RELATÓRIO

1.ª OFICINA DE TRABALHO SOBRE DISCURSO ACADÉMICO

Princípios teóricos e metodológicos para o uso de corpus numa investigação

Mafalda Mendes (CELGA-ILTEC, UC)

Mário Martins (UFERSA)

Marta Sitoe (UEM)

Emília Marrengula (UEM)

Lisboa/Leiria/Mossoró/Maputo, Novembro de 2022

FCT Fundação para a Ciência e a Tecnologia
CELGA-ILTEC Centro de Estudos de Linguística Geral e Aplicada da Universidade de Coimbra
POLITÉCNICO DE LEIRIA
REPÚBLICA PORTUGUESA

ISBN digital: 978-989-8797-74-2

DOI: https://doi.org/10.25766/txsg-pf19

Índice

1. Introdução

A oficina de trabalho *Princípios teóricos e metodológicos para o uso de corpus numa investigação* constitui a primeira oficina da série *Oficinas de Trabalho sobre o Discurso Académico*, uma iniciativa do Grupo de Trabalho Discurso e Práticas Discursivas Académicas (DPDA), do CELGA-ILTEC, da Universidade de Coimbra, e Escola Superior de Educação e Ciências Sociais (ESECS), do Instituto Politécnico de Leiria. Esta primeira oficina teve como objetivo principal instrumentalizar os participantes com princípios teóricos e metodológicos para o uso de corpora em projetos de investigação em linguística aplicada à educação. Intercalando-se momentos de exposição e discussão teórica com atividades práticas, os organizadores da oficina pretenderam proporcionar aos participantes um espaço de coconstrução gradual do conhecimento especializado sobre corpora, combinando teoria e prática.

Para a consecução do objetivo da oficina, o seguinte roteiro foi proposto (ver Anexo I para uma descrição detalhada):

- Apresentação dos organizadores, dos participantes e do plano de trabalho;
- Sessão 1: Princípios teóricos - atividades teórico-práticas;
- Sessão 2: Questões de exploração de corpus - atividades teórico-práticas;
- Sessão 3: Princípios de desenho de corpus - atividades teórico-práticas;
- Atividade da pergunta de trabalho;
- Balanço do trabalho realizado.

Após a realização da oficina, era esperado por parte dos participantes a consciencialização dos princípios teóricos e metodológicos da investigação linguística baseada em corpus e o acesso a ferramentas de trabalho de referência para a exploração das potencialidades operacionais na área da linguística de corpus.

2. Local e data

A oficina realizou-se no dia 27 de novembro de 2021, na modalidade mista (online e presencial). Os trabalhos foram conduzidos na plataforma Zoom, a partir da conta institucional da Escola Superior Educação e Ciências Sociais (ESECS), do Instituto Politécnico de Leiria. Foi aberta a possibilidade de participação nos trabalhos da oficina em regime presencial nas instalações da ESECS.

Os trabalhos tiveram início às 9h30 e terminaram às 17h00, com intervalo de uma hora quinze minutos para almoço e dois intervalos de quinze minutos, um na parte da manhã e outro na parte da tarde.

3. Destinatários

A oficina tinha como público-alvo quaisquer investigadores interessados em introduzir metodologias da linguística de corpus nos seus projetos de investigação em linguística aplicada. Por se tratar de uma oficina de natureza introdutória, não foram exigidos conhecimentos prévios na área da linguística de corpus.

Participaram na oficina 16 investigadores filiados em doze instituições académicas, de quatro países: Portugal, Brasil Moçambique e China:

Universidade de Aveiro	Portugal
Universidade de Coimbra	Portugal
Universidade de Lisboa	Portugal
Instituto Politécnico de Leiria	Portugal
Universidade Aberta	Portugal
Universidade de Trás-os-Montes e Alto Douro	Portugal
Universidade do Minho	Portugal
Universidade Federal do Tocantins	Brasil
Universidade Federal Rural do Semi-Árido	Brasil
Universidade de Estudos Internacionais de Xi'an	China
Universidade Eduardo Mondlane	Moçambique

4. Justificação

A realização desta oficina, centrada em princípios, procedimentos e técnicas fundamentais da linguística de corpus, justifica-se no fato de que permite a divulgação e a promoção, em contexto compartilhado, tanto no âmbito do DPDA, quanto no de outras instituições, de mais uma possibilidade de aparato metodológico em investigações sobre fenómenos linguísticos/textuais. A linguística de corpus é um sistema completo de métodos e princípios (McEnery *et al,* 2006) que pode ser aplicado para responder a questões relacionadas com o uso da língua e a sua variação nos diversos contextos em que é usada.

Para tanto, apresentamos, na oficina, um conjunto de princípios teóricos e metodológicos que, de acordo com a literatura na especialidade, importa ter em consideração no desenho de um projeto de investigação em que se opte pelo recurso à linguística de corpus como metodologia para responder a uma pergunta de investigação.

5. Sessões de trabalho

A oficina desenvolveu-se em 3 sessões de atividades teórico-práticas:

- Sessão 1: Princípios teóricos;
- Sessão 2: Questões de exploração de corpus;
- Sessão 3: Princípios de desenho de corpus.

Para as atividades práticas de exploração de dados de corpora, foi utilizada a ferramenta Sketch Engine[1] (Kilgarriff *et al*., 2014). Para esse efeito, na véspera da oficina, foi solicitado aos participantes que fizessem o registo na plataforma, tendo-lhes sido facultadas as instruções para esse efeito. O Sketch Engine é uma ferramenta de alojamento de corpora dotada de um conjunto de funções avançadas de exploração e análise de dados de língua disponibilizadas através de uma interface amigável e intuitiva. O acesso ao Sketch Engine é livre e gratuito para os membros de instituições académicas na União Europeia.

Foi também referida na oficina a ferramenta TEITOK[2], uma outra plataforma de armazenamento e exploração de corpora, utilizada em vários projetos de linguística de corpora apresentados, tais como o Corpus do Português Académico[3] e o MOZEA[4] (Silva, Santos & Sitoe, 2019) e ainda o DOESTE[5] (Martins & Mendes, 2021) .

Os participantes da oficina foram organizados em pares para efeitos de realização das atividades práticas propostas no decurso dos trabalhos.

5.1. Sessão 1: Princípios teóricos

A primeira sessão foi dedicada à apresentação de um conjunto de princípios teóricos subjacentes à linguística de corpus que decorrem do reconhecimento da existência de relações sistemáticas entre as regularidades da cultura e as regularidades da língua.

A sessão teve início com um exercício prático de reconhecimento intuitivo de padrões linguísticos evidenciados em fragmentos de língua descontextualizados e respetivo relacionamento com tipos de situação reconhecíveis pelos membros da cultura.

A partir de vinte fragmentos extraídos de sete textos de diferentes tipologias: horóscopo, receita de culinária, interação conversacional casual, notícia, acórdão de tribunal, artigo de divulgação científica, carta de motivação, pediu-se aos participantes que distribuíssem por sete colunas, uma por cada texto de origem dos fragmentos, os vinte fragmentos de língua aleatoriamente ordenados.

Tomando por referência os modelos teóricos da linguística funcionalista de raiz firthiana e a literatura específica da linguística de corpus (Biber & Conrad, 2019; Berber-Sardinha,

[1] https://auth.sketchengine.eu/#login?next=https%3A%2F%2Fapp.sketchengine.eu%2F
[2] http://teitok.corpuswiki.org/
[3] http://teitok2.iltec.pt/cpa/
[4] http://teitok2.iltec.pt/mozea/
[5] https://doeste.ufersa.edu.br/

2004; Weisser, 2015; Stefanowitsch, 2020, entre outros) foram trabalhadas as seguintes noções teóricas:

- Relação entre língua e cultura: projeção das regularidades dos significados da cultura ao nível da língua;
- Variação sistemática das estruturas e significados da língua em função da variação ao nível das atividades no contexto de cultura;
- Modelo hallidayano de cultura/língua, tipo de situação/registo, contexto de situação/texto;
- Probabilidade;
- Gramática como sistema probabilístico;
- Linguística de *corpus* como metodologia de investigação linguística;
- Linguística de *corpus* como método de validação do julgamento dos falantes;
- Definições de corpus.

No final da sessão, foram apresentados alguns exemplos de *corpora* disponibilizados em linha em regime de acesso livre (Corpus do Português Académico (Silva, Santos & Sitoe, 2019), Doeste (Martins & Mendes, 2021), Mozea (Silva, Santos & Sitoe, 2019).

5.2. Sessão 2: Questões de exploração de corpora

A segunda sessão da oficina foi dedicada à apresentação de três tipos básicos de análise de dados linguísticos proporcionados pela linguística de *corpus*:

- Análise de frequências;
- Extração de concordâncias;
- Extração de colocações.

Optou-se por uma metodologia expositiva sistematicamente ancorada na apresentação de resultados de pesquisas previamente realizadas em corpus e/ou em exercícios práticos realizados pelos participantes da oficina na ferramenta Sketch Engine.

Foi realizada a atividade prática preparatória de carregamento de um corpus para a plataforma do Sketch Engine, seguida da compilação do corpus: tokenização, lematização e anotação morfossintática.

Nesta sessão, os participantes tiveram oportunidade de se familiarizar com as ferramentas básicas de pesquisa no Sketch Engine - Wordlist, Concordance e Word Sketch -, e explorar as suas potencialidades em diferentes contextos de investigação linguística, tais como a linguística aplicada ao ensino de línguas, a análise de conteúdos ou construção de glossários, dicionários ou outras recursos de descrição da língua contextualmente situados.

5.2.1. Análise de frequências

Nesta parte da sessão 1, foi apresentada a noção básica de que, em última análise, tudo em linguística de corpus se baseia na verificação da presença ou ausência de sequências de caracteres. Tudo o que se faz em corpora envolve, qualquer que seja o nível de análise, a computação de frequências de ocorrências (Paquot & Gries, 2020), ou seja a computação do número de vezes que algum fenómeno da língua, registado como uma sequência de caracteres, acontece num corpus: nunca, algumas vezes, muitas vezes.

Os participantes da oficina foram convidados a validar intuições linguísticas acionadas para a resolução de exercícios de recuperação de padrões de língua em contextos linguísticos lacunares por confronto com análises de frequência de padrões observados no corpus Portuguese Web 2011 (ptTenTen1), disponível no Sketch Engine. Por exemplo, pediu-se a validação do termo "semelhança", proposto pela maioria dos participantes para o exercício de recuperação do elemento vocabular lacunar no contexto: "Assim foi a sociedade brasileira moldada **à imagem e** _____ da portuguesa."

A intuição foi validada pelo resultado da análise de frequências das unidades lexicais ocorrentes na vizinhança direita da sequência "à imagem e": o lema "semelhança" surge como o mais frequente neste contexto, com frequência relativa de 0.45, a uma distância de 0.33 do segundo elemento lexical na lista de ocorrências ordenada por ordem decrescente de frequência.

A partir da demonstração de como a informação linguística é codificada num corpus informatizado como sequências de caracteres, que podem variar desde as sequências de caracteres que instanciam as unidades da língua que compõem os textos (o texto como uma sequência de caracteres que representam as palavras, os sinais de pontuação e os espaços que separam as palavras) até às sequência de caracteres que compõem as diferentes camadas de marcação e anotação eventualmente adicionadas às unidades lexicais ou sintáticas que compõem os textos, os participantes foram levados a refletir sobre os diversos tipos de unidades linguísticas que podem ser alvo de análises de frequências: formas lexicais simples, lemas, unidades lexicais complexas, classes de palavras (*pos*), diferentes atributos linguísticos, etc.

Por fim, mediante a observação de tabelas de resultados de extração de dados de frequência no Sketch Engine, foram apresentadas as medidas básicas de apresentação de resultados de frequências:

- Frequência bruta: número absoluto de vezes que um elemento ocorre num corpus.
- Frequência normalizada por milhão de palavras: estimativa de número de vezes que, num corpus com a dimensão de um milhão de palavras, essa unidade ocorreria.

A frequência por milhão de palavras é uma medida que permite comparar frequências entre corpora de diferentes dimensões.

5.2.2. Análise de concordâncias

De seguida, ainda na sessão 1, foi apresentada a definição básica de concordância, como uma lista de ocorrências de um termo de busca, apresentadas dentro do contexto das suas ocorrências - usualmente um pequeno conjunto de palavras à esquerda e à direita do termo de busca, como se ilustra a seguir:

Foi apresentada e exemplificada, em resultados de pesquisas realizadas no Sketch Engine, a terminologia básica no trabalho de extração de concordâncias:

- Nó ou KWIC (*key words in context*);
- Janela ou contexto;
- Linhas de concordância;
- Concordância;
- Colocados (palavra imediatamente à esquerda ou à direita do nó).

Imagem 1: Ilustração da terminologia básica no trabalho de extração de concordâncias - exemplo de visualização de resultados de extração de concordâncias do lema "escrita" no Sketch Engine.

5.2.3. Extração de colocações

Por último, foi apresentado e discutido o conceito de colocação. O fenómeno da colocação funda-se no facto de, em certos contextos, certas palavras terem maior probabilidade de ocorrer em combinação com outras (Baker, Hardie & Mcenery, 2006). São considerados colocados palavras ou classes de palavras que tendem a ocorrer na vizinhança de outras.

A extração de colocações é obtida por aplicação de diferentes métodos estatísticos que consideram principalmente: a distância, a frequência e a exclusividade. As ferramentas estatísticas de extração de colocações constituem um importante contributo para a descrição de padrões de regularidade na língua.

A apresentação do conceito de colocação e o seu potencial descritivo foi exemplificada e discutida em exercícios práticos desenvolvidos pelos participantes na ferramenta Word Sketch, do Sketch Engine. O Word Sketch é uma poderosa ferramenta de extração de

colocações que fornece possibilidades avançadas de análise de dados e visualização de resultados. A extração de colocações para a língua portuguesa do Word Sketch tem por base o trabalho *Tanara Kuhn's Portuguese sketch grammar v5* (Khun 2016, 2017).

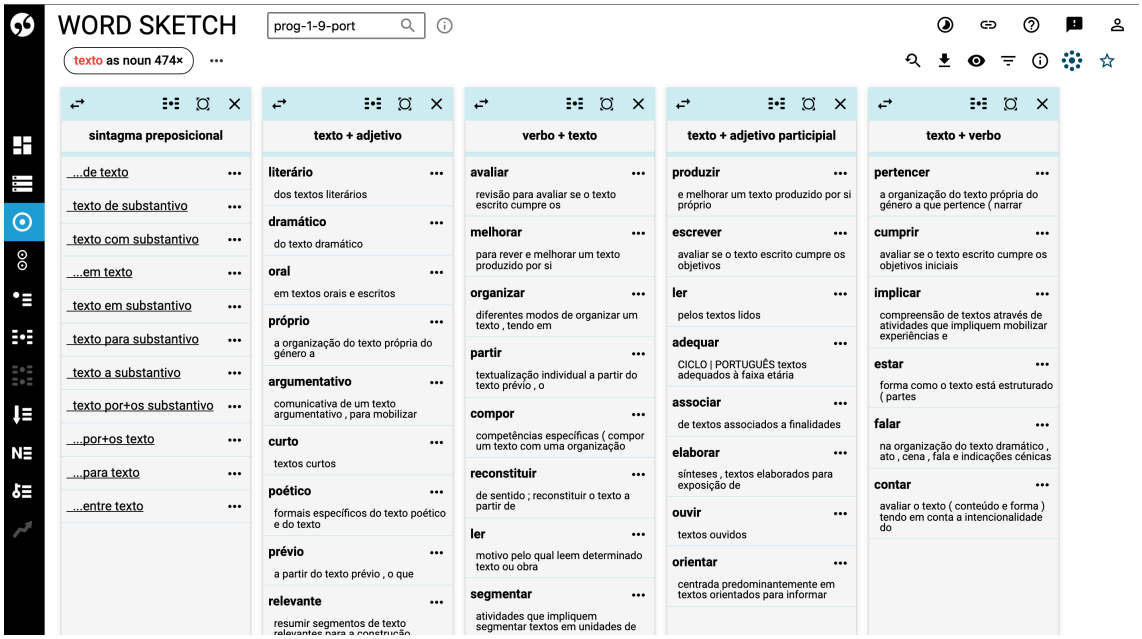

Imagem 2: Print screen do painel de resultados de extração de colocações do lema "texto", pelo Word Sketch, num corpus constituído pelo texto do Programa de Português do Ensino Básico.

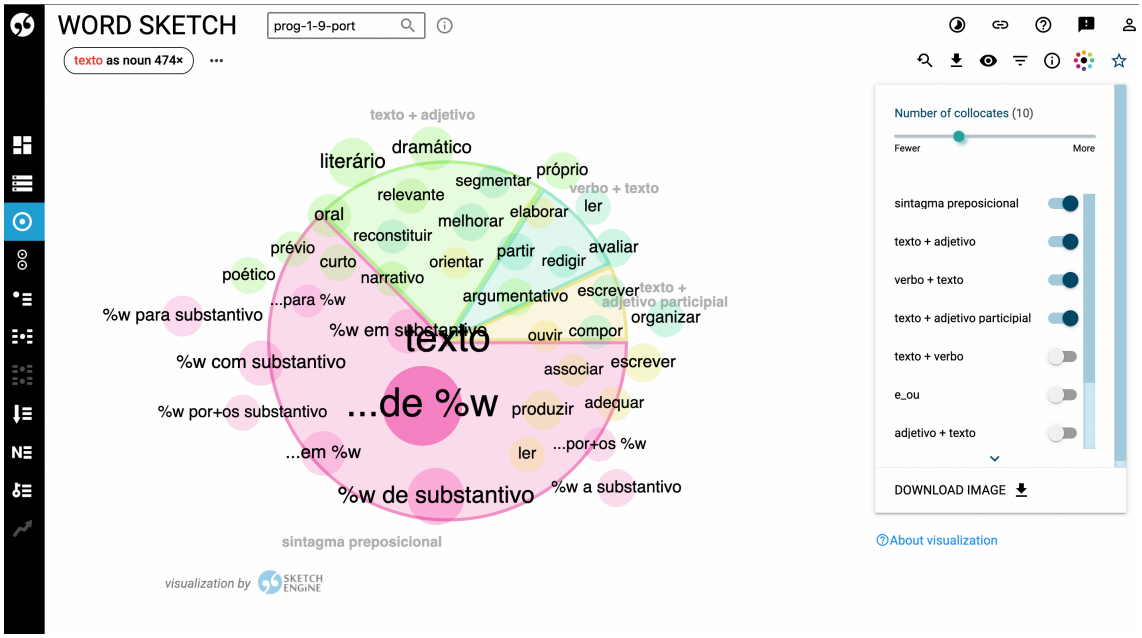

Imagem 3: Print screen do painel de visualização gráfica dos resultados da extração de colocações do lema "texto", pelo Word Sketch, num corpus constituído pelo texto do Programa de Português do Ensino Básico.

5.3. Sessão 3: Questões metodológicas de desenho de corpus

Na terceira sessão, de conformação expositiva, discutiram-se algumas questões seminais a serem consideradas no desenho de um corpus de investigação científica. A partir de Biber (1993, p. 243), primeiramente tratou-se da representatividade, entendida como a "extensão em que um corpus reflete a gama completa de variabilidade na população", e do balanço, enquanto "a proporção das diferentes amostras incluídas num corpus". Representatividade e balanço ligam-se ao facto de o objetivo na compilação de um corpus é que este deve ser uma prática maximamente representativa, o que se traduz numa amostra aceitavelmente representativa de uma população de utilizadores de línguas, de uma variedade linguística, ou de um tipo de discurso.

De seguida, ainda nesta sessão, apresentaram-se questões pragmáticas sobre o desenho do corpus intrinsecamente relacionadas com os condicionalismos associados à coleta de dados, tais como a quantidade de textos necessária para a composição da amostra, a tipologia dos dados de língua a coligir (questões de modalidade, suporte dos dados, acessibilidade, etc.) e a necessidade ou não de pré-tratamento (por exemplo, transcrição, normalização ortográfica, limpeza de dados extralinguísticos, uniformização, etc.). De importância central na linguística de corpus, foram discutidas nesta sessão questões de natureza ética, pelo que se resumiram orientações que expressam, entre outras possíveis, a necessidade de consentimento informado de uso de dados pelos informantes e eventualmente a aprovação das instituições (por exemplo, textos recolhidos em contexto educacional). Além disso, foram enfatizados pontos importantes da configuração ética do desenho de um corpus, tais como a necessidade de anonimização de dados ou de armazenamento seguro.

Em continuidade, discutiu-se sobre a documentação dos conteúdos do corpus, sob a forma de metadados, o que inclui, por exemplo, informação demográfica sobre a população representada ou sobre as situações de uso (objetivos da comunicação, atividade em curso, relações entre os participantes, modalidade de uso, data de produção, origem dos textos recolhidos, etc.), cruciais para a contextualização e interpretação dos dados linguísticos coligidos no corpus e também para o processo de divulgação do corpus e dos resultados das investigações a partir dele desenvolvidas. Ainda relativamente ao processo de desenho de um corpus, foram abordadas, na parte final da terceira sessão da oficina, as questões de formatação, marcação e anotação dos dados linguísticos, enquanto meios de registo de informação linguística e textual, intrinsecamente relacionadas com os objetivos investigativos que se pretendem atingir com a constituição do corpus.

Por fim, foi referida a importância de prever as questões relativas à disponibilização do corpus à comunidade, nomeadamente no que respeita às decisões relativamente a licenciamento do acesso aos dados, disponibilidade de espaço em servidores, escolha de plataformas de alojamento, gestão e exploração dos dados. Foi ainda referida a importância de assegurar a assessoria técnica em linguística computacional, tanto para a fase de compilação dos dados do corpus, como também para as tarefas de exploração dos dados.

6. Avaliação

6.1. Avaliação pela equipe dinamizadora da oficina

Após a oficina, a equipa dinamizadora reuniu-se a fim de realizar um encontro de autoavaliação. Uma primeira avaliação positiva diz respeito à integração da diversidade de contextos de trabalho dos integrantes da equipa dinamizadora. Essa integração só foi possível em virtude do caráter remoto da oficina. A avaliação pela equipa dinamizadora centrou-se em três grandes eixos relacionados à oficina: as estratégias didáticas, a seleção de conteúdos e o sistema de avaliação.

6.1.1. Estratégias didáticas

Toda a oficina foi organizada em torno da ideia de uma aprendizagem prática, graduada e em espiral (em que conteúdos seriam sempre acionados em diferentes tempos da oficina). Recorreu-se, assim, a uma distribuição de conteúdos que permitisse a um só tempo o contato com novas informações e a possibilidade de as manipular em situações práticas. Uma maior exposição dos participantes a atividades práticas de reforço das aprendizagens dos aspetos teóricos poderia ter sido um elemento impulsionador de mais interação entre equipa dinamizadora e participantes.

Ainda quanto à interação, o caráter remoto da oficina, por um lado, pareceu não contribuir muito para o estabelecimento de um espaço de diálogos mais intensos entre os participantes, como se costuma ter em contextos presenciais. Por outro lado, o ambiente virtual permitiu o uso de recursos digitais que facilitaram a exposição de conteúdos e de ferramentas de exploração de corpora.

6.1.2. Seleção de conteúdos

A seleção dos conteúdos foi cuidadosamente pensada à luz do público esperado, ou seja, investigadores sem conhecimentos prévios sobre princípios e metodologias vinculados à linguística de corpus. Desse modo, apresentaram-se aos participantes conteúdos e referências bibliográficas basilares que lhes permitiriam futuramente avançar em leituras mais especializadas e/ou lançar-se na utilização de softwares e plataformas computacionais de uso de corpora. Acreditamos que apresentar mais plataformas e softwares poderia ser uma mais-valia aos participantes.

6.1.3. Sistema de avaliação

Um exercício prático de apresentação, pelos grupos de participantes, do esboço de opções metodológicas a integrar o projeto de investigação para resposta à pergunta estímulo apresentada no início da oficina constava do planejamento da oficina, em que se poderia avaliar o atingimento ou não de algumas aprendizagens globais propostas pela equipa dinamizadora. Em virtude do tempo estendido das sessões anteriores, não chegou a realizar-se esse exercício prático.

6.2. Avaliação da oficina pelos participantes

Os participantes tiveram oportunidade de avaliar a oficina de trabalho mediante a resposta a um inquérito de satisfação disponibilizado, após o termo dos trabalhos, em formulário Google. Responderam ao inquérito 9 dos 16 participantes na oficina. Todas as respostas foram registadas no próprio dia da oficina e no dia seguinte.

O formulário propunha aos participantes que respondessem inicialmente a nove perguntas de resposta escalar (muito insatisfeito; insatisfeito; neutro; satisfeito; muito satisfeito). O gráfico abaixo sintetiza as respostas obtidas:

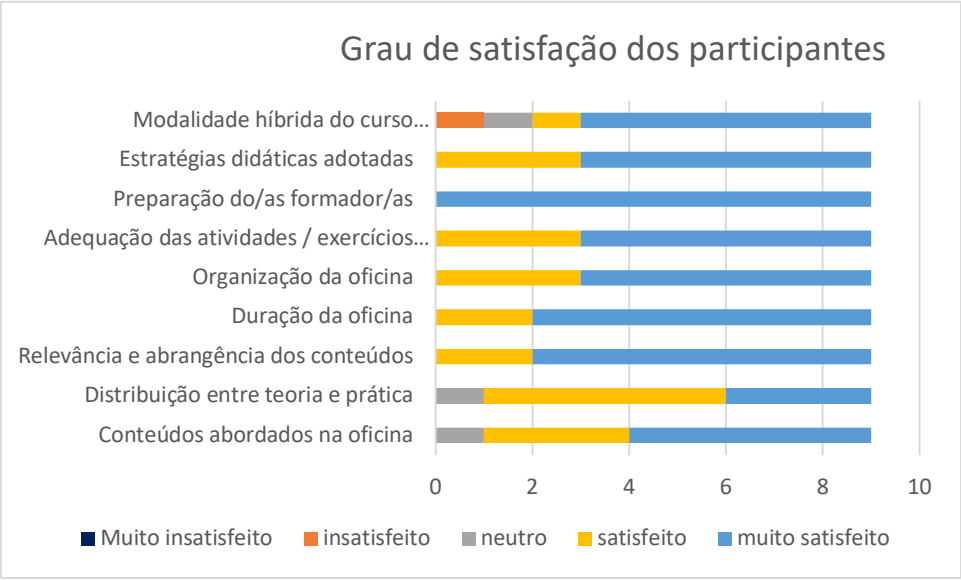

Gráfico: 1: Resultados das respostas sobre grau de satisfação relativamente a 9 parâmetros de avaliação da oficina.

O inquérito também incluiu os seguintes quatro itens de resposta aberta:

1. Que aspetos da oficina foram mais úteis?
2. Que aspetos podem ser melhorados nas próximas oficinas?
3. Sugestões de temas para as próximas oficinas.
4. Comentários / observações

Como aspetos mais úteis da oficina, foram destacados pelos participantes a oportunidade de familiarização com o Sketch Engine e o exercício de reflexão sobre as questões teóricas e metodológicas inerentes à linguística de corpus.

Relativamente à perceção do que poderia ser melhorado em oficinas futuras, os participantes referiram a possibilidade de alocação de mais tempo às atividades práticas, a abrangência das atividades a outras ferramentas de alojamento, gestão e análise de dados de corpora, a disponibilização de um guião aos formandos com as diferentes atividades a realizar e um resumo da informação a reter (quer para apoiar a realização das atividades durante o workshop, quer depois do workshop), a opção pela modalidade online para facilitar questões logísticas e a possibilidade de aumentar o número de vagas para participação na oficina.

Foram sugeridos para temas de oficinas futuras: Uso de Sketch Engine (avançado) e outras plataformas; anotação automática do português baseada em léxicos e/ou corpora existentes e o uso da linguística de corpus em pesquisa de abordagem qualitativa/quali-quantitativa.

No espaço aberto a comentários/observações, um dos participantes lamentou não ter havido tempo útil para, no fim da oficina, responder colaborativamente à questão de investigação proposta, por forma a aplicar os conhecimentos adquiridos. Um dos participantes lamentou que nenhum membro da organização tivesse participado presencialmente nos trabalhos. De modo geral, os participantes deixaram comentários positivos de agradecimento à equipa organizadora.

7. Agradecimentos

A organização da oficina contou com o apoio do grupo de trabalho Discurso e Práticas Discursivas Académicas, em especial na pessoa da IR, Doutora Marta S. Filipe Alexandre, e do Gabinete de Relações Públicas e Cooperação Internacional da ESECS-Politécnico de Leiria.

8. Referências

Baker, P., Hardie, A., & McEnery, T. (2006). *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Berber-Sardinha, T. (2004). Lingüística de Corpus. São Paulo: Manole.

Biber, D., & Conrad, S. (2019). Register, genre, and style. Cambridge: Cambridge University Press.

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, *8*(4), 243–257.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*(1), 7–36.

Halliday, M. A. K. (1978a). Language as social semiotic. In *Language as social semiotic* (pp. 108-126). London: Arnold.

Halliday, M. A. K. (1991[2007]). The Notion of Context in Language Education. In J. J. Webster (Ed.), *Language and Education* (pp. 269-290). London/New York: Continuum.

Halliday, M. A. K. (2003). Introduction: on the "architecture" of human language. In J. J. Webster (Ed.), *On language and linguistics* (pp. 1-29). London/New York: Continuum.

Halliday, M. A. K., & Hasan, R. (1985). *Language, context, and text: aspects of language in a social-semiotic perspective* (second edition ed.). Hong Kong: Oxford University Press.

Hasan, R. (2013). Choice, system, realisation: Describing language as meaning potential (L. Fontaine, T. Bartlett, & G. O'Grady, Orgs.). Cambridge University Press.

Kuhn, Tanara Zingano; KOSEM, Iztok (2016). “Devising a Sketch Grammar for Academic Portuguese.” *Slovenščina 2.0: empirical, applied and interdisciplinary research*, v.4, p.124 - 161, 2016 DOI: http://dx.doi.org/10.4312/slo2.0.2016.1.124-161

Kuhn, Tanara Zingano (2017). *A Design proposal of an on-line corpus-driven dictionary of Portuguese for university students* (doctoral dissertation). University of Lisbon, Lisbon, Portugal.

Martins, M., & Mendes, M. (2021). *DOESTE: corpora para a investigação de escrita escolar*. Comunicação apresentada em Linguistweets 2 (Abralin), Twitter.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London/New York: Routledge.

Paquot, M., & Gries, S. T. (2020). A practical handbook of corpus linguistics. Cham: Springer.

Silva, P. N., Santos, J. V., & Sitoe, M. Z. (2019). Itinerários da escrita académica no ensino superior: um projeto de investigação aplicada sobre textos e géneros. In F. Caels, L. Barbeiro, J. V. Santos (Orgs.), *Discurso académico: uma área disciplinar em construção* (pp. 277-297). Coimbra/ Leiria: CELGA-ILTEC & ESECS-IPL.

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Berlim: Language Science Press.

Weisser, M. (2015). Practical corpus linguistics: An introduction to corpus-based language analysis. Hoboken: Wiley-Blackwell.

ANEXO I: Equipa dinamizadora

Mafalda Mendes (CELGA-ILTEC, Universidade de Coimbra, Portugal, mafaldamendes@uc.pt)

É membro colaborador do Centro de Estudos em Linguística Geral e Aplicada, CELGA-ILTEC, no grupo de trabalho Discurso e Práticas Discursivas Académicas, DPDA, e membro do Grupo de Estudos em Linguística Educacional (LEd). É licenciada em Línguas e Literaturas Modernas, Estudos Portugueses e Espanhóis e doutorada em Linguística Aplicada pela Faculdade de Letras da Universidade de Lisboa. Iniciou a carreira de investigação no Instituto de Linguística Teórica e Computacional, ILTEC, no desenvolvimento de gramáticas formais para o processamento morfossintático do português. Na Verbalis Computação e Linguagem, Lda., coordenou e desenvolveu trabalhos de lexicografia computacional baseada em dados de corpora, com destaque para o projeto de desenvolvimento do *Dicionário Prático de Português-Caboverdiano*. No seu trabalho de doutoramento investigou o desenvolvimento da escrita infantil, num estudo baseado em dados de um corpus de escrita de alunos do 1.º ciclo do ensino básico. Em parceria com Mário Martins (UFERSA, Brasil), colabora no corpus desenvolvimental DOESTE.

Mário Martins (UFERSA, Universidade Federal Rural do Semi-Árido, Brasil, mario.martins@ufersa.edu.br)

É professor de Linguística e Língua Portuguesa na Universidade Federal Rural do Semi-Árido (UFERSA, Brasil). Leciona em cursos de graduação e pós-graduação. É formado em Letras-Português pela Universidade Federal do Amapá (UNIFAP) e possui mestrado e doutorado em Linguística Aplicada à Educação pela Universidade de Lisboa (UL, Portugal). É líder do Grupo de Estudos em Linguística Educacional (LEd). Seus interesses de pesquisa estão relacionados ao desenvolvimento da linguagem em idade escolar; a abordagens funcionais para o ensino de gramática; a letramento linguístico; a análise e produção de materiais didáticos baseados em uso/corpus; e a ensino e aprendizagem de português como L1 e L2. É autor do corpus desenvolvimental DOESTE, parte do projeto de investigação Repositório de Escrita Escolar, em parceira com Mafalda Mendes (CELGA/ILTEC).

Marta Sitoe (Universidade Eduardo Mondlane, Moçambique, martasitoe@gmail.com)

É doutoranda em Linguística do Português na Universidade de Coimbra (Portugal); Mestre em Português como Língua Estrageira e Língua Segunda pela mesma universidade (2018); Licenciada em Ensino do Português (2012), pela Universidade Eduardo Mondlane.

É docente na Faculdade de Letras e Ciências Sociais da UEM, onde leciona gramática do português e escrita académica em português no âmbito das seguintes unidades curriculares: Português I e II; Português III e IV e Técnicas de Expressão e Escrita Académica. É investigadora iniciante filiada à Cátedra de Português como Língua Segunda e Língua Estrangeira e ao Centro de Linguística Geral e Aplicada da Universidade de Coimbra (CELGA-ILTEC), onde é membro colaborador. As suas áreas de pesquisa são Didática do Português e Literacia Académica.

Participou na criação do Corpus de Português Académico, inserido no Projeto Texto, géneros e conhecimento – para o mapeamento dos usos disciplinares da língua nos diferentes níveis de ensino – iniciado no CELGA-ILTEC em setembro de 2016.

Emília Marrengula (Universidade Eduardo Mondlane, Moçambique, emiliamarrengula@gmail.com)

É professora e investigadora na Universidade Eduardo Mondlane (Moçambique), licenciada em Ensino de Português pela Universidade Pedagógica de Moçambique, com um mestrado em Linguística Aplicada ao Português como Língua Estrangeira e Língua Segunda frequentado na Faculdade de Letras da Universidade de Lisboa, actualmente doutoranda na mesma área do mestrado e na mesma instituição, com o financiamento da Fundação Calouste Gulbenkian, tem monitorado o domínio da expressão escrita entre estudantes universitários moçambicanos, tarefa que lhe permite definir algumas propostas metodológicas para a melhoria da produção de géneros textuais específicos.

ANEXO II: Plano de trabalho

9:30	Receção dos participantes e apresentação. Apresentação da oficina - proposta de tarefa global. Procedimentos preparatórios (distribuição de documentação e registo no Sketch Engine).
10:00 Sessão 1	Teórico-prática Princípios teóricos e atividades
11:00 Intervalo	
11:15 Sessão 2	Teórico-prática Questões de exploração de corpus e atividades
13:15	Almoço
14:30 Sessão 3	Teórico-prática Desenho de corpus: compilação, anotação e atividades
15:30 Intervalo	
15:45 Sessão 4	
16:30	Discussão e planeamento da próxima oficina