Expressões multipalavras no discurso de fóruns sobre saúde mental: desenvolvimento de um *corpus* em inglês e português brasileiro

Gabriela Berndt[1], Sofia Serrano[1], Mariana Jin [1], Adriana Pagano[1]

[1] Universidade Federal de Minas Gerais
{gabiberndt1, sofias, mjin, apagano}@ufmg.br

Abstract. *Multiword expressions (MWEs), word sequences displaying some degree of idiosyncrasy, are central to both linguistic studies and Natural Language Processing (NLP). This paper reports on the creation of a novel corpus of MWEs in English and Brazilian Portuguese in posts from Reddit two Reddit threads about depression — a domain underrepresented in existing MWE resources, predominantly based on newspaper texts. The corpus was manually compiled, automatically parsed using UDPipe, revised in Arborator Grew and manually annotated on the FLAT platform following the PARSEME guidelines. The analysis combines frequency distribution by MWE category and cross-linguistic contrastive examination. The results highlight domain-specific patterns in MWEs, as well as constructions that function as conventionalized resources for expressing feelings, evaluations, and subjective states in digital mental health discourse.*

Resumo. *Expressões multipalavras (multiword expressions, MWEs), isto é, sequências de palavras que apresentam algum grau de idiossincrasia, são centrais tanto para os estudos linguísticos quanto para o Processamento de Linguagem Natural (PLN). Este artigo apresenta a criação de um novo corpus de MWEs em postagens em inglês e português brasileiro de comunidades do Reddit dedicadas à saúde mental — um domínio ainda sub-representado nos recursos existentes de MWEs, predominantemente baseados em textos jornalísticos. O corpus foi compilado manualmente, anotado automaticamente com informações morfossintáticas e sintáticas por meio do parser UDPipe, revisado no Arborator Grew e anotado manualmente na plataforma FLAT, seguindo as diretrizes do PARSEME. A análise combina a distribuição de frequência por categoria de MWE e uma investigação contrastiva interlinguística. Os resultados evidenciam especificidades das MWEs nesse domínio e construções que funcionam como recursos convencionalizados para expressar sentimentos, avaliações e estados subjetivos no discurso sobre saúde mental nas redes sociais.*

1. Introdução

As expressões multipalavras (MWEs) constituem um fenômeno linguístico de grande relevância para o processamento de linguagem natural (PLN), uma vez que representam unidades lexicais complexas cujo significado é não-composicional, isto é, não derivável do significado individual de seus constituintes [Firth 1957], [Ramisch 2015].Essas combinações apresentam idiossincrasias sintáticas e semânticas que dificultam sua

interpretação automática, impactando tarefas como análise de sentimentos, tradução automática e outras aplicações em PLN. No domínio da saúde mental, a identificação e interpretação dessas expressões é particularmente importante, pois MWEs são frequentemente utilizadas para expressar diferentes graus de intensidade afetiva, estratégias de *coping* e estados dissociativos. [Coll-Florit 2021], ao estudar metáforas, observou que em ambientes como fóruns online, essas construções – as quais são compostas por e/ou atuam como modificadores – funcionam como recursos convencionados de significação, frequentemente mediando a construção discursiva da experiência psicológica.

Apesar da centralidade das MWEs, persistem lacunas importantes na literatura: os recursos existentes para tratamento das MWEs em português brasileiro concentram-se predominantemente em textos jornalísticos, de modo que não contemplam textos produzidos em redes sociais, classificados como vinculados à atividade sócio-semiótica compartilhar (*sharing*), segundo a classificação de [Matthiessen 2010]. A mesma lacuna se observa para o inglês no domínio específico de fóruns digitais. Em ambas as línguas, ainda, textos relacionados ao domínio da saúde mental são pouco contemplados. Embora não trate de MWEs nem inclua dados em inglês, [Wilkens et al. 2026]analisam postagens sobre saúde mental em português brasileiro, mostrando que padrões lexicais, sintáticos e psicolinguísticos distintivos podem ser associados a perfis de usuários com depressão. Essa perspectiva está em sintonia com o presente estudo, que investiga as MWEs como recursos convencionalizados para expressar sentimentos, avaliações e estados subjetivos no discurso digital sobre saúde mental.

Diante do cenário descrito, este trabalho teve como objetivo principal a criação de um *dataset* composto por dois subcorpora, um em português brasileiro e outro em inglês, de textos extraídos de fóruns digitais do Reddit sobre saúde mental, anotado com categorias de MWEs de acordo com o manual de anotação do projeto PARSEME. Com base no dataset, analisamos a distribuição tipológica das MWEs e as diferenças de uso dessas expressões entre o inglês e o português brasileiro. Embora ainda seja de dimensão reduzida, o *dataset* permitiu aplicar e refinar um protocolo de anotação para os objetivos da pesquisa. O corpus poderá apoiar a avaliação e a análise de erros de sistemas de identificação de MWEs.

2. Referencial Teórico e Fundamentos da Anotação

2.1. Definição de MWEs

MWEs são construções gramaticais complexas passíveis de ser analisadas em mais de um lexema e caracterizadas pela não composicionalidade [Baldwin and Kim 2010], [Constant et al. 2021], [Ramisch and Villavicencio 2018]. O elemento comum entre as definições consolidadas na área [Calzolari et al. 2002], [**?**] é a ideia de que o significado do conjunto não é derivado da análise das partes e nem das regras morfossintáticas que as combinam. Trata-se, então, de uma questão de grau: as MWEs podem ser mais ou menos composicionais [Cordeiro et al. 2019], distribuídas em um *continuum* que vai de construções transparentes a expressões completamente opacas.

2.2. O Projeto PARSEME e a Taxonomia das MWEs

No campo do PLN, a necessidade de sistematização levou ao desenvolvimento de tipologias operacionais de MWEs. Nesse contexto, destaca-se o projeto PARSEME

[Savary et al. 2017, Ramisch et al. 2020], [Savary et al. 2023], [Scholivet et al. 2026], [Savary et al. 2026], que propõe uma taxonomia padronizada com categorias voltadas à anotação multilíngue de MWEs, detalhada por [Savary et al. 2023] e [Savary et al. 2017]. Essa tipologia tornou-se referência para análises comparativas e para a construção de *corpus* anotados em múltiplas línguas, sendo adotada integralmente neste trabalho. As denominações em português dessas categorias são traduções/adaptações nossas, com base nas diretrizes PARSEME 2.0 ([Savary et al. 2026] e na taxonomia discutida por [Savary et al. 2023], retomada no capítulo de [Ramisch et al. 2024]. Neste estudo, foram consideradas 15 categorias da taxonomia adotada: expressões idiomáticas verbais (*verbal idioms*, VID); construções verbo-partícula (*idiomatic verb-particle constructions*, IVPC); construções verbo-suporte (*light-verb constructions*, LVC); verbos inerentemente reflexivos (*inherently reflexive verbs*, IRV); construções multiverbais (*multi-verb constructions*, MVC); expressões verbais inerentemente adposicionais (*inherently adpositional verbs*, IAV); expressões idiomáticas nominais (*nominal idioms*, NID); expressões idiomáticas pronominais (*pronominal idioms*, PronID); expressões nominais deverbais (*deverbal nominal idioms*, NV) (Ramisch et al. 2026); expressões idiomáticas adjetivais (*adjectival idioms*, AdjID); expressões idiomáticas adverbiais (*adverbial idioms*, AdvID); expressões idiomáticas determinantes (*determiner idioms*, DetID); expressões idiomáticas adposicionais (*adposition idioms*, AdpID); expressões idiomáticas conjuntivas (*conjunction idioms*, ConjID); e expressões idiomáticas interjeitivas (*interjection idioms*, IntjID). As categorias efetivamente identificadas e suas frequências são apresentadas na Tabela 3..

Este estudo examina as MWEs em um *corpus* comparável bilíngue, articulando duas dimensões analíticas: a distribuição tipológica das expressões, com base na classificação do PARSEME, e seus padrões de uso em discurso digital sobre saúde mental. A partir dessa abordagem, examinam-se a distribuição tipológica das MWEs e seus padrões de funcionamento discursivo nos dois subcorpus, com atenção às convergências e divergências entre o inglês e o português brasileiro no discurso digital sobre saúde mental.

3. *Corpus* e Metodologia

3.1. Composição do *Corpus*

O *corpus* constitui-se de textos coletados na plataforma Reddit, escolhida por sua natureza multilíngue e pelo caráter espontâneo e informal das interações, que propicia um discurso indireto e contextualizado [Wilkens et al. 2026] e que favorece a emergência de MWEs em contextos autênticos de uso. O *subcorpus* em português foi extraído da comunidade r/perguntas, a partir do tópico *O que vocês fazem para lidar com a depressão*; o *subcorpus* em inglês foi formado a partir de respostas ao tópico *how would you describe depression to someone who's never experienced it?/como você descreveria a depressão para alguém que nunca passou por isso?*, da comunidade r/AskReddit. As perguntas não são equivalentes: a pergunta em português solicita práticas utilizadas para lidar com a depressão, enquanto a pergunta em inglês solicita uma descrição da experiência. Essa diferença pode influenciar a distribuição das categorias e impede que os contrastes sejam atribuídos exclusivamente às línguas. Nos dois casos, privilegiaram-se ambientes discursivos em que as pessoas descrevem, explicam ou elaboram suas vivências de sofrimento psíquico. Os dados foram coletados exclusivamente de fontes públicas, com total supressão de metadados de identificação, em conformidade com boas práticas de pesquisa ética em *corpus*. O pré-processamento incluiu normalização ortográfica, expansão de siglas, remoção de

metadados e *emojis* não interpretáveis e segmentação manual de comentários extensos em unidades sentenciais. Sempre que possível, foi preservada a segmentação espontânea realizada pelos próprios autores dos comentários. De forma que apenas em casos de comentários extensos e sem pontuação ortográfica clara realizou-se segmentação manual, tomando-se as orações como unidades de significado.

3.2. Caracterização do corpus

A Tabela 1 apresenta uma caracterização geral do *corpus* nas duas línguas analisadas-quanto ao número de sentenças e palavras e à razão entre esses parâmetros. Os dados apresentados são aqui expostos para explicitar o escopo do *corpus* assim como demonstrar que os *subcorpora* são equiparáveis, o que também é verdade para a Tabela 2.

Tabela 1. Caracterização do *corpus*

Parâmetro	PB	EN	Razão PB/EN
Sentenças	263	261	1,01
Palavras	4.253	4.166	1,02

3.3. *Pipeline* de Anotação

O *pipeline* de anotação abrangeu duas etapas sequenciais. Na primeira, os *corpora* foram anotados morfossintaticamente por meio do parser UDPipe [Straka et al. 2016], utilizando o *corpus* Portuguese-Porttinari, identificado na plataforma como portuguese-porttinari-ud-2.17-251125, treinado a partir do treebank Porttinari-Base-ud [Duran et al. 2023] e o modelo English-GUM, identificado como english-gum-ud-2.17-251125 treinado a partir do treebank english-gum-ud [Zeldes 2017], gerando arquivos CoNLL-U. Na segunda etapa, os arquivos foram importados para a plataforma FLAT [Van Gompel et al. 2024] e as MWEs identificadas e classificadas manualmente por três anotadoras especializadas, sentença por sentença, nas duas línguas. Os critérios de identificação de MWEs contemplaram os testes propostos pelo PARSEME. A adjudicação foi conduzida por uma especialista sênior, que resolveu os casos de discordância entre as anotadoras.

4. Resultados

A Tabela 2 apresenta o número de ocorrências de MWEs e de tipos lematizados distintos; por exemplo, *fazer terapia* e *faço terapia* constituem duas ocorrências de um mesmo tipo lematizado.

Tabela 2. MWEs encontradas

Parâmetro	PB	EN	Razão PB/EN
MWEs anotadas	122	81	1,53
MWEs por lema	97	71	1,37

4.1. Distribuição por Categoria PARSEME

A Tabela 3 apresenta as categorias do PARSEME com ocorrências em cada *subcorpus* analisado (português e inglês). São apresentados o número de ocorrências e frequência relativa de cada categoria no total de ocorrências. A Tabela apresenta somente as categorias efetivamente encontradas no corpus.

Tabela 3. Categorias das MWEs nos *subcorpus*

Categoria	PB (n)	PB (%)	EN (n)	EN (%)
AdjID	3	2,46%	0	0,00%
AdpID	5	4,10%	0	0,00%
AdvID	25	20,49%	12	14,81%
ConjID	7	5,74%	4	4,94%
DetID	2	1,64%	3	3,70%
IAV	9	7,38%	7	8,64%
IntjID	3	2,46%	1	1,23%
IRV	11	9,02%	0	0,00%
IVPC	0	0,00%	31	36,71%
LVC	22	18,03%	5	6,17%
MVC	1	0,82%	1	1,23%
NID	19	15,57%	5	6,33%
NV-LVC	0	0,00%	1	1,23%
PronID	0	0,00%	1	1,23%
VID	15	12,30%	10	12,35%
Total	**122**	**100,00%**	**81**	**100,00%**

Considerando as 15 categorias analisadas anotadas na Tabela 3, 12 categorias apresentaram ao menos uma ocorrência no *subcorpus* em português brasileiro e 12 no *subcorpus* em inglês, assim como explicitado na Tabela 3. A análise feita apontou as categorias verbais como as mais numerosas nas duas línguas (47,5% e 66,67% no português e inglês) seguidas da etiquetas adverbial/adjetiva (22,95% e 14,81%) o que está de acordo com o esperado na literatura [Wilkens et al. 2026]. Os perfis de distribuição de categorias, no entanto, são distintos entre as duas línguas e essa diferença constitui o achado central deste estudo.

Três grandes diferenças se destacam e são todas teoricamente motivadas. A primeira, e mais expressiva, diz respeito às construções verbo partícula (idiomatic verb-particle constructions, IVPC): no inglês, essa categoria representa 38,27% das MWEs, enquanto ela não aparece no português. Essa divergência reflete uma propriedade tipológica bem documentada: o inglês é uma língua de enquadramento por satélite (*satellite-framed*), que expressa trajetória e aspecto por meio de partículas verbais separáveis, como ”give up”, ”break down”e ”fall apart”. Já o português, como língua de enquadramento verbal, codifica esses conteúdos no próprio verbo ou em construções perifrásticas [Talmy 2000]. No *subcorpus* analisado, essa tendência aparece em expressões verbais inerentemente adposicionais como ”lidar com”, ”contar com”, bem como em construções verbo-suporte como ”fazer terapia”, ”tomar decisão”.

A segunda diferença diz respeito aos verbos inerentemente reflexivos (inherently

reflexive verbs, IRV): o português registra 11 ocorrências (9,02%), enquanto o inglês não apresenta nenhuma instância dessa categoria. Esse resultado é compatível com diferenças estruturais entre as línguas e com os critérios da categoria PARSEME, uma vez que o inglês não dispõe de um sistema de clíticos reflexivos gramaticalmente obrigatório comparável ao do português, em que construções como ”queixar-se”e ”sentir-se”o verbo é morfossintática- e semanticamente indissociável do pronome reflexivo ”se”.

A terceira diferença está relacionada às construções verbo-suporte (light-verb constructions, LVC). Ambas as línguas se mostram produtivas, mas o português apresenta frequência proporcionalmente superior (18,03% vs. 6,17%). Em ambas as línguas, foram mapeados os verbos suporte mais frequentes (”ter”, ”tomar”, ”fazer”, ”estar”, ”dar”, em português; e ”take”, ”make”, ”get”em inglês). A razão da maior incidência de LVC no *subcorpus* do português precisa ser mais bem investigada.

Além da distribuição tipológica, os dados revelam que determinadas MWEs funcionam como recursos convencionados para a expressão de experiências emocionais e psicológicas no discurso digital sobre depressão [Coll-Florit 2021]. As MWEs do *corpus* organizam-se em torno de três funções discursivas complementares. A mais frequente é a nomeação metafórica de estados depressivos por NID e VID — expressões como ”fundo do poço”e ”rest in peace”, que a comunidade emprega como convenções estáveis de significação, com opacidade semântica parcialmente neutralizada pelo contexto discursivo compartilhado. A essa se somam os AdvID na função de calibradores da experiência subjetiva (”às vezes”, ”out of nowhere”), sinalizando frequência, imprevisibilidade e cronicidade dos estados emocionais. Por fim, as LVC constituem um vocabulário fraseológico de *coping*: construções como ”fazer terapia”e ”seek help”nomeiam práticas terapêuticas e de enfrentamento recorrentes no discurso de compartilhamento em saúde mental.

5. Conclusão

Este trabalho apresentou um corpus-piloto anotado segundo o PARSEME 2.0 e descreveu a distribuição das MWEs em dois tópicos do Reddit. O recurso poderá apoiar a avaliação e a análise de erros de sistemas de identificação de MWEs. No domínio da saúde mental digital, de crescente relevância para o PLN clínico, a diferença é especialmente significativa, reforçando a necessidade de *corpus* anotados manualmente e de recursos específicos para o português [Ramisch et al. 2024], [Savary et al. 2023]. O *corpus* compilado é de tamanho moderado e restrito a dois tópicos do Reddit, sendo assim, a generalização dos dados obtidos para outros registros e plataformas exige cautela.

6. Agradecimentos

As autoras agradecem os pareceristas anônimos pelas sugestões para o aprimoramento deste trabalho. Gabriela Berndt é bolsista de mestrado da CAPES (88887.204699/2025-00). Sofia Serrano é bolsista de monitoria de graduação da Pró-Reitoria de Graduação da UFMG. Adriana Pagano é Bolsista de Produtividade em Pesquisa e recebe apoio do CNPq (404722/2024-5; 313103/2021-6) e da FAPEMIG. Este trabalho recebeu apoio da Ação COST CA21167 UniDive, financiada pela European Cooperation in Science and Technology. A pesquisa afilia-se ao Instituto Nacional de Ciência e Tecnologia (INCT) em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (TILDIAR), financiado pelo CNPq (408490/2024-1).

7. Limitações

Como já apontado anteriormente, o tamanho dos *corpus* deste trabalho não é suficiente para a sua utilização em *fine tuning* de grandes modelos de linguagem (LLMs). Quanto à análise dos dados, uma outra limitação é que não foi feita uma comparação com um *corpus* anotado de referência em outro domínio, em decorrência de restrições de tempo. Apesar disso, esse trabalho está incluso em projeto maior de pesquisa, ainda em andamento, o qual se debruçará em algumas lacunas além de expandir o *dataset* com um *subcorpus* em francês canadense.

8. Declaração Ética

Este estudo utiliza exclusivamente dados publicamente acessíveis na plataforma Reddit, em conformidade com seus Termos de Uso. Reconhecendo a sensibilidade do domínio de saúde mental, adotaram-se as seguintes salvaguardas: (i) anonimização integral, com remoção de nomes de usuário e identificadores de fóruns; (ii) disponibilização do *corpus* apenas em formato CoNLL-U, sem trechos suficientes para reconstituição dos posts originais; (iii) ausência de qualquer inferência sobre estados clínicos de usuários individuais. Desencoraja-se o uso do *corpus* para fins de triagem ou diagnóstico clínico.

9. Uso de Ferramentas de IA Generativa

As autoras declaram o uso de ferramentas de IA generativa para a organização e limpeza das tabelas sem a alteração desses dados ou *tags* atribuídas. Todo o conteúdo gerado foi revisado, validado e adaptado pelas autoras, que assumem responsabilidade pelo resultado final. Não foram utilizadas ferramentas de IA generativa na anotação do *corpus*, na análise dos dados ou na redação do trabalho.

Referências

Baldwin, T. and Kim, S. N. (2010). Multiword expressions. In Indurkha, N.andDamerau, F. J., editor, *Handbook of Natural Language Processing*. CRC Press.

Calzolari, N., Fillmore, C. J., Grishman, R., Ide, N., Lenci, A., MacLeod, C., and Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons. In González Rodríguez, M. and Suarez Araujo, C. P., editors, *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. uropean Language Resources Association (ELRA).

Coll-Florit, M.; Oliver, A. R. S. (2021). Metaphors of mental illness: a corpus-based approach analysing first-person accounts of patients and mental health professionals. *Cultura, Lenguaje y Representación*, 25:85–104.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., and Todirascu, A. (2021). Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Cordeiro, S., Villavicencio, A., Idiart, M., and Ramisch, C. (2019). Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

Duran, M., Lopes, L., das Graças Nunes, M., and Pardo, T. (2023). The dawn of the porttinari multigenre treebank: Introducing its journalistic portion. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 115–124, Porto Alegre, RS, Brasil. SBC.

Firth, J. R. (1957). *A Synopsis of Linguistic Theory*, pages 1–32. Blackwell, Oxford.

Matthiessen, C. M. I. M.; Teruya, K. L. M. (2010). *Key Terms in Systemic Functional Linguistics*. Continuum, London.

Ramisch, C. (2015). *Multiword Expressions Acquisition: A Generic and Open Framework*. springer.

Ramisch, C., Savary, A., Guillaume, B., Waszczuk, J., Candito, M., Vaidya, A., Barbu Mititelu, V., Bhatia, A., Iñurrieta, U., Giouli, V., Güngör, T., Jiang, M., Lichte, T., Liebeskind, C., Monti, J., Ramisch, R., Stymne, S., Walsh, A., and Xu, H. (2020). Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*.

Ramisch, C. and Villavicencio, A. (2018). Computational treatment of multiword expressions. In Mitkov, R., editor, *The Oxford Handbook of Computational Linguistics*. Oxford University Press, 2nd edition. `https://doi.org/10.1093/oxfordhb/9780199573691.013.56`.

Ramisch, R., Ramisch, C., and Villavicencio, A. (2024). Expressões multipalavras: fogo de palha ou osso duro de roer? In Caseli, H. and das Graças Volpe Nunes, M., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, chapter 5. BPLN, 3 edition. `https://brasileiraspln.com/livro-pln/3a-edicao/parte-palavras/cap-mwe/cap-mwe.html`.

Savary, A., Ben Khelil, C., Ramisch, C., Giouli, V., Barbu Mititelu, V., Hadj Mohamed, N., Krstev, C., Liebeskind, C., Xu, H., Stymne, S., Güngör, T., Pickard, T., Guillaume, B., Bejček, E., Bhatia, A., Candito, M., Gantar, P., Iñurrieta, U., Gatt, A., Kovalevskaite, J., Lichte, T., Ljubešić, N., Monti, J., Parra Escartín, C., Shamsfard, M., Stoyanova, I., Vincze, V., and Walsh, A. (2023). PARSEME corpus release 1.3. In Bhatia, A., Evang, K., Garcia, M., Giouli, V., Han, L., and Taslimipoor, S., editors, *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., and Doucet, A. (2017). The PARSEME shared task on automatic identification of verbal multiword expressions. In Markantonatou, S., Ramisch, C., Savary, A., and Vincze, V., editors, *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Savary, A., Scholivet, M., Ramisch, C., Nakamura, T., Bilinski, E., Stymne, S., Giouli, V., Markantonatou, S., Pais, V., Mitrofan, M., Estève, L., Guillaume, B., Mititelu, V. B., Čibej, J., Hernández, R. D., Fendel, V., Gantar, P., Kanishcheva, O., Krstev, C., Liebeskind, C., Lobzhanidze, I., Marković, A. M., Nešpore-Bērzkalne, G., Pagano, A. S., Shamsfard, M., Stankovic, R., Tajalli, V., Tiberius, C., and Padhye, A. (2026). Parseme

2.0 multilingual corpus of multiword expressions. In Piperidis, S., Bel, N., van den Heuvel, H., Ide, N., Krek, S., and Toral, A., editors, *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 4819–4834, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Scholivet, M., Savary, A., Ramisch, C., Bilinski, E., Nakamura, T., Mitrofan, M., and Pais, V. (2026). Edition 2.0 of the PARSEME shared task on multilingual identification and paraphrasing of multiword expressions. In Ojha, A. K., Mititelu, V. B., Constant, M., Stoyanova, I., Doğruöz, A. S., and Rademaker, A., editors, *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, pages 254–275, Rabat, Marocco. Association for Computational Linguistics.

Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Talmy, L. (2000). *Toward a Cognitive Semantics*. MIT Press,.

Van Gompel, M., Başar, E., Neumann, A., and van der Klis, M. (2024). Proycon/flat: V0.11.5.

Wilkens, R., Caseli, H., Neris, V., and Villavicencio, A. (2026). The visible and the latent linguistic clues of mental health in Brazilian Portuguese textual posts. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 2*, pages 135–147, Salvador, Brazil. Association for Computational Linguistics.

Zeldes, A. (2017). The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.