CORPORART - um *corpus* de arte pública para a extração de léxico: representatividade e comparabilidade em *corpora* de especialidade

Chiara Barbero[1]
Universidade NOVA de Lisboa, CLUNL

Abstract:

This paper introduces the CORPORART, a bilingual *corpus* of Public Art. CORPORART intends to gather, in a single collection of bilingual data, representative samples of specialized language in European Portuguese and Italian. The compilation of this *corpus* is part of an ongoing doctoral project, which aims to integrate specialized lexical units into a pre-existing common language resource, WordNet.PT (Marrafa *et al.*, 2005), in the perspective of contributing to streamline communication between heterogeneous interlocutors (Amaro & Mendes, 2012). Assuming that the structure of the *corpus* heavily depends on the goals of the investigation, this paper presents the linguistic and extralinguistic parameters adopted for the construction and organization of the *corpus*, as well as the criteria for text processing. In particular, we will deepen the notion of representativity and comparability considering the specificity of this case study, outlining a work practice proposal oriented to guarantee these two flexible dimensions within the specialized languages context.

***Keywords*:** specialized *corpora*, specialized lexicon, *corpus* organization, representativity, Public Art

Palavras-chave: *corpora* de especialidade, léxico de especialidade, organização do *corpus*, representatividade, arte pública

1. Introdução

A necessidade de construir um *corpus* de arte pública surgiu no âmbito de um projeto de investigação em curso, que visa reunir um conjunto organizado de textos de especialidade num recurso linguístico funcional para a análise da relação entre léxico comum e léxico de especialidade[2] no quadro de um modelo relacional de organização do léxico, o da WordNet.

Para a constituição do *corpus* serão utilizadas, como ponto de partida, as diretrizes de Sinclair (Sinclair, 1996) e de Tognini-Bonelli (Tognini-Bonelli, 2001) quanto à noção de *corpus* e, em particular, de *corpus* de especialidade. Entender-se-á por *corpus* uma amostra representativa de dados linguísticos que refletem o uso real da língua ou, no caso dos *corpora* de especialidade, o uso da língua em contextos específicos, i.e. o discurso especializado.

No que diz respeito ao CORPORART, serão descritos a estrutura e os critérios de compilação do *corpus*, de acordo com os objetivos gerais da investigação. Em particular, serão analisadas as especificidades e os critérios de tratamento dos textos que visam garantir a representatividade da amostra selecionada e a qualidade dos dados linguísticos que serão extraídos, mas também as dificuldades e os desafios que derivam da complexidade da composição do domínio de especialidade selecionado.

[1] Este trabalho foi financiado pela FCT - Fundação para a Ciência e a Tecnologia - no âmbito da Bolsa de Doutoramento KRUse (PD/BD/128131/2016).

[2] A noção de léxico aqui considerada segue a definida por Salem (1987) (Salem, 1987), referindo-se ao conjunto virtual das palavras de uma língua, podendo ser de uso corrente – léxico corrente, ou de uso especializado, tipicamente de um domínio e usado por uma comunidade de especialistas – léxico de especialidade. Apesar de baseado em *corpora*, o trabalho apresentado não visa a descrição da lista exaustiva das palavras do *corpus* constituído - vocabulário (Guilbert, 1971), mas sim a extração de potenciais candidatos para integrar uma *wordnet*.

O CORPORART constituirá um recurso disponível tanto para a comunidade científica como para a sociedade em geral, revelando-se, à luz da análise levada a cabo neste artigo, um recurso, para além de original, também vantajoso por reunir informação dispersa e difícil de consultar no seu conjunto.

1.1. Arte pública: quem disse que a arte é só para especialistas?

Do que é que estamos a falar quando falamos de arte pública?

A resposta a esta pergunta não é tão trivial como seria esperado. De forma geral, podemos dizer que, embora não exista uma definição clara e unívoca, por arte pública entendemos uma modalidade de apresentação e fruição da arte que entra no tecido social e na estrutura urbana da cidade e que se manifesta no território público. («Fondazione Contrada Torino Onlus», sem data)

Dito de outra forma, é aquela dimensão artística que encontra a sua razão de ser fora das paredes elitistas dos museus e das galerias, é a arte que vem para a rua e "se funde progressivamente com a vida quotidiana, no próprio desenho integrado do espaço público". (Remesar & Brandão, 2010)

E é precisamente esta sua natureza democrática, mas não forçosamente massificada, e acessível a um público amplo e diversificado que orientou em grande medida a seleção deste domínio na ótica do trabalho de investigação.

Neste sentido o elemento "público", quer na aceção de *res publica* devido à localização das peças, quer enquanto conjunto de destinatários e beneficiários da comunicação artística (ou seja, os cidadãos em geral), adquire grande relevância na ponte que se estabelece entre universo especializado e não especializado. E, mais precisamente, entre os discursos que estes dois universos produzem sobre o mesmo objeto.

As manifestações de arte pública demonstram uma sensibilidade particular à participação coletiva, prática que permite exatamente que haja uma interação profunda entre atores profissionais e não profissionais, envolvendo assim o público na reflexão acerca do significado e do papel do espaço público, enquanto lugar com que a comunidade se identifica, e da comunidade, enquanto protagonista do processo criativo e de resignificação do espaço público (Bargna, 2012; Elias & Valente, 2017) .

> Thus, public art is art which has as its goal a desire to engage with its audiences and to create spaces—whether material, virtual or imagined—within which people can identify themselves, perhaps by creating a renewed reflection on community, on the uses of public spaces or on our behaviour within them. (Sharp *et al*., 2005, pp. 1003–1004)

No próprio Manifesto da Escultura Pública de Siah Armajani encontramos expressa de forma clara e explícita a importância da relação entre o mundo artístico e profissional com o contexto social não especializado:

> 14. The ethical dimensions of the arts are mostly gone and only in a newly formed relationship with a non-art audience may the ethical dimensions come back to the arts. (Armajani, 1995, p. 36)

Assim sendo, a escolha deste domínio garante a existência de pontos de contacto entre os discursos produzidos pelos dois grupos de falantes e, portanto, espera-se que seja um terreno propício para analisar as diferenças e as semelhanças entre as conceptualizações e lexicalizações que estes propõem acerca das mesmas entidades.

Ainda, tendo em conta o que foi dito até agora, chegamos à conclusão de que a arte pública é um domínio profundamente interdisciplinar, pois convoca necessariamente pelo menos três dimensões em simultâneo: a dimensão artística, a dimensão espacial e a dimensão social. Basta pensar, portanto, na variedade de disciplinas que estas três dimensões envolvem (desde as artes plásticas, à arquitetura, ao

urbanismo, ao *design*, passando também pela geografia, pela sociologia e pela educação, entre outros) para tomarmos consciência da variedade de discursos possíveis que, apesar de abordarem o mesmo tópico, refletem pontos de vista muito diferentes.

(…) esta acepção de arte pública não fecha o campo de intervenção no plano artístico, arquitetónico ou de design urbano, mas abre um território de conhecimento interdisciplinar, no qual se integram todos os elementos que compõem o espaço urbano. (Ochoa, 2011, pp. 75)

No contexto da demonstração da abrangência deste domínio, vale a pena citar, também, o caso da "Rede de Informação, Investigação e Intervenção em arte pública", projeto que surgiu em 2018 entre algumas unidades de investigação e universidades portuguesas, proposto e coordenando pelo CITAR (Universidade Católica do Porto), que visa juntar instituições ligadas a diferentes áreas de especialidade, estabelecendo objetivos e diretrizes de trabalho comuns (Abreu & Castro, 2018).

É, portanto, esta dupla vertente da arte pública, enquanto interface entre universo especializado e não especializado por um lado, e enquanto interface entre diferentes universos especializados, por outro, que constitui um campo de investigação particularmente interessante na perspetiva da investigação, tendo em conta que o projeto de investigação mais abrangente que envolve a constituição deste *corpus* procura analisar e codificar de forma sistemática a relação entre léxico de especialidade e léxico comum, a fim de integrar recursos lexicais de natureza diferente.

1.2. Prós e contras: riqueza ou excesso?

A vantagem mais evidente que deriva de trabalhar um domínio tão vasto e abrangente é, naturalmente, a riqueza lexical que este proporciona, requisito fundamental para a análise lexical que se pretende realizar.

Mas a outra face da moeda é, inevitavelmente, a complexidade, que tem fortes implicações na organização do *corpus* e, consequentemente, também no tratamento das unidades lexicais.

O facto de ser extremamente complexo organizar este *corpus* de forma orgânica e sistemática deve-se em grande medida à impossibilidade de traçar fronteiras claras entre as múltiplas subáreas que compõem o domínio que se pretende representar, subáreas que, pelo contrário, se sobrepõem e se complementam. Neste sentido, é muito frequente que num mesmo texto ocorram elementos de áreas diversas. É, portanto, expectável que esta fluidez se reflita em vários momentos e aspetos da análise que será realizada a partir deste *corpus* no futuro, como por exemplo no tratamento das unidades lexicais.

Na sequência destas considerações, ainda que não seja o foco deste trabalho, será pertinente apresentar uma reflexão acerca da construção das definições das unidades lexicais de especialidade a partir dos próprios textos, na medida em que este processo torna muito evidente a noção de variabilidade descrita acima, que implica a necessidade de tomar decisões e fazer escolhas dentro da multiplicidade de pontos de vista possíveis. Como podemos verificar em (1), dependendo da subárea de especialidade em que o texto se insere, teremos variações consideráveis no que diz respeito à definição dos conceitos que conseguimos inferir a partir dos textos.

Veja-se, no exemplo a seguir, o processo de extração da definição de *graffiti* a partir de três excertos de teses de mestrado realizadas no quadro de áreas científicas diferentes, nomeadamente: arte, património e teoria do restauro, o primeiro, cultura e comunicação, o segundo, e arquitetura, o terceiro.

(1)

a) "Graffiti – (it. *graffiare*) meio de comunicação não institucional, realizado manualmente com o uso de sprays ou outros materiais em suporte móvel ou fixo." (Simões, 2013, pp. 97)

b) "uma prática que encontra nos muros, nos transportes públicos ou no mobiliário urbano, suportes para a afirmação de identidades, para a marcação territorial ou simplesmente, para uma proclamação de existência." (Campos, 2010 *apud* Valente, 2016, pp. 49)

c) "discurso visual urbano, que consideramos activo na qualificação ou requalificação do espaço urbano" (Santos, 2016, pp. 19)

A partir desta pequena amostra[3] nota-se claramente como os três textos nos oferecem pistas muito diferentes para representar e definir *graffiti*, de acordo com as perspetivas que cada área privilegia.

Desta forma, no caso de elementos que não podem coexistir dentro de uma mesma definição/glosa, i.e hiperónimos diferentes que não correspondem a casos de polissemia regular compatível como "meio de comunicação" (1.a.) e "prática" (1.b.), será necessário estabelecer critérios sólidos para determinar a escolha da informação semântica que se quer privilegiar no quadro deste trabalho.

Uma vez que se escolheu colaborar com uma especialista da área[4], para validar tanto o *corpus* em si, como o mapeamento das redes lexicais, privilegiar-se-á a orientação indicada pela especialista.

2. CORPORART: compilação de um *corpus* de especialidade

No quadro deste trabalho, será tida em conta a definição de *corpus* estabelecida pelos autores pioneiros da linguística de *corpus*, Sinclair (Sinclair, 1991, 1996) e Tognini-Bonelli (Tognini-Bonelli, 2001), entre outros, que continua a demonstrar-se válida e atual nos dias de hoje.

Neste sentido, podemos dizer que por *corpus* entendemos uma amostra de dados linguísticos, selecionados e organizados de acordo com critérios pré-estabelecidos em conformidade com os objetivos da investigação linguística em causa, o mais representativa possível de uma determinada população[5].

Afunilando o foco, desde a noção geral de *corpus* para o CORPORART, podemos identificar alguns dos critérios que distinguem a compilação deste *corpus*, nomeadamente o facto de ser:

- bilingue – inclui dados de duas línguas naturais, nomeadamente o português europeu e o italiano;
- comparável – são aplicados os mesmos critérios de compilação e de seleção dos textos para ambas as línguas representadas;
- especializado – os textos recolhidos são produções de especialistas, ou especialistas em formação (e.g. estudantes de mestrado), destinadas à comunidade científica, e não à divulgação para o público geral, no âmbito de um domínio de especialidade específico, o da arte pública;
- fechado – os textos recolhidos datam de 2000 a 2018. Apesar da rápida evolução do domínio em questão, em particular, no que diz respeito ao surgimento de novas modalidades artísticas ligadas à aplicação das novas tecnologias (Regatão, 2016), a opção de deixar em aberto a possibilidade de ir

[3] Escolhemos limitar a amostra apresentada no corpo deste artigo ao exemplo de "*graffiti"*; no entanto é importante referir que encontramos inúmeros casos semelhantes no *corpus* ("espaço público", "público", "requalificação urbana" entre outros).

[4] Rita Ochoa, Professora Auxiliar no Departamento de Engenharia Civil e Arquitetura da Universidade da Beira Interior, onde leciona desde 2006. Arquiteta, co-fundadora da associação "Mulheres na Arquitetura". Licenciada em Arquitetura em pela Faculdade de Arquitetura da Universidade Técnica de Lisboa (1997), Pós-graduada em Qualificação da Cidade pela Universidade Católica Portuguesa (2002), Master em Desenho Urbano pela Universidade de Barcelona (2008) e Doutorada em Espaço Público e Regeneração Urbana, Arte e Sociedade pela Universidade de Barcelona (2011). Membro da comissão editorial e co-fundadora da Revista Branca, Revista de Arquitetura da Universidade da Beira Interior.

[5] "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" (Sinclair, 1996, pp. 4)

acrescentando textos a medida que forem produzidos e disponibilizados à comunidade científica não é viável, devido ao calendário da atividade de investigação que o *corpus* visa sustentar;

- anotado – os dados incluídos no *corpus* são lematizados e anotados de forma automática – com uma anotação larga que atribui uma categoria morfossintática a cada palavra do *corpus* – e de forma manual no que respeita a associação de um cabeçalho com metadados a cada texto do *corpus*, que permite agilizar a comparação de dados registados no *corpus* com base em determinados critérios (e.g. tipo textual; área científica, etc.).

Para a codificação deste cabeçalho foi usada uma linguagem de marcação standard, o XML, para associar a cada texto informação como o nome do ficheiro (ex. CART-PT-PhD01), o título, a data de publicação, a língua, o endereço web (URL), o tipo de texto, a área ou as áreas científicas em que o texto se inscreve.

- *research driven* - tipicamente, os *corpora* não são elaborados como um fim em si mesmos, mas sim em função de objetivos de investigação específicos. Utilizando as palavras de Tognini- Bonelli: "The decisions made in assembling a corpus, or the choice of what type of corpus to access if one is going to be just a user, depend on what is the use to which the corpus is going to be put" (Tognini-Bonelli, 2001, pp. 6).

Na sequência deste último ponto, importa, portanto, delinear brevemente os objetivos gerais da investigação em que a constituição do *corpus* aqui apresentado se enquadra e que são essencialmente três: em primeiro lugar, extrair e analisar o léxico especializado da área da arte pública para a construção de redes léxico-conceptuais de especialidade semi-independentes; estabelecer critérios estáveis para ligar redes de especialidade a uma rede de léxico comum pré-existente; e finalmente, criar uma metodologia de trabalho potencialmente aplicável a outras áreas e outras línguas.

2.1. Representatividade, equilíbrio e qualidade

Tendo em conta a definição de *corpus* aqui considerada, a representatividade revela-se um conceito central para o sucesso e para a qualidade dos resultados da investigação. Sendo impossível ter acesso, recolher e catalogar a totalidade de textos produzidos numa determinada língua ou variedade, salvo casos muito específicos, a amostragem é uma estratégia indispensável. Portanto, é fundamental garantir que a amostra selecionada seja o mais representativa possível da língua ou da variedade considerada, uma vez que a qualidade das generalizações inferidas a partir da análise dos fenómenos linguísticos observados é proporcional ao grau de representatividade da amostra selecionada. (McEnery *et al*., 2006, pp. 13)

Seria ingénuo considerar a representatividade apenas em termos de quantidade, ou seja, dando relevância apenas ao número de palavras que compõem o *corpus*. "Representativeness refers to the extent to which a sample includes the full range of variability in a population." (Biber, 1993, pp. 243)

Temos, pelo contrário, de considerar que a representatividade não é uma dimensão quantificável em termos absolutos e com parâmetros universalmente válidos, mas depende de múltiplas variáveis de caráter mais ou menos linguístico ou, de acordo com a terminologia de Biber (1993), de fatores situacionais (ou externos) e de fatores linguísticos (ou internos).

Entre estes fatores destacamos aqui alguns, a título de exemplo:

a) a população que se pretende representar, o que implica delimitar claramente o universo da análise;
b) a variedade textual que é possível encontrar dentro do universo estabelecido, ou seja, os diferentes tipos de texto produzidos num determinado âmbito;
c) os recursos e os meios à disposição, mas também a facilidade de acesso aos textos;
d) o tipo de *corpus* (de referência, de especialidade, etc.) e, portanto, indiretamente, os objetivos que estão na base da constituição do mesmo, uma vez que, se se tratar de um *corpus* de referência, a

representatividade será determinada pela maior cobertura possível de géneros textuais, enquanto, no caso de *corpora* de especialidade, a representatividade será determinada pelo grau de saturação (pelo menos no que diz respeito ao nível lexical) (McEnery *et al*., 2004, pp. 15);

e) o fator diacrónico que se aplica maioritariamente aos casos de *corpora* contemporâneos que precisam de atualizações constantes sob pena de, caso contrário, se tornarem rapidamente pouco representativos do uso corrente da língua (Hunston, 2002).

No caso específico deste projeto, devido a alguns constrangimentos situacionais, como será argumentado de forma mais aprofundada na secção 5, decidiu-se privilegiar a representatividade em sentido ontológico, mais do que estritamente formal. Isto significa dar prioridade à cobertura do máximo de subáreas de especialidade possível dentro do domínio considerado, não obstante a dificuldade em encaixar os textos em rótulos estáveis devido à fluidez dos tópicos abrangidos, em parcial detrimento do equilíbrio formal de acordo com a distribuição dentro das tipologias textuais consideradas.

Ou seja, se por um lado é inegável o facto de não podermos apontar para o equilíbrio formal como ponto de força deste recurso, por outro este fornece os instrumentos para tirar proveito da riqueza lexical procedente das diferentes áreas e subáreas do domínio, o que terá implicações muito positivas na cobertura das redes lexicais que constituem o objetivo central do projeto de investigação mais abrangente que determina a necessidade de constituição do *corpus* que aqui se apresenta.

Apesar dos limites descritos até aqui, garantir-se-á a qualidade do CORPORART através do respeito rigoroso dos critérios de compilação e seleção dos textos e da validação do trabalho desenvolvido por especialistas. Com isto pressupõe-se uma colaboração próxima com especialistas do domínio de investigação em questão para (i) avaliarem a pertinência e a relevância dos textos selecionados dentro do panorama dos textos existentes e (ii) detetarem tipologias textuais, bem como autores e/ou obras importantes que possam ter sido omitidas na fase de recolha, por falta de conhecimento e sensibilidade específica na área.

2.2. Organização do *corpus*: critérios de seleção dos textos e balanço das imperfeições

Uma vez estabelecido o domínio de especialidade de interesse e definida a população a ser representada, realizou-se uma seleção dentro do universo textual existente.

De acordo com os objetivos da investigação que motivaram a constituição do CORPORART foram selecionados textos que respondessem aos seguintes critérios:

- Recorte temporal: pretendendo analisar a língua contemporânea, foram utilizados apenas textos produzidos a partir do ano 2000. No entanto, vale a pena mencionar o facto de a maioria dos textos recolhidos terem sido produzidos do ano 2010 para a frente.

- Originalidade: para evitar ao máximo qualquer tipo de artificialidade e de interferência nos dados decorrente do processo de tradução, foram considerados apenas textos escritos originalmente em Português e Italiano. Desta forma pretende-se salvaguardar a naturalidade do uso das unidades lexicais que irão popular as *wordnets* para cuja constituição serão utilizados os dados do CORPORART.

- Especialização: querendo garantir um alto nível de especialidade das redes, escolheu-se utilizar só textos produzidos por especialistas (incluindo dissertações de mestrado) e direcionados a um público-alvo especializado.

- Formato: tanto para uniformizar o trabalho, como para economizar tempo e recursos, escolheu-se utilizar apenas textos já existentes em formato digital e acessíveis de forma gratuita. Isto implica, portanto, que não foi realizada qualquer tarefa de digitalização de textos em papel. Os capítulos de livros editados em formato papel que constam do *corpus* foram retirados, já em formato digital, de

repositórios académicos, de revistas do setor ou de plataformas em linha. No entanto, por questões de compatibilidade de formato com as ferramentas de tratamento semiautomático de *corpora*, todos os textos passaram por uma ferramenta de reconhecimento ótico de carateres (o ABBYY Fine Reader) de forma a poderem ser transformados de formato de imagem (jpg; png) ou de texto não editável (pdf) para o formato de texto simples (txt). A escolha da ferramenta foi determinada pelo desempenho da mesma com as línguas de trabalho.

Portanto, na sequência da seleção dos textos orientada pelos critérios acima descritos, foram deixados de fora: (i) as publicações em papel (livros, manuais, catálogos de exposições, revistas etc.), (ii) qualquer produção destinada à vulgarização e divulgação para o público geral (artigos de jornal), (iii) trabalhos académicos produzidos no contexto de ciclos de estudos graduados (e.g. teses de licenciatura), (iv) textos de acesso restrito (artigos de revistas de acesso pago), (v) textos noutras línguas e (vi) textos traduzidos a partir de outras línguas.

Depois de ter estabelecido os parâmetros formais e teóricos a seguir na compilação do *corpus* e na seleção dos textos, é também fundamental fazer uma avaliação dos problemas e das imperfeiçoes que o *corpus* apresenta, uma vez que "a natureza intrinsecamente desequilibrada dos *corpora* não lhes tira o mérito de serem uma fonte rica de evidências linguísticas, desde que tal fonte seja utilizada de forma consciente e com a devida cautela" (Lenci *et al*., 2016, pp. 42) [tradução minha].

Portanto, como já mencionado anteriormente, trabalhar neste domínio de especialidade significa ter em consideração as dificuldades inevitáveis que derivam da abrangência de uma realidade tão multifacetada e interdisciplinar quanto a da arte pública, no que diz respeito à procura de um equilibro tanto formal como ontológico entre os diferentes tipos de textos e as subáreas tocadas. Contudo, houve outro fator determinante que tem adquirido cada vez mais relevância durante a fase de recolha dos textos, e que teve grande impacto em termos de comparabilidade entre os dois *corpora*: trata-se da acessibilidade dos textos e das fontes. Como sabemos as políticas de acesso, mais ou menos aberto, e de partilha de conhecimento científico variam muito consoante a zona geográfica e a área científica. Neste caso específico, o contexto académico italiano foi o que se demonstrou mais problemático e que influenciou de forma particularmente negativa a recolha dos textos, uma vez que o número de trabalhos académicos disponíveis (tanto teses de mestrado como de doutoramento, mas também de artigos científicos) é muito reduzido em comparação com a realidade portuguesa. É evidente que este fenómeno afetou em grande medida o grau de comparabilidade dos *corpora* em termos quantitativos (ver Tabela 1.). Ainda assim, consideramos que a comparabilidade continua garantida, embora necessariamente de forma parcial, pela aplicação dos mesmos critérios de recolha e seleção para ambas as línguas (cf. secção 2.4).

Finalmente, chegando às questões mais práticas relativas ao próprio estilo e formatação dos documentos, é preciso ter em conta a grande variedade de formatos que os ficheiros apresentam e, em contrapartida, a necessidade de garantir um tratamento uniforme e *standard* para todos os tipos de texto.

Relativamente a este último ponto, foi necessário transformar todos os documentos de partida em ficheiros de texto simples, o que implica reduzir qualquer estrutura textual mais complexa e criativa em texto corrido. Depois procedeu-se a uma "limpeza" manual, através de um editor de texto, para eliminar todas as fontes de ruído e os elementos que pudessem originar incompatibilidade com o formato de texto simples.

Nomeadamente, foram retirados: (i) os elementos pessoais, como por exemplo os agradecimentos ou as referências de bolsas de investigação associadas aos autores, (ii) os elementos técnicos relativos ao texto, como fichas técnicas ou normas de direitos de autor, (iii) as biografias dos autores, (iv) as divisões ou fragmentos escritos em outras línguas, nomeadamente resumos ou citações, (v) as referências bibliográficas, tanto no fim como no corpo do texto e (vi) as imagens.

Adicionalmente foram convertidas (vii) as notas de rodapé em notas de fim, para não interferir na ordem do texto e (vii) as colunas emparelhadas, típicas das revistas ou das atas de conferência, em colunas únicas, mais uma vez para não alterar a ordem do texto na passagem a texto simples.

Segue que, em termos gerais, o que se mantém após a tarefa de limpeza é: (i) o corpo do texto, (ii) os resumos e as palavras-chave, quando nas respetivas línguas de trabalho, típicos dos trabalhos académicos e dos artigos científicos, (iii) os índices analíticos, (iv) as legendas de figuras e tabelas, por serem normalmente ricos em terminologia relevante, (v) as tabelas, (vi) os anexos, quando avaliados como interessantes relativamente ao domínio em questão.

Apesar de termos de considerar uma margem de erro humana, uma vez que esta tarefa foi realizada de forma manual, os resultados (nomeadamente, as listas de palavras) resultam certamente beneficiados desta limpeza, em comparação com os resultados que se conseguiriam sem este tipo de pré-edição humana.

2.3. Integralidade dos textos

Ainda no âmbito dos critérios de utilização e de tratamento dos textos, foi avaliada a questão da integralidade dos documentos, em particular, para definir a utilização de textos extensos com muitas ocorrências, como as teses de doutoramento e os livros.

Diferentes autores apresentam opiniões divergentes acerca desta questão. Sinclair (2004), por exemplo, defende a importância do uso de documentos integrais, ou o mais perto possível de uma versão integral, independentemente de o *corpus* apresentar documentos de tamanho muito diferente: "The integrity and representativeness of complete artefacts is far more important than the difficulty of reconciling texts of different dimensions." (Sinclair, 2004, pp. 11). Se por um lado é preferível não manipular os textos de forma a os resultados refletirem o uso o mais real da língua possível, em muitos casos, esta ideia entra em conflito com as políticas e normas de direitos de autor – caso que sustenta a opção de amostragem em *corpora* de referência, por exemplo.

Por outro lado, Biber (1993) defende a opinião diametralmente oposta, tendo em conta a distribuição, tipicamente estável ao longo do texto, de muitos dos fenómenos linguísticos (entre as quais o léxico): "Common linear linguistic features are distributed in a quite stable fashion within texts and can thus be reliably represented by relatively short text segments" (Biber, 1993, pp. 252). De acordo com esta perspetiva, seria necessário estabelecer, também, normas de seleção e amostragem dentro dos próprios textos.

Contudo ambos os autores reforçam a ideia de que não há regras universais, mas é sempre preciso ter em conta o tipo de *corpus* e a finalidade da investigação: "Claims of corpus representativeness and balance, however, should be interpreted in relative terms and considered as statement of faith rather than as fact, as presently there is no objective way to balance a corpus or to measure its representativeness. " (McEnery *et al*., 2004, pp. 21)

No caso do CORPORART, a questão acerca da utilização dos textos na sua forma integral surgiu quer por razões de equilíbrio e harmonia dentro do *corpus*, quer por razões de pertinência em relação ao fio condutor da amostragem.

As teses de doutoramento, em particular, foram os tipos de texto que levantaram o maior número de questões, devido às estruturas muito heterogéneas que apresentam, e devido à extrema diversidade de focos relacionadas com as várias áreas que integram (e.g. metodologias de investigação, enquadramento histórico/social, etc.).

Nessa perspetiva, apresenta-se aqui uma pequena experiência realizada com o intuito de responder a esta questão e de determinar a escolha de uma ou outra abordagem, de acordo com as especificidades deste *corpus*.

Selecionou-se, portanto, um texto em Português entre os pertencentes à categoria tese de doutoramento - CART-PT-PhD01[6] - que foi limpo de acordo com as diretrizes descritas na secção anterior (cf. Secção 2.2.). Este foi inserido na sua forma integral no programa de extração automática das ocorrências e das frequências Sketch Engine (Kilgarriff *et al*., 2014). Foi aplicada a lista de *stop words*[7] para eliminar palavras funcionais (e.g. preposições, conjunções, determinantes, numerais, pronomes, etc.) das posições das frequências mais altas, pois nesta fase não acrescentam informação interessante à análise que está a ser realizada.

Num segundo momento, repetiu-se o mesmo processo com o mesmo texto, desta vez, não na sua forma integral, mas depois de eliminadas as secções que em termos de conteúdo, se afastavam mais do tópico da arte pública.

A resposta que obtivemos do programa foi de 71 819 ocorrências totais (*word tokens*) e 4 753 formas distintas (*word types*) para o texto completo, 54 073 ocorrências totais e 4 155 formas distintas para o texto selecionado.

(2) a. Texto completo b. Texto selecionado

	Lemma	↓ Frequency
1	estrutura	990
2	público	912
3	água	804
4	frente	785
5	arte	706
6	espaço	701
7	articulação	551
8	cidade	497
9	urbano	372
10	elemento	338
11	poder	286
12	colocação	278
13	ir	273
14	ter	251
15	forma	226

	Lemma	↓ Frequency
16	físico	215
17	monumento	207
18	obra	199
19	percurso	197
20	outro	196
21	lisboa	195
22	praça	191
23	relação	169
24	todo	158
25	fator	155
26	diferente	155
27	constituir	153
28	eixo	148
29	considerar	145
30	território	126

	Lemma	↓ Frequency
1	público	800
2	arte	665
3	estrutura	621
4	água	572
5	frente	549
6	espaço	491
7	articulação	401
8	cidade	378
9	elemento	297
10	urbano	290
11	colocação	271
12	ir	226
13	poder	218
14	ter	204
15	monumento	203

	Lemma	↓ Frequency
16	obra	188
17	forma	182
18	percurso	179
19	lisboa	154
20	físico	153
21	praça	152
22	outro	144
23	relação	135
24	todo	121
25	diferente	121
26	considerar	116
27	constituir	112
28	fator	108
29	partir	101
30	investigação	100

Figura 1 Comparação das listas de palavras mais frequentes extraídas com o Sketch Engine do texto CART-PT-PhD01

[6] Martins Ochoa de Castro, A. R. (2011). *Cidade e frente de água. Papel articulador do espaço público*. Dissertação de doutoramento, Universitat de Barcelona.

[7] Lista de *stop words* utilizada: a,à; à; ainda; ao; aos; aquela; aquelas: aquele; aqueles; aquilo; as; às; assim; até; através; com; como; da; das; de; dela; delas; dele ; deles; depois; do; dos; e; ela; elas; ele; eles; em; então; entre; essa; essas; esse; esses; esta; estas; este; este; estes; eu; foi; isso; isto; já; já; lhe; lhes; mais; mas; me; mesmo; meu; meus; minha; minhas: muito; na; não; nas; nem; no; nos; nós; nossa; nossas; nosso; nossos; num; numa; o; os; ou; para; pela; pelas; pelo; pelos; por; porque; qual; quando; que; quem; se; sem; ser; será; seu; seus; só; sua; suas; também; te; tem; têm; teu; teus; tu; tua; tuas; um; uma; você; vocês; vos.

Este pequeno exemplo apresenta resultados por um lado surpreendentes, na medida em que contrariam as expectativas prévias, mas por outro parece oferecer respostas claras à pergunta inicial, ou seja, qual seria a melhor abordagem quanto à utilização dos textos no contexto específico deste trabalho.

Como se pode verificar pelas imagens retiradas a partir do próprio programa, as duas listas refletem frequências de ocorrências absolutas muito semelhantes. Embora haja uma diferença de quase 20 000 ocorrências entre o texto completo e o texto truncado, as diferentes formas não estão sujeitas a grande variação. Nomeadamente, as primeiras 27 entradas são exatamente as mesmas, com algumas diferenças apenas na ordem em que surgem.

Repetiu-se a mesma experiência com um segundo texto - CART-PT-PhD02[8] - mais uma vez pertencente à categoria tese de doutoramento, para excluirmos a hipótese de se tratar de um fenómeno casual ligado apenas a um texto específico. Os resultados confirmaram os resultados observados com o primeiro exemplo.

A resposta que obtivemos do programa foi de 25 5375 ocorrências totais e 10 197 formas distintas para o texto completo, 206 571 ocorrências totais e 9 177 formas distintas para o texto selecionado, sendo que, a semelhança do exemplo anterior, as formas que aparecem entre as frequências mais altas são essencialmente as mesmas em ambas as listas[9].

(3) a. Texto completo b. Texto selecionado

Figura 2 Comparação das listas de palavras mais frequentes extraídas com o Sketch Engine do texto CART-PT-PhD02

[8] Andrade Marques, I.M. (2012), *Arte e habitação em Lisboa 1945-1965. Cruzamentos entre desenho urbano, arquitetura e arte pública*. Dissertação de Doutoramento, Universitat de Barcelona.

[9] Importa realçar que as maiores diferenças se encontram entre os últimos elementos da lista devido às ligeiras diferenças na ordem (que reflete a frequência de ocorrência). Logo, muitos destes elementos têm correspondências em posições mais baixas daquelas apresentadas. No entanto, por razões de espaço teve de ser selecionada uma amostra reduzida da lista.

Com esta análise preliminar demonstrou-se que, embora as expectativas iniciais apontassem para uma dispersão lexical muito grande, no caso do texto integral, consoante a heterogeneidade dos tópicos dos diferentes capítulos, ambas as versões do texto apresentam um nível muito próximo de saturação lexical.

Além disso, é preciso, também, ter em conta que a manipulação do texto para a seleção dos capítulos implica decisões em grande medida subjetivas, que podem alterar a natureza dos textos e ter repercussões negativas na extração do léxico.

Quanto à frequência absoluta das ocorrências, o facto de serem textos grandes tem alguma consequência na repetição de um conjunto específicos de unidades lexicais em detrimento de outras, devido às escolhas e a preferência pessoal do autor em questão. No entanto este fator será tido em conta na seleção das entradas da *wordnet*, pois para além da frequência serão, também, avaliadas a distribuição ao longo do *corpus* e a pertinência relativamente ao domínio, validada pelo especialista. Relembramos que o *corpus* servirá de base à extração de unidades lexicais potencialmente pertencentes ao domínio de especialidade, mas não diretamente e apenas a partir da análise de frequências, como descrito em outras secções deste artigo.

Chegou-se, portanto, à conclusão de que, vista a qualidade dos resultados obtidos a partir dos textos integrais quando considerados em comparação com os obtidos com os fragmentos mais pertinentes, e tendo em conta naturalmente o investimento de tempo necessário para selecionar manualmente os textos, o esforço de preservação do equilíbrio em termos de tamanho dos textos incluídos no *corpus* não teria como resultado mais-valias substanciais para investigação. Assim sendo optou-se por manter os textos integrais.

2.4. Descrição quantitativa do CORPORART

Uma vez analisadas as especificidades e as questões problemáticas envolvidas na constituição do CORPORART, cabe agora fazer uma descrição mais detalhada dos conteúdos do *corpus*.

A tabela apresenta uma primeira aproximação da quantificação das ocorrências, catalogadas por tipo de texto. A contagem é ainda aproximativa pois, nesta fase, a limpeza dos textos ainda não se encontra completada. Assim, espera-se que o cômputo total, numa fase mais avançada do trabalho, possa sofrer algumas alterações moderadas, mantendo-se, no entanto, as proporções, inalteradas.

TIPO DE TEXTO	PT (nº palavras)	IT (nº palavras)	Abreviatura
Tese Mestrado	1 027 646	550 332	MA
Tese Doutoramento	1 242 155	68 338	PhD
Cap. Tese Doutoramento	11 408	17 981	CPhD
***Paper* - Conferência**	69 001	0	Pap
Artigo - Revista	183 689	143 042	Art
Revista completa	133 688	105 446	Jou
Livro	52 241	55 419	Book
Capítulo de livro	73 684	107 543	Cbook
Catálogo	8 784	8 438	Cat
Relatório	0	8 482	Law
Lei/Regulamento	8 524	9 142	Rep
Edital de concurso	18 364	21 938	Tend
Tot.	**2 829 184**	**Tot. 1 096 101**	

Tabela 1 Análise quantitativa do CORPORART por tipo de texto e por língua representada

Como se pode verificar na tabela 1, nota-se uma tendência geral do *corpus* para ter mais ocorrências para o português, sendo que a evidente discrepância que existe entre a contagem total das ocorrências nas duas línguas representadas é em grande medida atribuível à falta de trabalhos académicos e artigos científicos em italiano disponíveis, cujos motivos foram descritos anteriormente (cf. secção 2.3).

Através desta comparação em termos quantitativos, resulta de forma muito evidente quais as implicações que as políticas de acesso fortemente restritivas no contexto italiano tiveram na recolha da amostra, em particular no que diz respeito aos textos académicos. Sem dúvida, é preciso também ter em conta o facto de, hoje em dia, a predominância do Inglês nas publicações académicas fazer com que haja cada vez menos textos especializados produzidos noutras línguas. Neste contexto é preciso fazer ainda mais uma distinção relativamente às duas línguas representadas no CORPORART: os trabalhos realizados em Português conseguem ter alguma visibilidade, pois existe uma comunidade lusófona razoavelmente ampla (com o PT a ser língua oficial em nove países), sendo que para o Italiano a comunidade de falantes é significativamente mais restrita. Isto implica uma redução radical do uso do italiano para a divulgação científica. Somando, então, todos os fatores descritos neste artigo decorre que o trabalho de recolha dos dados foi muito complexo e desafiante.

Quanto às áreas disciplinares abrangidas pelo *corpus*, há uma sobreposição nos dados de ambas as línguas. As principais áreas cobertas são: belas artes e estudos artísticos; arquitetura; urbanismo e *design* urbano; sociologia e antropologia; comunicação e cultura visual; conservação e restauro dos bens culturais, gestão do património e mercado das artes.

Portanto, tendo em conta as evidências que este estudo ressalta, terá de ser avaliada novamente a metodologia de investigação que depende fortemente dos dados apresentados pelo *corpus*. Nomeadamente terá de ser revista a questão da comparabilidade dos *subcorpora*, pois a disparidade de dados disponíveis não suporta a hipótese de um trabalho em paralelo nas duas línguas.

Ainda assim, este contexto de trabalho, para além de ser desafiante, oferece a possibilidade de avaliar de forma concreta a relação em termos de proporção que existe entre o número de ocorrências e a efetiva cobertura dos dados.

Segue que, da mesma forma que ao processar os textos chegamos à conclusão de que o aparente "ruído" inicialmente considerado como potencialmente contraproducente em termos de dispersão lexical, na prática não teve nenhum efeito, nem positivo nem negativo, também o facto de um dos *corpora* ser substancialmente mais pequeno que o outro poderá não ter um impacto negativo proporcional à diferença quantitativa no que toca à extração do léxico de especialidade. Neste sentido, uma vez que verificamos que ao omitir uma parte dos textos a lista de ocorrências não sofreu alterações significativas, poder-se-ia colocar a hipótese de que é possível atingir um nível de comparabilidade razoável em termos de cobertura lexical, apesar do inegável contraste que existe em termos de quantidade de dados para as duas línguas representadas no *corpus* (Morin & Hazem, 2014).

Uma vez que a frequência é um dos critérios levados em consideração para a extração das unidades lexicais (cf. secção 2.3), será necessário recorrer a algumas medidas estatísticas simples para uniformizar os dados e torná-los numericamente comparáveis – e.g. a frequência relativa normalizada por um milhão[10]. Veja-se os exemplos a seguir com as unidades multilexicais *espaço urbano* (CORPORART-PT) e *spazio urbano* (CORPORART-IT); *arte pública* (CORPORART-PT) e *arte pubblica* (CORPORART-IT) que, de acordo com frequência de ocorrência e com a distribuição homogénea ao longo do *corpus*, serão candidatos potenciais para integrar a *wordnet*:

[10] Frequência relativa: $\frac{\text{frequência de ocorrência absoluta} \cdot 1\,000\,000 + 1}{\textit{corpus} \text{ total}}$

(4) a. e*spaço urbano*:
Frequência absoluta: 1060 ocorrências
Frequência relativa: 374,6
b. s*pazio urbano*:
Frequência absoluta: 408 ocorrências
Frequência relativa: 372,2

Se pelas frequências absolutas seria impossível comparar dados tão distintos, as frequências relativas fornecem uma base partilhada para comparar os dados dos dois *corpora*, sendo que neste caso o valor muito próximo sugere que o peso do elemento lexical em análise é praticamente equivalente em ambos os lados.

Curiosamente, com *arte pública* e *arte pubblica* os resultados obtidos não confirmam o paralelismo que seria expectável, à semelhança do exemplo anterior.

(5) a. *arte pública*:
Frequência absoluta: 6823 ocorrências
Frequência relativa: 2411
b. *arte pubblica*:
Frequência absoluta: 854 ocorrências
Frequência relativa: 779

Ao contrariar as expectativas, estes resultados oferecem pistas de análise muito interessantes, como por exemplo avaliar se: (i) em italiano é preferido o uso do equivalente inglês (*public art*) à expressão italiana, (ii) coexistem muitas variantes ortográficas (*arte pubblica*, *Arte pubblica*, *Arte Pubblica*, *Public Art*, *public art* etc.) que podem influenciar as contagens, ou ainda (iii) se privilegia o uso de sinónimos ou merónimos mais específicos (*arte urbana, street art, estatuária urbana, arte site-specific, land art,* etc.).

3. Considerações finais

Neste trabalho foi apresentado o CORPORART, um *corpus* bilingue de arte pública, desde o processo de definição de projeto de compilação dos dados, que teve em conta o contexto de investigação mais abrangente em que o *corpus* se enquadra, até ao produto final.

Considerada a complexidade e extensão que o domínio de especialidade em questão apresenta, realizou-se primeiramente uma caracterização da área da arte pública, de forma a sustentar as escolhas feitas ao longo do processo de seleção dos textos e a descrever os futuros desafios que se prevê poderem surgir no que diz respeito ao tratamento das unidades lexicais extraídas a partir do *corpus*. Procedeu-se depois à descrição dos critérios de seleção dentro do universo potencial e dos critérios de organização dos conteúdos do *corpus*, em função dos conceitos de representatividade, equilíbrio e qualidade, adaptados aos objetivos e à perspetiva de trabalho considerada.

De acordo com os dados apresentados verificou-se um contraste muito grande, nomeadamente em termos quantitativos, entre o material recolhido no âmbito do contexto português (sobretudo nas produções académicas) e o contexto italiano. No entanto, os resultados observados a partir da análise de alguns textos, nomeadamente duas teses de doutoramento, revelaram pistas muito interessantes para a definição da noção de representatividade dos *corpora*, em particular dos *corpora* de especialidade, tendo em conta a cobertura lexical abrangida. Em concreto, embora a análise quantitativa do CORPORART realce uma diferença entre os dois *subcorpora* (CORPORART-IT e CORPORART-PT) em termos proporcionais de cerca de 1 para 3, a primeira aproximação aos dados aqui apresentada oferece elementos que permitem levantar a hipótese de que

a cobertura lexical potencialmente alcançada com o CORPORART-IT poderá não espelhar exatamente esta proporção, já que a experiência feita com o truncamento de textos extensos indica que a quantidade de dados necessários para se obter uma cobertura lexical representativa de um domínio de especialidade poderá ser mais reduzida do que inicialmente antecipado. Neste sentido o CORPORART adquire relevância para a comunidade científica, quer enquanto recurso de especialidade disponível e de livre acesso, quer enquanto contributo para o debate acerca da noção de representatividade no âmbito da linguística de *corpus*, sobretudo no que concerne à recolha e à utilização de *corpora* de pequenas dimensões para fins de extração lexical.

Note-se ainda que o CORPORART, para além de representar uma fonte de dados fundamental para o projeto de investigação mencionado, tenciona também ser um recurso disponibilizado online e de livre acesso, mediante licença de download para utilização para fins não comerciais, (www.clunl.fcsh.unl.pt), tanto para a comunidade científica como para utilizadores não especializados. Neste último caso, avaliar-se-á a disponibilização de ferramentas de suporte adicionais como, por exemplo, um manual de utilização.

Considerada a complexidade do desafio de reunir informação tão dispersa e construir um *corpus* de arte pública, sem obviamente ignorar as limitações ao nível da sua estrutura e organização anteriormente mencionados, acreditamos poder ser este um recurso útil e proveitoso, tanto para consultas individuais como para investigação futura. A partilha do trabalho científico é fundamental para um contínuo avanço na investigação, mas para que isto seja possível é necessária uma descrição detalhada e transparente dos conteúdos dos recursos. Neste sentido, o CORPORART tem para cada texto um cabeçalho com a informação necessária para a sua identificação, facilitando assim a utilização e extensão futura do *corpus*.

Referências

Abreu, J. G., & Castro, L. (2018) Rede de Informação, Investigação e Intervenção em Arte Pública. *Encontro Ciência '18*. Lisboa.

Amaro, R., & Mendes, S. (2012) Towards merging common and technical lexicon wordnets. Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon: 24th International Conference on Computational Linguistics. Mumbai, pp. 147-160.

Armajani, S. (1995) Manifiesto: La escultura publica en el contexto de la democracia norteamericana. *AA. VV. Espacios de lectura*. Barcelona: MACBA, pp. 35-37.

Bargna, I. (2012) Nessuna partecipazione senza distanza. Quel che l'arte pubblica e partecipativa mettono in gioco. *Africa e Mediterraneo* 21, pp. 2-5.

Biber, D. (1993) Representativeness in Corpus Design. *Literary and linguistic computing 8(4)*, pp. 243-257.

Elias, H., & Valente, C. (2017) Da Utopia à Distopia. O Mural como Ferramenta Participativa nos Espaços Públicos da Cidade. *Convocarte, REVISTA DE CIÊNCIAS DA ARTE* 5, pp. 1-19.

Fondazione Contrada Torino Onlus. (sem data) Obtido de https://contradatorino.org/arte-pubblica/

Guilbert, L. (1971). De la formation des unités lexicales. *In Grand Larousse de la langue française, 1*.

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

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014) The Sketch Engine: ten years on. *In Lexicography 1(1)*, pp. 7-36.

Lenci, A., Montemagni, S., & Pirrelli, V. (2016) *Testo e computer*. Roma: Carocci editore.

Marques, I. M. A. (2012) *Arte e habitação em Lisboa 1945-1965. Cruzamentos entre desenho urbano, arquitetura e arte pública*. Dissertação de doutoramento. Universitat de Barcelona.

Marrafa, P., Amaro, R., Chaves, R. P., Lourosa, S., Martins, C., & Mendes, S. (2005) WordNet.PT – Uma rede léxico-conceptual do Português on-line. *XXI Encontro da Associação Portuguesa de Linguística*.

Martins Ochoa de Castro, A. R. (2011) *Cidade e frente de água. Papel articulador do espaço público.* Dissertação de doutoramento, Universitat de Barcelona.

McEnery, T., Xiao, R., & Tono, Y. (2006) Representativeness, balance and sampling. *In Corpus-based Language Studies: An Advanced Resource Book*, Routledge, pp. 1-8.

Morin, E., & Hazem, A. (2014) Looking at unbalanced specialized comparable corpora for bilingual lexicon extraction. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics 1, pp. 1284-1293.

Regatão, J. P. (2016) Prólogo. *Convocarte, REVISTA DE CIÊNCIAS DA ARTE 1*, pp. 12-13.

Remesar, A., & Brandão, P. (2010) Prólogo. *In Arte Pública e Cidadania: novas leituras da cidade criativa*. Lisbon: Caleidoscópio, pp. 5-11.

Salem, A. (1987) *Pratique des segments répétés: essai de statistique textuelle*. Klincksieck éditions.

Santos, F. M. P. G. (2016) *Graffiti e Arquitetura: intervenções no espaço e na memória da cidade.* Dissertação de mestrado. Universidade Lusíada.

Sharp, J., Pollock, V., & Paddison, R. (2005) Just Art for a Just City : Public Art and Social. *Urban Studies, 42*, pp. 1001–1023.

Simões, M. C. (2013) *Graffiti e Street Art em Portugal*. Dissertação de mestrado. Faculdade de Letras, Universidade de Lisboa.

Sinclair, J. (1991) *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. (1996) *Preliminary recommendations on Corpus Typology*. Birmingham.

Sinclair, J. (2004) Corpus and Text — Basic Principles. In *Developing Linguistic Corpora: a guide to good practice*. Oxford: Oxbow Books, pp. 3–22.

Tognini-Bonelli, E. (2001) *Corpus linguistics at work*. Amsterdam-Philadelphia: John Benjamins Publishing Company.

Valente, C. (2016) *A Street Art no femenino. O lugar da mulher na Arte Pública.* Dissertação de mestrado. Faculdade de Letras, Universidade de Lisboa.