics." em *International Journal of Corpus Linguistics*. 4: 2. pp. 243-259.

Beaugrande, R. De; A. Shunbaq & M.H. Heliel (1994). *Language, Discourse and Translation in the West and Middle East*. Amsterdam/Philadelphia: John Benjamins.

Bright, W. (ed.) (1990). *The Collected Works of Edward Sapir, V: American Indian Languages (1)*. Berlin: Mouton de Gruyter.

Caravedo, R. (1999). *Lingüística del corpus. Cuestiones teórico-metodológicas aplicadas al español*. Salamanca: Universidad de Salamanca.

Čermák, F. (2002). "Today's corpus linguistics. Some open questions". em *International Journal of Corpus Linguistics*. 7:2. pp. 265-282.

Chesterman, A. & R. Arrojo (2000). "Shared Ground in Translation Studies". em *Target*. 12:1. pp. 151-160.

Chomsky, N. (1962). palestra dada na University of Texas 1958, 3rd Texas Conference on Problems of Linguistic Analysis in English. Austin: University of Texas.

Chomsky, N. (1970). *Aspectos de la teoría de la sintaxis*. Madrid: Aguilar.

Chomsky, N. (1984). *Modular Approaches to the study of the Mind*. San Diego: San Diego University Press.

Church, K.W. & R.L. Mercer (1993). "Introducing to the Special Issue on Computational Linguistics Using Large Corpora". em *Computational Linguistics*. 19:1. pp. 1-24.

Codesido Garcia, A.I. (sem publicar). "Tecnologias de la lengua y Traducción: recursos y herramientas". em M. Cal; P. Núñez & I. Palacios (eds.). *Aplicaciones de las nuevas tecnologías a la Lingüística, Traducción y Enseñanza de lenguas*. Santiago de Compostela: Univrsidade de Santiago de Compostela.

Copyright Licensing Agency em http://www.cla.co.uk (09-06-2005).

Coseriu, E. (1988). *Sincronía, Diacronía e Historia. El problema del cambio lingüístico*. 3ª edición. Madrid: Gredos.

Delisle, J.; C. Lee-Jahnke & M. Cormier (1999). *Terminologie de la Traduction. Translation Terminology. Terminología de la Traducción. Terminologie der Übersetzung*. Amsterdam/Philadelphia: John Benjamins.

Derose, S. (1999). "XML and the TEI" em *Computers and the Humanities*. 33:1-2. pp. 11-30.

Díaz Labrador, J.; I.J. Taquet; F. Quintana Hernández; J. Abaitua Odriozola; G. Araolaza & G. Barrutieta Anduiza (2003) "Gestión de traducciones mediante metadatos TEI y XLIFF". apresentado em *CLiP 2003. Computers, Literature and Philology*. *Università degli Studi di Firenze, Florencia (Italia), 4 a 6 de diciembre de 2003*. (http://www.deli.deusto.es/AboutUs/Publications/CLIP_03_articulo.pdf).

Dini, L. & V. Di Tomaso (1998). “Corpus Linguistics for Application Development”. em *International Journal of Corpus Linguistics*. 3:2. pp.305-318.
Doorslaer, L. van (1995). “Quantitative and Qualitative Aspects of Corpus Selection in Translation Studies” em *Target*. 7:2. pp. 245-260.
EAGLES Intering Report (1994). 2.1 “Corpus Typology, a Framework for Classification” (http://www/ilc/pi/cnr.it/EAGLES96/coprustyp/).
Escandell Vidal, M.V. (2002). *Introducción a la Pragmática*. Barcelona: Ariel.
Even-Zohar, I. (1990 [1972]). "Polysystem Theory". *Poetics Today*. 11:1. pp. 9-26 (http://www.tau.ac.il/~itamarez/works/papers/trabajos/ps-th_s.htm).
Even-Zohar, I. (1999 [1990]). "La posición de la literatura traducida en el polisistema literario". em M. Iglesias Santos (org.). *Teoría de los Polisistemas*. Madrid: Arco/Libros. pp. 223-231. (http://www.tau.ac.il/~itamarez/works/papers/trabajos/poslit_es.htm).
Ferreira, M. (1976). “A Literatura Africana de Expressão Portuguesa. Uma Literatura Ignorada”. em *Revista de la Universidad Complutense*. XXV:103. pp. 231-254.
Fernández Casas, Mª. X. (2004). *Edward Sapir en la Lingüística actual. Líneas de continuidad en la historia de la Lingüística*. Anexo 54. *Verba, Anuário Galego de Filología*. Santiago de Compostela: Universidade de Santiago de Compostela.
Fernández Pérez, M. (1986). *La Investigación Lingüística desde la Filosofía de la Ciencia. (A propósito de la lingüística chomskiana)*. Anexo 28. *Verba, Anuário Galego de Filología*. Santiago de Compostela: Universidade de Santiago de Compostela.
Fernández Pérez, M. (1996). “El Campo de la Lingüística Aplicada. Introducción”. em M. Fernández Pérez (coord.). *Avances en Lingüística Aplicada*. Santiago de Compostela: Universidade de Santiago de Compostela. pp. 11-45.
Fernández Pérez, M. (coord.) (1996). *Avances en Lingüística Aplicada*. Santiago de Compostela: Universidade de Santiago de Compostela.
Firth, J.R. (1957). *Papers in Linguistics 1934-1951*. Oxford: Oxford University Press.
Francis, W. N. (1982). “Problems of assembling and computerizing large corpora”. em S. Johanson (ed.). *Computer corpora in English language research*. Bergen: Norwegian Computing Centre for the Humanities.
Francis, W. N. (1992). “Language Corpora B.C.” em Svartvik, J. *Directions in Corpus Linguistics*. *Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991*. Berlin/New York: Mouton de Gruyter. pp. 17-32.
Fries, C.C. (1952). *The Structure of English: an introduction to the construction of English sentences*. New York: Harcourt, Brace & World.
Garsid, R.; G. Leech & T. McEnery (1997) *Corpus Annotation. Linguistic Information from Computer Text Corpora*. London/New York: Longman.
Gaussier, É.; D. Hull & S. Aït-Mokhtar (2000). “Term alignment in use”. em J. Véronis (ed.). *Parallel Text Processing*. Dordrecht/Boston/London: Kluver Academic Publishers. cap 13.
Gómez Guinovart, X. & A.M. Lorenzo Suárez (eds.) (1996). *Lingüística e informática*. Santiago de Compostela: Tórculo Edicións.
Gómez Guinovart, X. & E. Sacau Fontenla (sem publicar). “Técnicas de procesamento linguístico-computacional de corpus paralelos no CLUVI (Corpus Linguístico da Universidade de Vigo) apresentado no *VI Congreso de Linguística General, Santiago de Compostela, 3-7 maio 2004*. (http://webs.uvigo.es/sli/).
Gómez, J. (29-03-2005). “Una Guía al TMX (Translation Memory Exchange)” em http://sirio.deusto.es/abaitua/deli/xtrabi-e341.htm.
Grefenstette, G. & F. Segond (1997). “Multilingual Natural Language Processing”. em *International Journal of Corpus Linguistics*. 2:1. pp. 153-162.

Hallebeek, J. (1999). "El Corpus paralelo". em *Procesamiento del Lenguaje Natural*. 24. pp. 5-16. (http://www.sepln.org/revistaSEPLN/revista/24/24-articulo5.pdf).
Hauenschild, C. & S. Heizmann (eds.) (1997). *Machine translation and translation theory*. Berlin: Mouton: University of Texas Press.
Hill, A.A. (ed.) (1962). *The Third Texas Conference on Problems of Linguistic Analysis in English*. Austin: University Texas Press
Hjelmslev, L. (1969 [1943]). Prolegomena to a Theory of Language. Madison: University of Wisconsin Press.
Hockey, S. (1998). "Textual databases". em J.M. Lawler & H.A. Dry (eds.). *Using Computers in Linguistics. A Practical Guide*. London/New York: Routledge. pp. 101-115.
Hoard, J. E. (1998). "Language understanding and the emerging alignment of linguistics and natural language processing". em J.M. Lawler & H.A. Dry (eds.). *Using Computers in Linguistics. A Practical Guide*. London/New York: Routledge. pp. 197-201.
Hunter, D. (2001). *Iniciación a XML*. Barcelona: IFORBOOK'S. cap 1.
Hutchins, W. J. (1992). "Why Computers do not Translate Better" em VV.AA. *Translating and the computer. A marriage of convenience?*. London: Aslib. pp. 3-17
Hutchins, W.J. (ed.) (1997). *Early years in Machine Translation. Memoirs and biographies of Pioneers*. Amsterdam/Philadelphia: John Benjamins.
Hutchings, W.J. & H.L. Somers (1995). *Introducción a la Traducción Automática*. Madrid: Visor.
Iglesias Santos, M. (org.) (1999). *Teoría de los Polisistemas,* Madrid: Arco.
Instituto António Houaiss (2001). *Dicionário Houaiss da Lingua portuguesa*. Versión (electrónica) 1.0. Editora objectiva: Rio de Janeiro.
Johanson, S. (ed.) (1982). *Computer corpora in English language research*. Bergen: Norwegian Computing Centre for the Humanities.
Jones, D. (1996). *Analogical Natural Language Processing*. London: University College London Press/Centre for Computational Linguistics.
Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. New York: Longman.
Koller, W. (1995). "The Concept of Equivalence and the Object of Translation Studies". em *Target*. 7:2. Amsterdam: John Benjamins. pp. 191-222.
Kristeva, J. (1988). *História da Linguagem*. Lisboa: Edições 70.
Language Technology (1987). "Rosetta. Rock-Aided-Translation" em Meer, J. van der (ed.). *Language Technology. The Magazine of the Language Industries*. 1. p. 11.
Language Technology (1988). "Jean Gachot Resurrecting Systran". em *Language Technology. The Magazine of the Language Industries*. 6. pp. 26-29.
Labov, W. (1982 [1966]). *Social Stratification of English in New York City*. Washington D.C.: Center for Applied Linguistics.
Lawler, J. M. & H.A. Dry (eds.) (1998). *Using Computers in Linguistics. A Practical Guide*. London/New York: Routledge. pp. 1-3.
Leech, G. (1991). "The State of the Art in Corpus Linguistics." em K. Aijmer & B. Altemberg (eds.). *English Corpus Linguistics*, London/New York: Longman. pp. 8-29.
Leech, G. (1992). "Corpora and theories of linguistic performance". em J. Svartvik. *Directions in Corpus Linguistics*. *Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991*. Berlin/New York: Mouton de Gruyter. pp. 105-122.
Leech, G. (1993). "Corpus annotation schemes". em *Literary and Linguistic Computing*. 8:4. pp. 257-281.
Leech, G. (1997). "Introducing Corpus Annotation". em R. Garsid; G. Leech & T. McEnery. *Corpus Annotation. Linguistic Information from Computer Text Corpora*. London/New York: Longman. cap 1.

LISA http://www.lisa.org.
Mahoney, A. (2003). "Talking about Meter in SGML" em *Computers and the Humanities*. 37:4. pp. 469-473.
Malinovski, B. (1935). *Coral Gardens and their Magic*. 2. London: Allen & Unwin.
Marrafa, P. & M.A. Mota (org.) (1999). *Linguística Computacional. Investigação fundamental e aplicações. I Workshop sobre Linguística Computacional da APL, FLUL, Maio de 1998*. Lisboa: Edições Colibri/Associação Portuguesa de Linguística.
Martí Antonín, Mª.A. (cord.). (2003). *Las tecnologías del lenguaje*. Barcelona: Editorial UOC.
Martí Antonín, M.ªA. & J. Llisterri (eds.) (2002). *Tratamiento del lenguaje natural*. Barcelona: Edicions Universitat de Barcelona.
Martín, R. (1999). *Alineación automática de corpus paralelos: una propuesta metodológica y su aplicación a un dominio de especialidad*. Tesis doutoral. Universidad de Deusto.
Mata, I. (1995). "A periferia da periferia. O estatuto periférico das literaturas Africanas de Língua Portguesa e a adupla perifericidade das literaturas são-tomense e guineense" em *Discursos*, 9. pp. 27-36.
McEnery, T. (2003). "Corpus Linguistics" em R. Mitkov (ed.). *The Oxford Handbook of Computational Linguistics*. New York: Oxford University Press. cap. 24.
McEnery, T. & A. Wilson (2001). *Corpus Linguistics. An Introduction (Second Edition)*. Edinburgh: Edinburgh University Press.
Melby, A.K. (2000). "XML and the Translator" em *Translation Journal*. 4:2. (http://accurapid.com/journal/12xml.htm).
Mitkov, R. (ed.) (2003). *The Oxford Handbook of Computational Linguistics*. New York: Oxford University Press.
Moreno Fernández, F. (1990). "Lingüística Informática e Informática Linguística". em *Lingüística Española Actual*. 12:1. pp. 5-16.
Moreno Sandoval, A. (1998). *Lingüística Computacional. Introducción a los modelos simbólicos, estadísticos y biológicos*. Madrid: Síntesis.
Moure, T. & J. Llisterri (1996). "Lenguaje y nuevas tecnologías: El Campo de la Lingüística Computacional". em M. Fernández Pérez (coord.). *Avances en Lingüística Aplicada*. Santiago de Compostela: Universidade de Santiago de Compostela. pp. 147-227.
Nagao, M. (1984). "A framework of a mechanical translation between Japanese and English by analogy principle". em A. Elithorn & R. Banerji (eds.). *Artificial and Human Intelligence*. Brussels: NATO Publications.
Nagao, M. (1989). *Machine Translation. How far can it go?*. New York: Oxford University Press.
Nagashima, Y. (ed.) (1968). *Selected writings of Edward Sapir*. Tokio: Nan'un-do.
Och, F.J. & H. Ney (2004). "The Alignment Template Approach to Statistical Machine Translation". em *Computational Linguistics*. 30:4. pp. 417-449.
Olohan, M. (2004). *Introducing Corpora in Translation Studies*. Oxfordshire/New York: Routledge.
Olsen, M. & A. Music McLean. (1993) "Optical Character Scanning: A Discusion of Effiency and Politics" em *Computers and the Humanities*, 27: 2. pp. 121-127.
Palmer, F.R (1968). *Selected Papers of J.R. Firth 1952-59*. London/Harlow: Longman.
Pogson, G. (1988). "Maghi King on Machine Translation". em *Language Technology. The Magazine of the Language Industries*. 5. pp. 11-12.
P. Rodrigues, J.H. (2000). *Introduçom à Lingüística com Corpora*. Ferrol/Santiago de Compostela: Artábria Fundaçom/Laiovento.

Rabadán, R. & F.J. Fernández Polo (1996). "Linguística Aplicada a la Traducción". em M. Fernández Pérez (coord.). *Avances en Lingüística Aplicada*. Santiago de Compostela: Universidade de Santiago de Compostela. pp. 105-145.

Rafel i Fontanals, J. & J. Soler i Bou (2003). "El procesamiento de corpus. La lingüística empírica". em Mª.A. Martí Antonín (cord.). *Las tecnologías del lenguaje*. Barcelona: Editorial UOC. cap. 2.

Riera, M. (ed.) (1992). *Quimera*. 112-114. p. 3.

Rodríguez Prado, M. F. (2001). "Literaturas Africanas de Língua Portuguesa no Estado Espanhol: Uns poucos livros." em *Cadernos Vianenses*. 30. Viana do Castelo: Câmara Municipal de Viana do Castelo. pp. 81-92.

Rossini Favretti, R. (2000) *Linguistica e Informatica. Corpora, Multimedialità e Percorsi di Apprendimento*. Roma: Bulzoni Editore.

Santos, D. (1999). "Disponibilização de corpora de texto através da WWW". em P. Marrafa & M.A. Mota (org.). *Linguística Computacional. Investigação fundamental e aplicações. I Workshop sobre Linguística Computacional da APL, FLUL, Maio de 1998*. Lisboa: Edições Colibri/Associação Portuguesa de Linguística. pp. 323-335.

Santos, D. (2000). "The translation network". em J. Véronis (ed.). *Parallel Text Processing*. Dordrecht/Boston/London: Kluver Academic Publishers. cap 8.

Sapir, E. (1929). "Central and North American languages" em W. Bright (ed.) (1990). *The Collected Works of Edward Sapir, V: American Indian Languages (1)*. Berlin: Mouton de Gruyter. pp. 95-104

Sapir, E. (1933). "Language" em Nagashima, Y. (ed.) (1968). *Selected writings of Edward Sapir*. Tokio: Nan'un-do. pp. 42-93

Saussure, F. de (1998). *Curso de Lingüística General*. Madrid: Alianza Editorial.

SDL Internacional e-nabling global business (15-03-2005). "Knowledge-based Translation". Em http://www.sdl.com/localization-information/white-papers-articles/premium-white-papers/premium-white-papers-knowledge-based-translation.htm.

SDL Internacional e-nabling global business (15-03-2005). "The Importance of TMX". em http://www.sdl.com/localization-information/white-papers-articles/white-papers-list/white-papers-importance-of-tmx.htm.

Schuetz, J. (1987). "A computer Translation System for Authors". em *Procesamiento del Lenguaje Natural*. 5. Monográfico sobre las III Jornadas de la SEPLN. pp. 93-97. (http://www.sepln.org/revistaSEPLN/revista/5/5-Pag93.pdf).

Simões, A.; X. Gómez Guinovart & J.J. Almeida (2004). "Distributed Translation Memories implementation using WebServices". em *Procesamiento del Lenguaje Natural*. 33. pp. 89-94. (http://193.144.40.227/search*gag/jprocesamiento+del+lenguaje+natural/jprocesamiento+del+lenguaje+natural/1%2C1%2C1%2CB/l856&FF=jprocesamiento+del+lenguaje+natural&1%2C0%2C%2C1%2C0).

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

Sinclair, J. (2000). "Current Issues in Corpus Linguistics." em R. Rossini Favretti. *Linguistica e Informatica. Corpora, Multimedialità e Percorsi di Apprendimento*. Roma: Bulzoni Editore. pp. 29-38.

Slocum, J. (ed.) (1988). *Machine translation systems*. Cambridge: Cambridge University Press.

Snell-Hornby, M. (1986). Übersetzungwissenschaft – eine Neuorientierung: Zur Integrierung von Theorie und Praxis. Tübingen: Narr. [Uni-Taschenbücher, 1415].

Soto Vázquez, A. L. (ed.); B. Crespo & P. Cancelo López. (eds. associados) (2001). *Insights into Translation*, Vol. III. A Coruña: Tórculo.

Stoddard, J. (1987). "My Computer Speaks to Me in Pidgin. Ruminations on Natural Language Interfaces" em *Language Technology. The Magazine of the Language Industries*. 4. p. 7.
Streiter, O. (2000). "Learning Lessons from Bilingual Corpora: Benefits for Machine Translation". em *International Journal of Corpus Linguistics*. 5:2. pp. 199-230.
Svartvik, J. (1992). *Directions in Corpus Linguistics*. *Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991*. Berlin/New York: Mouton de Gruyter.
Svartvik, J. (1996). "Corpora are becoming mainstream" em J. Thomas & M. Short. *Using Corpora for Language Research*. *Studies in the Honour of Geoffrey Leech*. London/New York: Longman. pp. 3-13.
TEI Guidelines em http://www.tei-c.org.
Thomas, J. & M. Short (1996). *Using Corpora for Language Research*. *Studies in the Honour of Geoffrey Leech*. London/New York: Longman.
Tognini-Boneli, E. (2001). *Corpus Linguistics at work*. Amsterdam/Philadelphia: John Benjamins.
Valero Garcés, C.; D. Sale; B. Soto & M. El-Madkouri (2004). "Panorama de la Traducción de Literatura de Minorías en la España de Comienzos de Siglo: Literatura de la India, Literatura Árabe, Literatura Magrebí y Literatura de Países Africanos" em *Revista Electrónica de Estudios Filológicos* (http://www.um.es/tonosdigital/znum8/estudios/18-tradumin.htm).
Vanhoutte, E. (2004) "An Introduction to the TEI and the Tei Consortium" em *Literary and Linguistic Computing*. *Journal of the Association for Literary and Linguistic Computing*. 19:1. pp. 9-16.
Váradi, T. & G. Kiss (2001). "Equivalence and Non-equivalence in Parallel Corpora". em *International Journal of Corpus Linguistics*. 6 (Special Issue). pp. 167-177.
Véronis, J. (2000). "From the Rosetta stone to the information society". em J. Véronis (ed.). *Parallel Text Processing*. Dordrecht/Boston/London: Kluver Academic Publishers. cap 1.
Véronis, J. (ed.) (2000). *Parallel Text Processing*. Dordrecht/Boston/London: Kluver Academic Publishers.
Verschueren, J. (1987). "The pragmatic perspective". em J. Verschueren & M. Bertucceli-Papi (eds.). *The pragmatic perspective: selected papers from the 1985 International Pragmatics Conference*. Amsterdam/Philadelphia: John Benjamins. pp. 3-8.
Verschueren, J. & M. Bertucceli-Papi (eds.) (1987). *The pragmatic perspective: selected papers from the 1985 International Pragmatics Conference*. Amsterdam/Philadelphia: John Benjamins.
Villayandre Llamazares, M. (1997). "Lingüística Computacional: Heterogeneidad e Interdisciplinaridad" em *Interlingüística*. 7. pp. 259-265.
Vuyst, J. de (1990). "Knowledge Representation for Text Interpretation". em *Journal of the Association for Literary and Lingistic Computing*. 5:4. pp. 296-302.
VV.AA. (1992). *Translating and the computer. A marriage of convenience?*. London: Aslib.
Wils, W. (1994). "Translation as a Knowledge-Based Activity. Context, Culture and Cognition" em R. de Beaugrande; A. Shunbaq & M.H. Heliel. *Language, Discourse and Translation in the West and Middle East*. Amsterdam/Philadelphia: John Benjamins. pp. 35-43.
World Wide Web Consortium http://www.w3.org/.
Zarechnak, M. (1987). "Space Age and Machine Translation" em *Language Technology. The Magazine of the Language Industries*. 4. p. 6.