Human Interaction and Emerging Technologies (IHIET-AI), Vol. 201, 2026, 507–516
https://doi.org/10.54941/ahfe1007204

A Bilingual Study of Multi-Word Expressions in Journalistic Texts: Fine-Tune BERT With Head-Based Masking Technique

Christina Valavani and Stavros Giannakis

European Communication Institute, Athens, GR 15780, Greece

ABSTRACT

While Machine Translation (MT) systems integrate the advantages of Artificial and Human Intelligence, achieving "human parity" requires overcoming persistent challenges in modeling complex linguistic structures. This study explores the representation of financial Multi-Word Expressions (MWEs) in the German-Greek language pair to improve the accuracy of Neural Machine Translation (NMT) systems, addressing challenges in translating complex linguistic structures. Observed errors in financial terminology translation serve as the basis for refining the numerical representation process (vectorization). Particular emphasis is placed on the computational modeling of domain-specific and general language to address financial language issues and terminology. The study focuses on optimizing the numerical vectorization of multi-word terms to solve the "Distributed Semantic Problem" often observed in German separable verbs. By introducing a novel Head-Based Masking technique (HBMt), the study demonstrates a 56% improvement in semantic clustering compared to standard baselines. These findings confirm that enhanced vector handling of MWEs establishes a superior linguistic foundation, responding to a key challenge for the next generation, precision-oriented AI applications. The main goal is to resolve challenges associated with distributed semantic representation by compelling the model to treat distant components as a single semantic unit, producing more compact vector clusters for domain-specific terminology. The model is tested via an evaluation script using a curated test set (defined in evaluation.py) of 14 MWE pairs across four categories: Financial Causality, Functional Verbs, Separable Verbs, and Journalistic Phrasing.

Keywords: Word embedding, Neural machine translation, BERT, Double masking technique, AI

INTRODUCTION

The first part of the study establishes the theoretical framework and a classification of German and Greek multi-word expressions (MWEs), followed by a brief overview of common issues arising in their machine translation. Subsequently, an empirical analysis examines how the DeepL Translator handles various MWE categories in both languages, illustrated through three indicative translation examples. While no systematic analysis or categorization of machine translation errors is undertaken, in the final

Received January 19, 2026; Revised February 27, 2026; Accepted March 10, 2026; Available online April 1, 2026

section, these specific examples provide the basis for the subsequent steps in the numerical representation (vectorization) process. Specifically, these inaccuracies illustrate the need for the proposed double-masking embedding technique, with the aim of addressing challenges related to financial language and terminology through enhanced computational modeling.

MULTI-WORD EXPRESSIONS AND MACHINE TRANSLATION INACCURACIES

This study presents categories of multi-word expressions and analyzes machine translation inaccuracies related to their processing. Verbal multi-word expressions constitute a key category of MWEs. Specifically, German functional verbal phrases (*Funktionsverbgefüge*) and phrasal verbs are examined below, accompanied by examples of their German and Greek equivalents in financial terminology. An example of financial jargon, as processed by an automated translation system (DeepL-Translate) followed by the correct translation provided by a specialist translator in the third row, with corresponding glosses included in the fourth row, is presented below:

Table 1: "Σημειώνει πτώση" two-word phrase V+Adv. in German and V+ N in Greek.

LPKF Laser *notiert* **ebenfalls** *leichter* **(minus 2,14 Prozent).** *[LPKF Laser also falls (by 2.14%)]*
*Η LPKF Laser διαπραγματεύεται επίσης χαμηλότερα (πτώση 2,14%). [Η LPKF Laser σημειώνει επίσης πτώση (κατά 2,14%)
Η LPKF Laser *σημειώνει επίσης πτώση* (μείον 2,14%).
LPKF Laser is also *trading lower* (down 2.14 percent). [LPKF Laser also *falls* (by 2.14%)

In Table 1, the German phrase *[…] notiert […] leichter […]* is examined. The verb in this example could be rendered in Greek as *εμφανίζει πτώση* or *έχει πτώση*.

Another category of particular interest to this paper consists of particle (or phrasal) verbs such as *anheben*. The primary challenge regarding such verbs is that they are syntactically and morphologically separable in certain contexts, e.g., *Die DZ Bank hebt den fairen Wert für die Papiere der Lufthansa (Lufthansa Aktie) von 6 auf 10 Euro an [DZ Bank raises the fair value for Lufthansa securities (Lufthansa shares) from 6 to 10 euros]*. Table 2, presented below, provides an example of a particle (or phrasal) verb and compares its translation by both a human specialist and an automated machine translation system.

Table 2: Particle verb *anheben [raise]* in financial context.

FRANKFURT (dpa-AFX Analyser) - Die DZ Bank *hat* den *fairen Wert* für die Papiere der Lufthansa (Lufthansa Aktie) von 6 auf 10 Euro *angehoben*, die Einstufung aber auf “Verkaufen” belassen.
*[1]ΦΡΑΝΚΦΟΥΡΤΗ (dpa-AFX Analyzer) - Η DZ Bank αύξησε την *εύλογη αξία για το χαρτί της* Lufthansa (μετοχή της Lufthansa) από 6 σε 10 ευρώ, αλλά άφησε την *αξιολόγηση σε “πώληση”*.
ΦΡΑΝΚΦΟΥΡΤΗ (dpa-AFX Analyser) - Η DZ Bank *αύξησε την εύλογη αξία για τα* χρεόγραφα της Lufthansa (μετοχή της Lufthansa) από 6 σε 10 ευρώ, αλλά *διατήρησε* την *αξιολόγησή* τους στο επίπεδο “Πώληση”.
DZ Bank has raised the fair value for Lufthansa securities (Lufthansa share) from 6 to 10 euros, but maintained its rating at “Sell”.

Verbs like *gegenüberstehen* “(lit. *stand opposite, ‘there are’*), *schönreden* (lit. *flatter*, ‘*sugarcoat*’) or *totarbeiten* (lit.*dead work, ‘work to death’*) [2]” that may be considered as compounds since they involve lexical stems rather than prefixes, are in fact particle verbs because they are separable (Schlücker, 2019:81). Other categories include multi-word termini, mainly nominal or verbal phrases. These are subject-specific, but often lack a strict definition, e.g., draw a bill of exchange, draw up a will, *[einen Wechsel ziehen, ein Testament aufsetzen/* εκδίδω συναλλαγματική, συντάσσω διαθήκη*]* etc. (Stojić et al., 2018:363).

Additionally, adverbial collocations, typically comprising two adverbs, are frequently attested in German financial texts. The modifiers in these phrases express gradation, i.e. intensifying or diminishing the property denoted by the head adverb (Schlücker, 2019:81).

Table 3: Adverbial collocation *etwas stärker [a bit higher].*

Der TecDAX startete *etwas stärker*, rutschte dann allerdings tief in die Verlustzone ab.
* Ο TecDAX ξεκίνησε κάπως ισχυρότερα, αλλά στη συνέχεια διολίσθησε βαθιά στο κόκκινο.
Ο TecDAX ξεκίνησε *ανοδικά*, αλλά στη συνέχεια διολίσθησε βαθύτερα στη ζώνη των απωλειών.
The TecDAX opened a bit higher, but then slipped deeper in the red.

Regarding nominal MWEs, the N+N is a common structure of multi-word phrasing in both German and Greek (Schlücker, 2019:86) (Ralli, 2007).

Specifically, the structures observed in two-word phrases include: (a) AN [black market], (b) NN [shell company, lawyer politician, framework law, member state], and (c) NNgen [glasses case] (Ralli, 2007). In cases two nouns share the same case, they may be in a relation of dependency (e.g., framework law, shell company, giant corporation) or in a relation of parataxis [placed side by side] (e.g., lawyer politician) (Koliopoulou, 2012:7).

[1]The asterisk * denotes a linguistic error within the sentence, indicating that either the entire structure or a specific component is incorrect.
[2](Schlücker, 2019:81).

Table 4: Nominal Phrase N+PP *(von Dativ)* in German and N+Ngen in Greek.

Anteilscheine von Anlagefonds
*Μερίδια επενδυτικών κεφαλαίων
Μερίδια Αμοιβαίου Κεφαλαίου
Mutual Fund Shares

Building upon the machine translation issues identified above, this study proposes an embedding-based technique designed to contextualize lexical units and enrich them with additional semantic information.

The Proposed Approach

The study introduces a novel Head-Based Masking technique and a 4-Component Embedding Architecture ($E_{token}+P_{intra}+E_{phrase}+P_{inter}$) to improve the numerical representation of Multi-Word Expressions (MWEs) in German financial and journalistic texts. The objective is to resolve the "Distributed Semantic Problem" (as in separable verbs such as *brach... ein*) by compelling the model to treat distant components as a single semantic unit, thereby producing more compact vector clusters for domain-specific terminology. This approach introduced by Levy and Goldberg (2014) utilizes word embeddings following the parsing of each sentence (spaCy) (Honnibal & Montani, 2017). Similarly, this study parses sentences, with the distinctive technique of masking tokens based on their syntactic positions. Subsequent to sentence parsing (Levy et al., 2014:305), each word is trained on its respective phrase using an adaptable window size (implements Fixed-Window Padding: ensures every phrase block is normalized to 10 tokens) (dependency parsing) (Goldberg, 2017:129).

This approach aims to accurately predict both proximate word-context pairs within predefined phrasal dependencies and long-distance syntactic dependencies (Krimpas & Valavani, 2023) across entire sentences (Goldberg, 2017:129). Utilization of the dual-masking embedding technique further enables a focus on syntactic information; for instance, in the case of German separable verbs - where one part of the verb occupies the initial or second sentence position and the other is placed at the end - the first embedding step captures the verb in its entirety (the complete verb is embedded within its corresponding verbal phrase, including the end part of the verb). Training each word within its respective phrase (e.g., nominal phrase or verbal phrase) ensures that the vectors of those words are positioned more closely within the vector space. Multi-word phrases, whether verbal or nominal, are more tightly integrated when using the proposed dual masking embedding technique. This process is reiterated at the sentence level.

Custom Masking Strategy Based on Syntactic Relations

Early work on word representations, such as the Skip-gram model introduced by Mikolov et al. (2013), focused on the efficient learning of word embeddings from large-scale textual corpora. BERT employs a multi-layer bidirectional transformer encoder architecture, which is primarily designed to process each word within the context of all other constituent elements of a sentence, rather than one-by-one in order. BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) undergoes pre-training on two key tasks:

1. Masked Language Modeling (MLM): This approach entails concealing specific words within a given sentence and predicting them solely based on the context provided (Devlin et al., 2019) (Joshi et al., 2020). Variations in masking strategies have also been proposed by Liu et al. (2019).
2. Next Sentence Prediction (NSP): This process entails the prediction of whether one sentence logically follows another (Devlin et al., 2019) (Liu et al., 2019).

Sentences are tokenised and parsed into the constituent phrases (spaCy) (Honnibal & Montani, 2017). Each token is assigned a unique identifier (ID) from a specific vocabulary (V) (Manning et al., 2008). This vocabulary functions as a comprehensive mapping system, associating all possible tokens with distinct numerical values. The model determines the positions to be masked, such as the locations of all heads within each sentence, and then masks each token using its corresponding ID (Manning et al., 2008). Conversion to IDs (assuming a hypothetical vocabulary): [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,15, 16, 17, 18, 19, 20].

The implementation of the masking strategy for BERT entails two fundamental steps. First, phrase extraction is initiated, and special tokens (Liu et al., 2019) are introduced for phrase boundaries, such as "[PHRASE_START]" and "[PHRASE_END]" to clearly demarcate the start and end of a phrase.

The masking rules are applied during the data pre-processing stage of model training by modifying the training data generation scripts to incorporate custom masking logic. Within each phrase, lexical items are masked jointly with their respective syntactic governors, in accordance with dependency parsing principles. In each case, two words are masked concurrently (double masking) (Joshi et al., 2020). By masking each word together with its syntactically dependent component, the model enhances its ability to predict both the content and the structure of the phrases. This approach embeds words that exhibit long-distance dependencies (Krimpas & Valavani, 2023) while explicitly encoding their relationships, thereby enabling the model to capture and learn these dependencies more effectively.

Instead of arbitrary token selection, this approach employs syntactic double masking to ensure that only syntactically linked word pairs are masked concurrently, thereby encoding their morphosyntactic dependencies. For instance, in the German sentence *„Nach diesen starken Zuwächsen haben sich jedoch die Aktienkurse von Teladoc, Veeva und Medtronic wieder auf normales Niveau eingependelt."* each masked pair corresponds to a direct syntactic dependency, such as *nach–Zuwächsen*, *diesen–Zuwächsen*, *Aktienkurse–eingependelt*, *haben–eingependelt*, *wieder–eingependelt*, *auf–Niveau*, *normales–Niveau*, and so on. By way of illustration, when predicting the preposition *"nach"*, both *"nach"* and its syntactic complement *"Zuwächsen"* are masked, yielding the input: "[MASK] diesen starken [MASK]". This strategy is consistently applied across the sequence, generating a structured series of syntactically dependent masking pairs such as:

Table 5: Syntactic double masking technique.

[MASK] diesen starken [MASK] […] (nach - Zuwächsen)
Nach [MASK] starken [MASK] […] (diesen - Zuwächsen)
Nach diesen [MASK] [MASK] […] (starken - Zuwächsen)
Nach diesen starken Zuwächsen [MASK] [MASK] jedoch die Aktienkurse […] [MASK] (haben- sich - eingependelt)
[…] sich jedoch die [MASK] […] [MASK] (die Aktienkurse - eingependelt)
[…] sich jedoch die Aktienkurse […] [MASK] [MASK] (wieder - eingependelt)
[…] Aktienkurse [MASK] [MASK] Veeva und Medtronic […] (von - Teladoc)
[…] Aktienkurse [MASK] Teladoc [MASK] […] (von - Veeva)
[…] Aktienkurse [MASK] Teladoc, Veeva und [MASK] […] (von – Medtronic)
[…] eingependelt [MASK] normales [MASK] (auf - Niveau)
[…] eingependelt auf [MASK] [MASK] (normales - Niveau)

Additionally, entire phrases (Joshi et al., 2020) may be masked, with the model trained to predict the masked phrase based on the bilateral context provided by the preceding and succeeding phrases. For example, given the phrases "[Nach diesen starken Zuwächsen]", "[die Aktienkurse von Teladoc, Veeva und Medtronic]", and "[auf normales Niveau]", masking the intermediate phrase yields the input "[first phrase^{1position}] [MASK^{2position}] [third phrase^{3position}]". This process entails assigning a unique embedding to the phrase as a single unit, which can be achieved by:

- Assigning a dedicated positional embedding (Vaswani et al., 2017) to the entire phrase or employing a separate embedding layer that captures the semantic representation of the phrase as a whole.
- Using a separate neural network layer or a pooling strategy (Goodfellow, 2016), such as mean or max pooling, across the token embeddings within a phrase to generate a single vector representation of the entire phrase.

Unlike conventional word-level or arbitrary span masking, our approach restricts the masking operation only to those phrase pairs that exhibit direct syntactic dependencies, as identified by dependency parsing. This strategy ensures that the model learns from semantically and syntactically coherent units. Furthermore, the positional embedding of each token remains unaltered, thereby preserving sentence order and grammatical structure.

For each syntactically linked phrase pair, double masking is applied: both the dependent and its governing phrase are masked simultaneously. This approach enables the model to infer their internal semantic content and external syntactic relationships, thereby enhancing its understanding of multi-word expression structures and functional phrase composition. A representative example of syntactically dependent phrase pairs and double masking strategy for the sentence: *„Nach diesen starken Zuwächsen haben sich jedoch die Aktienkurse von Teladoc, Veeva und Medtronic wieder auf normales Niveau eingependelt*"is presented in Table 6:

Table 6: Syntactically dependent phrase pairs and double masking strategy.

Dependent Phrase	Head Phrase	Double Masked Format Example
„Nach diesen starken Zuwächsen“	„eingependelt“	[[MASK_PHRASE] haben sich … [MASK_ PHRASE]]
„die Aktienkurse“	„eingependelt“	[[MASK_PHRASE] haben sich … [MASK_ PHRASE]]
„von Teladoc, Veeva und Medtronic“	„die Aktienkurse“	[[MASK_PHRASE] [MASK_ PHRASE]] eingependelt
„auf normales Niveau“	„eingependelt“	eingependelt [[MASK_PHRASE] [MASK_ PHRASE]]

Each pair described above meets:

- A direct syntactic dependency (e.g., subject → verb, PP → verb, NP modifier → NP)
- Logical phrase coherence (multi-word units that form constituents)

This phrase-level masking approach maintains syntactic granularity while significantly enhancing the model's capacity to capture the semantic compositionality of syntactically bound phrase pairs, such as prepositional phrases, noun phrases, or clausal complements, within their dependency context. Joint masking of syntactically linked phrases leads to strengthened multi-word representations and a more refined comprehension of inter-phrase functional relationships.

Training

The training script automatically downloads the Europarl (DE-EL) dataset, applies filtering criteria for complex journalistic sentences, and trains the custom German encoder. To mitigate the risk of catastrophic forgetting, a two-phase training strategy is employed (Giannakis, 2024):

- Phase 1 (Epoch 1): The Base BERT model is FROZEN. Only the new Phrase Embeddings and CLS head are trained. This aligns the new architecture without destroying pre-trained knowledge.
- Phase 2 (Epoch 2+): The Base BERT is UNFROZEN. The entire model is fine-tuned using differential learning rates (Lower LR for BERT, Higher LR for Phrase Embeddings).

Technical Architecture Details: MWEProcessor.py

Robust Chunking: Identifies noun chunks via spaCy and treats remaining verbs/prepositions as single-token phrases, ensuring 100% sentence coverage.

Attention Masks: Generates masks to ensure BERT ignores the padding tokens used to enforce the fixed phrase window (10 tokens).

Dual Masking: Implements the core hypothesis by simultaneously masking the Governor and the Dependent (e.g., masking both “Inflation” and “erhöhen”).

RESULTS

The project modifies the standard BERT architecture to improve the processing of financial multi-word expressions (MWEs) in German. It introduces a head-based masking technique and a 4-component embedding summation – comprising token + intra-phrase + phrase + inter-phrase - to capture causal and dependency relationships between distant words in financial texts (Giannakis, 2024).

The objective is to address the “Distributed Semantic Problem” (e.g., separable verbs such as *brach... ein*) by forcing the model to treat distant lexical components as a single semantic unit. The evaluation script generates a PCA projection that illustrates how the custom model clusters semantic pairs more effectively than standard BERT. The left panel shows vector proximity under Standard BERT, whereas the right panel displays the proposed MWE-BERT.

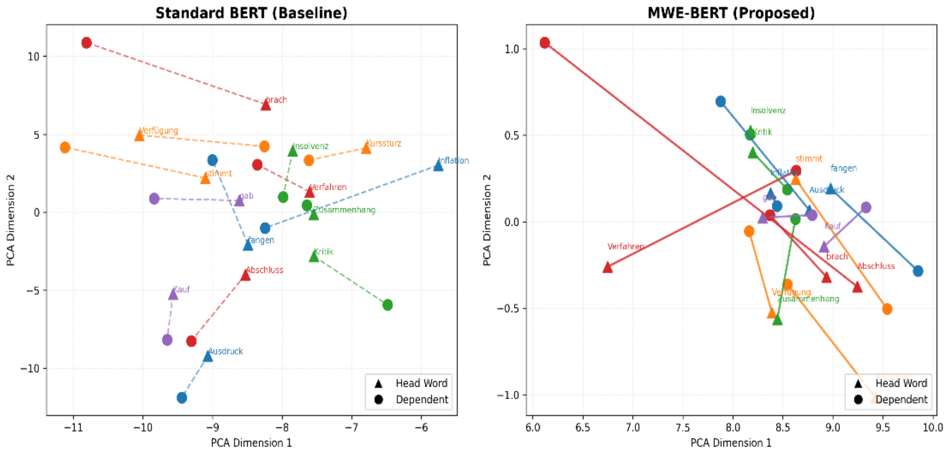

Figure 1: Vector proximity in Standard BERT vs. in the proposed MWE-BERT (Giannakis, 2024).

This script bridges the gap between linguistic theory (Dependency Parsing) and Neural Tensors.

Key Findings: The MWE-BERT model achieves a significant average improvement of +56.13% in cosine similarity compared to standard BERT.

Separable Verbs show massive gains due to the head-based masking technique, which connects distant words. Notably, *fangen - an* improved by +129.81% and *stimmt - ab* by +128.86%.

Financial Terminology also benefited greatly, with the fixed expression *Insolvenz - anmelden* (file for bankruptcy) improving by +51.62%, validating the model’s domain adaptation capabilities.

CONCLUSION

The first part of this paper examined multi-word expressions within financial language and terminology, and in light of the challenges observed in machine translation, it is concluded that they continue to pose a significant challenge, particularly in the case of domain-specific, multi-word expressions and financial terminology. In the second part, an enhanced contextualized vectorization technique is proposed, bringing complex NLP tasks closer to improved performance. The proposed approach aims to optimize the numerical representation of contextualized word embeddings in vector space, with the objective of not only improving machine translation systems but also benefiting natural language processing applications in general. Nevertheless, the automatic recognition and extraction of MWEs remains a persistent challenge for NLP applications.

REFERENCES

Cyprus Securities and Exchange Commission. (n.d.) Financial terms and interpretations. GetFile.aspx (cysec.gov.cy)

DeepL. (n.d.). DeepL Translator. https://www.deepl.com

Devlin, Jacob, Chang, Ming Wei, Lee, Kenton, Toutanova, Kristina. (2019). "BERT: Pre training of Deep Bidirectional Transformers for Language Understanding", in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

European Parliament. (n.d.). Europarl corpus (DE–EL). https://www.statmt.org/europarl/

Giannakis, Stavros. (2024): Bilingual MWE-BERT: Fine-tuning with Head-Based Masking. https://github.com/cgs23/mwe-bert

Goldberg, Yoav. (2017). *Neural Network Methods for Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Goldberg, Yoav. (2017). *Neural Network Methods for Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Goodfellow, Ian, Bengio, Yoshua, and Courville, Aaron. (2016). *Deep Learning*. MIT Press.

Goodfellow, Ian, Bengio, Yoshua, and Courville, Aaron. (2016). *Deep Learning*. MIT Press.

Google AI Blog. (2018, November 2) "Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing". Google AI Blog. https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

Honnibal, Matthew, Montani, Ines. (2017). "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing". spaCy 2 technical report. Available from spaCy documentation and community sources.

Joshi, Mandar, Chen, Danqi, Liu, Yinhan, Weld, Daniel S., Zettlemoyer, Luke, and Levy, Omer. (2020). "SpanBERT: Improving Pre-training by Representing and Predicting Spans". *Transactions of the Association for Computational Linguistics*.

Koliopoulou, Maria. (2012). Between compounds and phrases. The 10th International Conference of Greek Linguistics. Edited by Zoe Gavriilidou, Angeliki Efthymiou, Evangelia Thomadaki, Penelope Kambakis-Vougiouklis. Democritus University of Thrace, Komotini.

Krimpas, Panagiotis G., Valavani, Christina. (2023). "Attention mechanism and skip-gram embedded phrases: short and long-distance dependency n-grams for legal corpora". Comparative Legilinguistics, 52, 318–350. https://doi.org/10.14746/cl.52.2022.14

Levy, Omer, Goldberg, Yoav. (2014). "Dependency-Based Word Embedding", in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 302–308, Baltimore, Maryland. Association for Computational Linguistics. dependency-based-word-embeddings-acl-2014.pdf (wordpress.com).

Liu, Nelson F., Gardner, Matt, Belinkov, Yonatan, Peters, Matthew E., Smith, Noah A. (2019). "Linguistic Knowledge and Transferability of Contextual Representations", in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1112

Manning, Christopher D., Raghavan, Prabhakar, Schütze, Hinrich. (2008). Introduction to Information Retrieval. Cambridge University Press.

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S., Dean, Jeff. (2013a). "Efficient estimation of word representations in vector space". ICLR Workshop.

Ralli, Angeliki. (2007). Word composition. Cross-linguistic morphological approach. Athens: Ed. Patakis.

Schlücker, Barbara. (2019). Compounds and multi-word expressions in German. This work is licensed under the Creative Commons Attribution-Non Commercial-No Derivatives 4.0 License. De Gruyter. https://doi.org/10.1515/9783110632446-003

spaCy. (n.d.) Industrial-strength Natural Language Processing in Python. https://spacy.io

Stojić, A., Košuta, N. (2018). Zur Abgrenzung von Mehrwortverbindungen. Universität Rijeka.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Łukasz, Polosukhin, Illia. (2017). "Attention Is All You Need". arXiv preprint. https://doi.org/10.48550/arXiv.1706.03762