
Volume 38

—

Number 1

—

2025

—

ISSN 2190-6858



**Journal for
Language Technology and
Computational Linguistics**

Edited by Christian Wartena



**Gesellschaft
für Sprachtechnologie
und Computerlinguistik**

Imprint

Editor

Christian Wartena

Publication supported by the Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL)

Board of Directors

Committee and Advisory Board of the GSCL

Current issue

Volume 38 2025 Number 1

Address

Christian Wartena

Hochschule Hannover

Expo Plaza 12

D-30539 Hannover

info@jlcl.org

ISSN

2190-6858

Publication

Mostly 2 issues per annum

Publication only electronically on jlcl.org

License

Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

Contents

Editorial

Christian Wartena

iii

I Dissertation Abstracts

Measuring the Contributions of Vision and Text Modalities

Letitia Parcalabescu

1

II Research Articles

Can we Operationalize Conceptual Metaphor Cross-Lingually?

Maria Berger, Tatjana Scheffler

4

Enhancing ASR-based error analysis for atypical speech: post hoc integration of aphasic and dialectal phonetic features

Eugenia Rykova, Elisabeth Zeuner, Susanne Voigt-Zimmermann, Mathias Walther 37

With this issue, we present Volume 38, Number 1 of the Journal of Language Technology and Computational Linguistics (JLCL). The journal continues to provide a forum for research at the intersection of computational methods and linguistic inquiry. The contributions in this issue reflect several current directions in computational linguistics and natural language processing, addressing questions of multimodal understanding, meaning, and language variation.

The issue opens with a dissertation abstract by Letitia Parcalabescu on vision and language models. The work investigates how multimodal systems combine information from different sources and how their capabilities can be evaluated and interpreted. As such models become increasingly important in language technology, questions concerning their understanding and transparency remain central research topics.

The two research articles highlight the importance of combining computational approaches with linguistic analysis. Maria Berger and Tatjana Scheffler investigate the cross-lingual operationalization of conceptual metaphors, exploring how linguistic expressions can provide insights into underlying conceptual structures. Eugenia Rykova, Elisabeth Zeuner, Susanne Voigt-Zimmermann, and Mathias Walther address challenges in automatic speech recognition for atypical speech by incorporating knowledge about aphasic and dialectal phonetic features into an ASR-based analysis pipeline.

Together, these contributions illustrate the continued need for methods that account for the complexity and diversity of human language. They also demonstrate how linguistic insights can contribute to the development and evaluation of language technologies.

We thank the authors and reviewers for their contributions to this issue and hope that the articles will stimulate further research and discussion within the JLCL community.

Christian Wartena

Measuring the Contributions of Vision and Text Modalities in Multimodal Transformers

Understanding natural language requires machines to grasp more than just the surface-level form of text; they must also comprehend the underlying meaning, which involves knowledge of the world humans inhabit. While text provides substantial information, humans acquire much of their knowledge through other modalities like vision and sound. For machines to develop a better understanding of the world and the natural language that describes it, they need access to multiple modalities. A great starting point is the visual modality, given its relevance in human perception and its rich contribution to our understanding of the world.

This dissertation (Parcalabescu, 2024) explores vision and language models, which are multimodal systems that take vision and text modalities to produce outputs. Specifically, it develops computational tools to assess the effectiveness of vision and language models in combining, understanding, using, and explaining information from these two modalities. We structure our investigation into three key goals: (i) measuring specific and task-overarching capabilities of vision and language models, (ii) interpreting these models to quantify how much they leverage and integrate information from both modalities, and (iii) evaluating their ability to self-consistently explain their outputs to users.

In the first part of this dissertation, we introduce VALSE, a benchmark dataset designed to assess the visio-linguistic grounding capabilities of vision and language models across specific linguistic phenomena. This benchmark challenges models to differentiate between correct image captions and so-called foils – captions that contain subtle errors targeting specific linguistic phenomena grounded in vision: existence, plurality, counting, spatial relations, actions, and entity coreference. We propose four automated strategies to construct VALSE, ensuring the development of reliable and valid foils. Our evaluation of five widely-used vision and language encoder models and three decoder models (generating text from vision and language inputs) reveals that, while these models effectively identify objects and their presence in images, they generally struggle with more complex phenomena such as actions and spatial relations. This benchmark establishes a critical, ongoing challenge for modern vision and language models, aiming to track the progress of pretrained vision and language models from a *linguistic perspective*, complementing traditional task-centred vision and language evaluations in the field.

In the second part of the dissertation, we analyse how much vision and language models integrate and use information from both modalities. To quantify this integration, we introduce a multimodality score called MM-SHAP, which we designed to complement performance metrics, such as accuracy. This score is based on Shapley values, offering a performance-agnostic method to reliably determine the extent to which a multimodal model leverages individual modalities. With MM-SHAP, we assess different model

architectures for their overall degree of multimodality and evaluate the specific contributions of each modality within individual models on specific datasets and samples. Our findings challenge the belief that unimodal collapse – where a model predominantly relies on one modality – occurs uniformly in one direction. Instead, we observe that unimodal collapse can manifest in varying degrees and in different directions. Based on these insights, we recommend MM-SHAP for interpreting multimodal models and tasks, for diagnosing and guiding progress towards true multimodal integration.

In the third part of this dissertation, we explore whether vision and language models can give self-consistent explanations for their predictions. But the utility of these explanations hinges on their faithfulness, i.e., their accuracy in reflecting the model’s inner workings. Therefore, we need to test explanations for faithfulness. We clarify the status of existing faithfulness tests (developed almost solely for language-only models) in view of model explainability, characterising them as self-consistency tests instead. We compare all previous tests using the same models on the same datasets, and show that the predictions differ widely. We argue that the overall result at least questions the commonly-held view that these tests measure faithfulness, because they yield highly diverse predictions. While existing tests require input edits to test whether the model output changes, we propose CC-SHAP, an edit-free and interpretable measure, that analyses how model outputs relate to *how* the model processes the input. We compare CC-SHAP for 11 language models on 5 tasks against all other tests and show its advantages supported by individual examples for language only models. We find that chat language models show higher self-consistency than their base variants.

Finally, we extend CC-SHAP to vision and language models, and we are first to evaluate the self-consistency of vision and language models in both *post-hoc* and *chain-of-thought explanation* settings. We assess the self-consistency of 3 vision and language models on 11 datasets with CC-SHAP. We also apply the existing language-only self-consistency (faithfulness) tests in our multimodal setting. We find that vision and language models are less self-consistent than language-only models. Furthermore, the contributions of the image are significantly larger for explanation generation than for answer generation, and the difference is even more pronounced in chain-of-thought compared to the post-hoc explanation setting. This added complexity in the behaviour of vision and language models, as compared to their language-only counterparts, opens up new avenues for future research into the explainability of multimodal models.

We expect that the research contributions presented in this dissertation will continue to help measure the progress of vision and language research, and inspire future research on model benchmarking, interpretability and explainability.

References

Parcalabescu, L. (2024). *Measuring the contributions of vision and text modalities in multimodal transformers* (Doctoral dissertation). Universität Heidelberg.

Correspondence

Letitia Parcalabescu 

Universität Heidelberg
Institut für Computerlinguistik
Heidelberg, Germany
parcalabescu@cl.uni-heidelberg.de

Thesis Information

Doctoral thesis defended on July 19, 2024, at the Department of Computational Linguistics at Heidelberg University. Supervised by Prof. Dr. Anette Frank. Full text available at the Heidelberger Dokumentenserver: <https://archiv.ub.uni-heidelberg.de/volltextserver/35753/>.

Can we Operationalize Conceptual Metaphor Cross-Lingually?

Abstract

The conceptual nature of metaphorical expression is a long-discussed phenomenon, highly investigated by linguists, psychologists, translators, and philosophers, amongst others. In theoretical work, distinctions are made between conceptual metaphors (a phenomenon of human cognition) and linguistic metaphors (their concrete realizations in language), while most computational approaches have only addressed the latter. In the age of massive language models, metaphor and other phenomena of figurative speech are earning new attention as more and more textual analyses are built on top of neural-networking tools that do not necessarily make a distinction between the lexicalization of a concept and the concept itself. Hence, an investigation of conceptual metaphor using a more linguistics-driven perspective is of much importance. In this work, we investigate the conceptuality of metaphoric expressions across two languages utilizing a parallel corpus of news commentaries from the web. We assume that a conceptual metaphor is represented by many instances of linguistic metaphors. This idea presupposes linguistic metaphor as an operationalization of conceptual metaphor. We perform several tests on how metaphors are translated between the languages, to assess whether distinct lexicalizations of a metaphor form conceptual clusters, and whether the usage of words in a metaphorical context is distinguishable from their usage in literal contexts.

We find that we are able to group linguistic metaphors in one language into semantically related sets by clustering their translations in another language. We argue that these semantically related sets constitute an operationalization of conceptual metaphors. In English, the clusters are formed by fewer, but more diverse lexelts (linguistic types), while in German we find more and bigger clusters composed primarily of derivatives and compounds. We also find that when a lexelt is translated similarly in unannotated instances to known metaphoric usages, then its contextual sense tends to be figurative as well.

1 Introduction

Among experts, there is consensus that the phenomenon of metaphor requires that the interpreter of the figure creates a link to the metaphoric concept (overtly denoted by the expression of the figure). This phenomenon is not only considered a linguistic one (Lakoff & Johnson, 1980), it also is well-approved as a mental one (Borbely, 1998). The connection between metaphors and concepts actually goes back to (Ricoeur, 1978),

who sees language itself as a resource to regenerate ideas in new ways, and argues that metaphor is a driver of that process.

1.1 Conceptual Metaphor Theory

Following Lakoff and Johnson (1980), conceptual metaphor theory, at its core, understands metaphor as a matter of cognition. Metaphors (e.g., “LIFE IS A JOURNEY”) are processed by mapping “concrete” source domain concepts (e.g., journey) onto “abstract” target domain concepts (e.g., life). Zoltán Kövecses writes in Semino and Demjén (2016): “A conceptual metaphor is understanding one domain of experience (typically abstract) in terms of another (typically concrete).” In this definition, conceptual metaphor is, on the one hand, a cognitive process of understanding a domain, on the other hand, the concept is a product. In the linguistics literature, these two are not necessarily differentiated. In this article, we won’t distinguish them either as we work based on textual lexicalization of metaphors “in the wild”.

Experts in general agree on the opinion that metaphors play a role in the process of verbal conceptualization. In the field of cognitively influenced research on conceptual metaphor, many researchers also argue that conceptual metaphor resides in thought (Lakoff, 2009) and that metaphorical concepts can be built and enriched by linguistic instances (Semino & Demjén, 2016). Lakoff even claims that the human brain works effortlessly, precisely because it is processing language in a metaphorical way, hence, metaphor must be established at a cognitive level (cf. Holyoak & Stamenković, 2018). Lakoff and Johnson (1980) see metaphors not as a pure linguistic phenomenon, but instead argue that language is the means that exhibits metaphor’s work. In their view, metaphor is an embodied understanding of the world, relating the mind to the body, and the body to the world. The relationships in this scenario (e.g., space, orientation, movement) are driven by metaphors while the embodied concepts (e.g., container, path) are language independent (cf. Veale, Shutova, & Klebanov, 2016).

Borbely (1998) moves from a linguistic metaphor concept to a supra-linguistic one, a psychoanalytic concept of metaphor. Starting from Ricoeur (1978), who defines metaphor as “seeing something in terms of something else”, a psychoanalytic approach sees the present in terms of the past and the past in terms of the present. That understanding broadens the linguistic metaphor concept. Metaphor is not only a phenomenon of language, but of mentation. In fact, Borbely (1998) highlights a specific psychoanalytic concept of metaphorical process between primary (unconscious) and secondary (rational) thinking. Experiments showed that the mind functions metaphorically when faced with challenges requiring creativity.

Finally, in a comprehensive, empirical study, Holyoak and Stamenković (2018) elaborate three main theories in the research field of metaphor comprehension: analogy, categorization, and conceptual mapping. While analogy and categorization are simply represented by a model of information processing; conceptual mapping calls for an embodied understanding of metaphor comprehension. Holyoak and Stamenković wish for a greater consideration of a metaphors’ contextual function and its emotional

impact. They propose ways to combine analogy and conceptual mapping to broaden metaphor understanding. One way, they suggest, is considering conceptual metaphor based on individual representations of word pairs (source domain and target domain of a metaphor). Following Kintsch (2000), who says that metaphor comprehension is a real-time generation of a specific meaning adapted to the specific context in which a metaphor is used, a metaphor instance might also be understood as an analogical alignment process.

1.2 Computational Approaches to Conceptual Metaphor Theory

Work by Tsvetkov, Boytsov, Gershman, Nyberg, and Dyer (2014); Tsvetkov, Mukomel, and Gershman (2013) and Bizzoni and Ghanimifard (2018) shows that the conceptual metaphor idea is made use of in computationally-driven cross-lingual metaphor detection setups. Especially, Schuster and Markert (2023) build their work in cross-lingual metaphor detection on the idea that a concept remains stable across languages while its lexicalization does not necessarily do so.

Kintsch (2000) develops an approach to determine the semantic representation of a metaphor computationally. Kintsch argues that when individual instances of the source and target domains of a metaphor¹ are “merged”, we receive a set of paired “concepts” (corresponding to lexicalized words). These pairs could then be “conceptually combined” via a so-called metaphor vector (such as the procedure described in Kintsch, 2000). This again, creates a context-specific semantic representation of a metaphor. Kintsch (2000) further makes an attempt to distinguish target domains used in a figurative meaning from a literal one: A comparison between the metaphor vector “my lawyer is a shark” towards typical attributes of the source domain “shark” (e.g., “fish”) shows a strong distance between the metaphor vector and the “fish”-attribute. In contrast, the non-metaphor “my lawyer is young” shows a strong similarity between the phrase vector and the source-domain attribute (“young”). This shows that the “source domain” can have an impact to the contextual meaning of a (metaphoric) phrase.

Based on the understanding that the individual instances of conceptual metaphors are used in similar situations, they build similar contexts across languages. We can represent these contexts by using distributional semantics, a theory that has been pivotal in computational processing of language (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). We can adapt distributional embedding representations for multiple languages for example by using multi-lingual embedding models. Bilingual vector models might enable us to measure/compute these metaphoric contexts. They were successfully applied in NLP before (Luong, Pham, & Manning, 2015; Upadhyay, Faruqui, Dyer, & Roth, 2016; Vulić & Moens, 2015) and can help to investigate the phenomenon more deeply.

Related work often only considers the extracted linguistic metaphors in fixed POS patterns such as NV, VN, and AN (Köper & Schulte im Walde, 2017; Shutova, 2010;

¹An instance of a metaphor for Kintsch always contains a metaphor target and a metaphor source, e.g., “my lawyer (target) is a shark (source)”.

Tsvetkov et al., 2014, 2013) and translated independently of their context. Alternatively, metaphors are represented based on external knowledge (Bizzoni & Ghanimifard, 2018; Tsvetkov et al., 2014) while language-specific lexicalization of the metaphor is not accounted for on a precise level. The applied techniques also do not necessarily consider strongly divergent cultures between source and target languages.

In this research, we do not focus on automated metaphor detection in general, instead, we make use of the comparable lexicalization of conceptual metaphor across languages. We therefore do not review and present relating literature of automated metaphor recognition. The interested reader, however, may be referred to an extensive study on neural automated metaphor recognition (Choi et al., 2021), a procedure that also considers multi-lingual metaphor recognition (Sanchez-Bayona & Agerri, 2022), a contrastive learning approach (Lin, Ma, Yan, & Chen, 2021) that is based on the literal and figurative context of a given lexelt considering pre-trained neural models and a silver data augmentation procedure, and a verb metaphor detection procedure treated as a relation extraction problem (Song, Zhou, Fu, Liu, & Liu, 2021) in which the authors investigate a metaphoric verb in relation to its context.

1.3 Non-universal Conceptual Metaphors

Conceptual metaphor, not only its lexicalization, does not necessarily need to be universal. Concepts may be modified or lost, in particular across different cultures. As such, Kövecses (2018) writes that the universality of conceptual metaphors can vary from culture to culture, since they experience different contextual influences and are elaborated in different ways, which also can be influenced by the given context. Also, Yan, Noël, and Wolf (2010) investigate cross-lingual metaphor transfer asking how a “target language culture” affects the lexicalization of a metaphor and how that lexicalization, again, affects the concept of the metaphor to diverge from the source language. Yan et al. (2010) find that the way a metaphor concept is modified in a target language, relies strongly on the entrenchment rank. This means that a metaphor either remains preserved, or, when not shared in the target language’s system, it might be replaced by another metaphor of a higher entrenchment rank.

1.4 Our Contribution - Investigate Conceptual Metaphor via Translation

We approach the problem differently, from an empirical perspective, to find evidence for conceptual metaphor across languages. We are interested in how a metaphor’s lexicalization is translated. We build on the hypothesis that similarly translated linguistic metaphors share or at least somehow delimit a concept. We use the translations’ contextual representations and attempt to form clusters of words that share similar meanings, and therefore represent separate instances of the same conceptual metaphor. We present several experiments to evaluate translation embeddings and study whether translation representation can help to distinguish metaphorical and literal context for metaphor candidates.

We also build our work on the understanding that conceptual metaphor tends to remain rather preserved across languages, even though this is not always the case, while this does not necessarily apply to linguistic metaphor. Our overarching research question reads:

RQ: “How is conceptual metaphor represented by means of linguistic metaphor?”

As we understand conceptual metaphors as essentially universal (Kövecses, 2018; Lakoff & Johnson, 1980) and thus language independent, we can use data from different languages to investigate to what extent a conceptual metaphor is expressed using the same or different linguistic metaphors. In doing so, we gain new insights into how well contextualized representations support capturing conceptual metaphor.

In the following, we refer to a source language (SL) and a target language (TL) as is common in machine translation studies, and we refer to the metaphoric source domain (SD) and the metaphoric target domain (TD).

2 Background

2.1 Conceptual Metaphor in Translation Studies

The conceptuality of metaphor, in fact, is critical to analyzing metaphor across languages. Finding and automatizing ways to cross-language metaphor detection, hence, represents an important complement to studies in conceptual metaphor theory. As such, one of the major problems of current approaches is that cross-lingual metaphor detection, in languages culturally divergent, does not necessarily work based on simple dictionary translation. Investigating this phenomenon more deeply, Yan et al. (2010) investigate the linguistic perspective of metaphor transfer, i.e., the effect of the target language’s culture on the translated expression, and the conceptual perspective, i.e., the effect of the translation on the source language’s metaphoric concept. In their empirical categorization of FEAR metaphors, the authors make use of Vukovic (2007)’s assumption that a metaphor forms a general concept, which appears as specific lexical instances, e.g., FEAR IS AN OPPONENT vs. “conquered by fear”, “besieged by fear”. To retrieve metaphors in native/translated English and translated/native Chinese, the authors apply metaphorical patterns assuming that a metaphor forms a multi-word expression from a given SD towards more specific lexical items of a TD (Bisang, Hock, Winter, Stefanowitsch, & Gries, 2006). The authors find that whether a metaphor remains lexically stable or is replaced by a conceptually similar one, highly relies on its familiarity (entrenchment, Langacker, 1987) in the target language. Piccioni (2013) also finds some oddities in lexicalization during the translation of metaphors. The author compares the collocation variety of metaphors in Spanish original texts and English-to-Spanish translations and finds several unconventional metaphors in the translations, e.g., “see black” vs “see gray”. The author refers to that effect as “shining through”, cf. (Teich, 2003).

Piccioni (2013) performs a corpus study to derive a methodology for the analysis of translations via a comparison of metaphor use in original and translated texts. It uses a monolingual corpus of corporate sustainability reports consisting of (i) a sub-corpus of Spanish originals and (ii) a sub-corpus of translations from English to Spanish. The author analyses verb-noun metaphors to compare collocation variety, collocation regularity and metaphorical conventionality (depending on the appearance of a verb and a noun in its basic meaning or in its metaphoric usage) of the verb-noun pairs in original and translated texts. Results show that metaphors in translated texts show a tendency towards normalization (frequent metaphors make up a very high portion of all metaphors) as well as a preference for unconventional uses, which means that metaphors that do not necessarily appear in an original language text, do appear in a translation because words might be translated too literally. Very recently, Y. Liu et al. (2023) used conceptual features to help predict which language family a metaphor belongs to. As the conceptual features used are culture-independent, this approach successfully predicted the correct TL.

2.2 Representing Conceptual Metaphor by Embedding Vectors

Approaches to automated metaphor detection often utilize a form of vector representation to encode the meaning of words, individually or in context. Today, semantic information mainly is encoded by word embeddings (Mikolov et al., 2013). For example, Bizzoni and Ghanimifard (2018) and Maudslay, Pimentel, Cotterell, and Teufel (2020) present architectures of artificial neural networks that process the semantic representation of texts to recognize metaphors. Input representations might also be enriched with concreteness scores from relevant dictionaries and together passed into the network as features. However, these embeddings often come with a certain degree of bias (Papakyriakopoulos, Hegelich, Serrano, & Marco, 2020) and cannot necessarily distinguish metaphoric from literal use of lexical items. In earlier work, Kintsch (2000) says that when individual instances of metaphor are “merged”, we receive a set of paired “concepts” (corresponding to lexicalized words). These pairs could then be “conceptually combined” via a so-called metaphor vector to create a context-specific semantic metaphor representation.

3 A Data-driven Perspective on Conceptual Metaphor

It is tricky to study a cognitive notion such as conceptual metaphor directly on the basis of textual evidence. Indeed, a linguistic metaphorical expression in one text in one language does not give good evidence for an analysis of conceptual metaphor. However, we propose that parallel corpora may enable an operationalization for identifying conceptual metaphors by using cross-lingual translations of individual metaphoric expressions. Precisely, we imagine the process of metaphor translation as one of transferring the meaning of a metaphor represented by many events of lexicalization.

Our overarching research questing is: “How can conceptual metaphor be operationalized by means of linguistic metaphor?”

As shown in the literature, there is a tendency to view metaphor as something language-independent. However, from a lexicalized perspective, metaphors are not always translated literally, nor is the same linguistic metaphor necessarily used in the same context in all languages. This contributes to a range of challenges (variability, uncertainty, conceptual preservation) when metaphors are translated across languages. Descriptionist Van den Broeck (1981), for example, addresses this by introducing substitution and paraphrasing next to literal translation of a metaphor.

In the following, we list a more granular operationalization based on the different data types that helps us to address our main research question. We address metaphor variation across languages in a parallel corpus, consider their figurative and literal contexts and look at metaphorical compound nouns, which are very frequent in German.

3.1 Research Questions

When translating metaphor, we encounter two problems: (i) Translating the linguistic metaphor, and (ii) translating the conceptual metaphor (i.e., that also carries the metaphoric act; cf. Yan et al. (2010)). Following Schäffner (2004), translation studies experienced a strong shift from pure linguistics towards focusing on more contextual factors that can affect a translation.

For example, equivalence-based approaches say that a metaphor should be transferred intact from a SL to a TL, but different cultures can have a strong impact on such a transfer. For this purpose, Schäffner (2004) establishes conceptualization as relevant to translating an individual metaphor, hence, raises the problem of translatability from the lexical/linguistic level to the level of the cultural system of SL and TL.

Van den Broeck (1981) reflects on the boundaries of translatability exemplified by the translation of metaphor. Borrowing from (Even-Zohar, 1971), Van Den Broeck lists some rules that need to be considered when translating metaphor. One such rule says that “translatability is high when the general cultural evolution of SL and TL” is similar. Another rule correspondingly says that translatability is higher the fewer kinds of information need to be delivered. Concerning the first point, we shall be fine concerning the translatability of metaphor from English to German in the news commentary domain. The second point can be applied to the diversity of possible translations of a word in another language. We can infer that depth-2 entropy of a metaphor makes translatability harder.

In our dataset, it further often happens that a metaphor’s SD and TD (Lakoff & Johnson, 1980) from the English language text are translated together as the same unique compound word in the German language part of our corpus. Examples are:

1. gun(TD) control(TD) battle(SD) : Waffenkontrollkampf
2. gun(TD) war(SD) : Waffenkrieg
3. poverty(TD) card(SD) : Armutskarte

Van den Broeck (1981) also points to the challenges that accompany the translation of compounds. As metaphor as such typically violates contextual preference in the SL, such a violation in the TL’s system might also be expected. Van den Broeck (1981) argues this is not always the case as, for example, the condensed nature of the German compound metaphor “Tränenbäume” (tear trees) is translated in French as “arbres de larmes” (trees of tears), however, larmes has other translations (droplet, drop, water), hence it is “rendered more analytically” (Van den Broeck, 1981, p. 80). Even though, in our data, poetic jewellery is not predominant, we still are interested in how compounds affect translation across two languages and whether the distribution of the translations of compounds is similar to that of single words.

We refine our research to the following **Questions**:

- RQ1:** To what extent are linguistic metaphors preserved during translations (cf. Experiment 1)?
- RQ2:** How does linguistic metaphor volatility contribute to the same conceptual metaphor, i.e., can conceptual metaphor be encoded by contextualized representations of linguistic metaphor (cf. Experiment 2)?
- RQ3:** Can contextualized representations of conceptual metaphor help to distinguish metaphorical and literal contexts of linguistic metaphor candidates (cf. Exp. 3)?
- RQ4:** Does lexicalization of linguistic metaphor differ between German compounds and simplicia (cf. Experiment 4)?

Experiments and Data Overview: We measure how a specific metaphoric word is translated, how diverse the German translation set, cf., the different possible translations of the metaphor in the TL is, and measure how dominant one specific translation is in our corpus compared to others.

The data used in the study are metaphoric words from our metaphor corpus together with their German translations. To control the experiment, we also look into the translation distributions of the metaphoric word when it is not used in a metaphoric context, i.e., not annotated for metaphor using alignment data. The alignment data reveal information on how words are associated with each other across two languages. When a lexelt is a compound, the associated translation is a multi-word-unit (cf. Fig. 1).

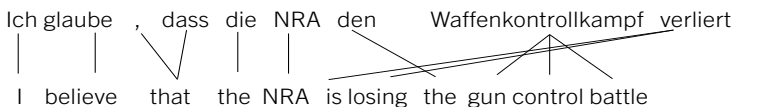


Figure 1: Example of bilingual alignment containing compound words

3.2 Metaphor Data

Our dataset of parallel metaphors initially originates from the English language corpus by Gordon et al. (2015)². Gordon et al. (2015) semi-automatically gathered metaphors pertaining to target concepts in the domains of governance, economic inequality, and the gun control-debate. The sentences collected for that purpose are from different genres of writing including press releases, news articles, weblog posts, online forum discussions, and social media. The authors justify their choice of sentences in that they “exhibit the diversity of ways that people can conceptualize the same set of target concepts”.³

Most metaphors in our data are of the form where “source + target” are directly syntactically related (e.g., “fighting poverty”, or “kills democracy”). They very frequently are formed by words that belong to the “fighting” domain, such as “combat hunger” and “assault of poverty”, but also the “securing” domain is represented often: “poverty alleviation” and “protecting democracy”

We use a machine-translated and manually evaluated subset of that corpus (Berger, 2022) for our study. It consists of 1000 English sentences. In every sentence, one metaphoric SD (one lexelt) and one TD (one lexelt) was annotated.⁴ For every lexelt that was used in a metaphoric context, we also retrieved contexts in which it has been used literally from the same corpus. This explains why the (literally used) types are a subset of the metaphorically used types. Tab. 1 gives an overview of the corpus statistics.⁵

Table 1: Corpus statistics

			(b) Metaphor tokens distributed by part-of-speech using spacy's POS-tagger (https://spacy.io/)		
(a) Words used metaphorical in the corpus					
contexts	EN	DE		EN	DE
types metaphorical	276	384	VERB	583	525
out of which literal	101	102	NOUN	318	376
tokens metaphorical	961	961	ADJ	34	24
tokens literal	209	396	ADV	0	21
			PROPN	16	9
			X	0	1
			tokens (total)	951	956

²Accessed: July 2025 from: https://figshare.com/articles/dataset/A_Corpus_of_Rich_Metaphor_Annotation/6179210?file=11188490 This effort makes use of the IARPA-funded Metaphor Program USC/ISI annotated metaphorical language collection, release iarpa_metaphor_isi.edu_metaphor_corpus_2015-04-03. <http://purl.org/net/metaphor-corpus>

³The exact references are not provided. The annotation procedure the authors are following originates from Mohler, Tomlinson, and Rink (2015).

⁴The entire parallel corpus is published in Berger, Kiwitt, and Reimann (2024).

⁵Note that token frequency differs between the label distribution and the corpus representation, because of translation quality, labeling procedure and multi-word expressions.

3.3 Contextualized Representations of the Data Used

We use contextual embedding representations from a pre-trained multilingual BERT model (Vaswani, 2017). This way, we retrieve a vector for each word that is used metaphorically in our corpus, and also for each of its contextual translations. Contextual embeddings can represent all words of the vocabulary of a language within one high-dimensional numerical space (Mikolov et al., 2013). In our work, these embeddings allocate the vocabulary of English and German words into the same vector space. We obtained the representation by aggregating the last four layers of a pre-trained multilingual BERT model hosted on huggingface.⁶

4 Metaphor Variation Across Languages

4.1 Entropy & Stability of Linguistic Metaphor (Exp. 1)

Y. Liu et al. (2023) propose a method to define the degree of conceptual similarity across languages. The authors create a bipartite alignment graph between the concepts of a SL and a set of TL tokens. Based on that graph, the authors align cross-lingual conceptuality in over 1,000 languages. Our study focuses on two languages only, and presumes the language agnostic conceptuality of a metaphor. Working based on a rather small dataset, we can also consider the translation procedure as a graph (bipartite or tripartite depending on the data format), and then retrieve translations in individual events. Inspired by Y. Liu et al. (2023), we define *stability* $S(M)$ of a linguistic metaphor. We consider translating a metaphor into one language and back as a tripartite graph $\mathcal{G}$ starting from one given lexelt l . Then, the number of paths of depth two—represented as a triple of nodes (y,x,z) —starting from that lexelt and ending at the same lexelt is divided by all other paths starting from that lexelt, and ending up in any lexelt of the SL.

$$S(M) = \frac{|\{(l, x, l) \in \mathcal{G}\}|}{|\{(l, x, y) \in \mathcal{G}\}|} \quad (1)$$

We measure depth-2 *entropy* $E_{depth-2}(M_l)$ of a linguistic metaphor by retrieving metaphoric words that still are contained in the translation set after one translation forward, and one backward pass (again, path depth two, cf. Y. Liu et al. (2023)). We calculate $E_{depth-2}(M_l)$ in the following way: We first calculate the entropy $E(l)$ of a given lexelt l in a SL by considering all ratios of all its translations (TL) in the total of possible translation as they appear in our data. We build the total $\sum$ over all these ratios (cf. Formula 3). Then, we calculate the entropy of all the translations of a lexelt l , namely $E(T(l))$. Here, we follow exactly the same procedure as above, simply starting with words in the TL. We build the average of these values $E_{av}(T(l))$. We balance $E(l)$ with .5 and the averaged entropy of all its translations $E_{av}(T(l))$ also with .5. This way, we obtain $E_{depth-2}(M_l)$. This measure especially helps us to gain insights into

⁶<https://huggingface.co/google-bert/bert-base-multilingual-cased>, acc. Nov. 2024

how diversely a lexelt and its translations are translated, hence how volatile metaphoric words are in our dataset.

$$E_{depth-2}(M_l) = \frac{E(l) + E_{av}(T(l))}{2}, l \in SL, T \in TL \quad (2)$$

While the entropy $E(x)$ of many events for x is defined as (Gray, 2011):

$$E(x) = - \sum_{x \in X} p(x) * \log p(x) \quad (3)$$

We present the stability-depth-2 *entropy* matrix in Figure 2. At first glance, the inverted correlation between depth-2 *entropy* and stability is especially visible for the German data rather than for the English data. Further, especially the ($y=1$) and ($y=.5$) lines show that there are many cases in which only one or two single translations exist for a given metaphor in both languages. We can also see that English and German metaphors are located a bit apart from each other in the matrix. German words are located more towards lower stability, while English words have higher stability. German intuitively has higher depth-2 *entropy* values due to the higher number of derivatives and compounds (cf. Sec. 4.2).

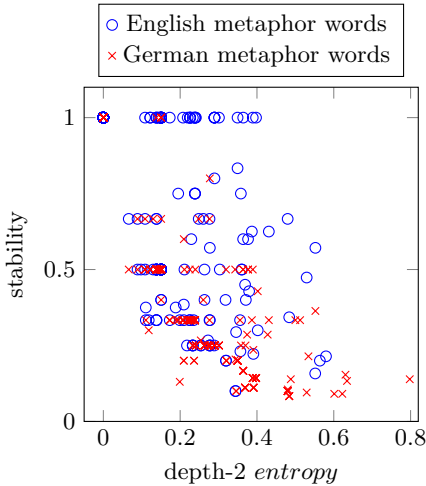


Figure 2: depth-2 *entropy* versus stability in English and German metaphoric words

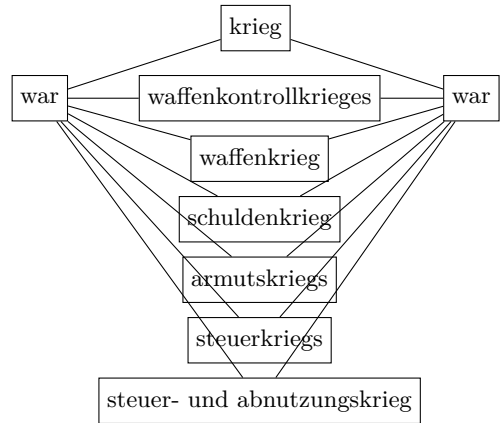


Figure 3: Translations graph of “war”

Figure 3 (and Fig. 8 to 10 in Appendix A) show samples of translation graphs of some of our English language linguistic metaphors. Often (about 25% of cases) an

English linguistic metaphor is alternatively translated by two different German verbs. This mainly happens because the distribution of words used in a text follows a power distribution (cf. Sec. 6.2).

Further, the English metaphoric words are often translated with noun compounds in German. Because we only consider metaphoric source words, German translations form a more diverse translation set, because the source and the target of an English language metaphor are combined into the same word in the German translation. This influences the stability and depth-2 *entropy* figures of English linguistic metaphors, because it leads to a high depth-2 *entropy* of words such as “war”, with a high stability at the same time (cf. Fig. 2).

4.2 Metaphor Distribution across Languages (Exp. 2)

Using the annotated English language metaphors and their contextual German translations, we investigate the translation distributions of lexelts in a language. The translation distribution of a lexelt is represented by a set of words in the TL. For example, the translation distribution of the English lexelt “oppose” is represented by the German lexelts “lehne (ab)” and “widersetzen”. For this set of words, we retrieve contextual embedding representations (CER) from a pre-trained multilingual BERT model⁷. When we average the vectors built by the contextual embedding representations V_{CER} of each word in the translation set into one single vector, we retrieve a translation vector TV for each word that is used metaphorically in our corpus:

$$TV_{oppose} = 0.5 * (V_{CER_{lehne}} + V_{CER_{widersetzen}}) \quad (4)$$

A clustering based on the 768 parameters of each vector of each lexelt can now help us to find lexelts that belong together in a—as we assume—conceptual sense. Throughout our experiments, when we talk about metaphor or metaphoric words, we only consider lexelts that were originally annotated as metaphor in the English language corpus, and their direct translations.

We use k-means clustering with a cluster size of about 10% of the total number of data-points resulting in 33 clusters for English and 38 clusters for German.⁸ We use initial centroid sampling based on the data points’ weights to the overall distribution, which is run one hundred times ($n_{init} = 100$). We cluster according to the 768 dimensions of the embedding representations. That is, because the 768 dimensions enable us to make use of a much more detailed parameter space, distinguishable and unifying characteristics are better usable. Tables 2 and 3 show all the clusters of metaphoric words. Clusters are often formed by inflections and compounds of the same word (for German). We also encounter a lot of synonymous and conceptual relations within the same cluster. We cannot determine the metaphor TD from the clusters, but, we can delimit the conceptual SD.

⁷<https://huggingface.co/google-bert/bert-base-multilingual-cased>, acc. Nov. 2024

⁸For English we can allow more clusters (.12) as vocabulary diversity is lower than in German.

Manual analysis shows that about 75% of the English and about 50% of the German language clusters represent moderately stable concepts in the sense that the vast majority of terms contained belong to an overarching topic. We “allow” this fuzziness, because we consider high contextual flexibility in today’s language usage. The other half however, contains terms that appear too random considering semantic and domain relatedness. Be aware that the purity of clusters strongly depends on their granularity. Hence, there always will be words that need to be assigned to clusters they “rather” belong to even if they do not match the concept very well. This trend especially becomes clear as smaller, more distinguishable conceptual clusters are much cleaner. For example, “Protect” is wrongly assigned to the “attack” cluster (cf. Tab. 2, cluster no. 11). An explanation can be that “Protect” is upper-cased, so that a very specific context, such as the beginning of a sentence or a heading, probably misguided the model (cf. Tables 2 and 3).

For visualizing purposes, we transform the 768 dimension down to a 2D plot, as shown in Figure 4a and 4b. Discarding the long tail of clusters, to have a clearer view at the data⁹, we encounter a representation that gives us some understanding of the arrangement of clusters in the space of contextual representations. The main purpose of visualizing the results of a clustering technique is to demonstrate that the properties (translations) used to describe an item (lexelt) are suitable to locate/arrange the semantics of that lexelt in relation to others. In our data, the plotted clusters show well-defined boundaries and the translations are thus suitable to describe the meaning of a lexelt.

We use principal component analysis (PCA Jolliffe, 2005) for dimensionality reduction, which aims to reduce dimensionality while retaining as much data variance as possible making it best suitable for our problem. In principle, PCA identifies the principal components in which the data varies the most.

Another common dimensionality reduction algorithm is multidimensional scaling. However, it is less suitable for our purpose. In contrast to PCA, MDS mainly focuses on preserving initial distances between each two data points. It also commonly is applied to problems situated in a lower-dimensional space. We also applied MDS to our clusters to visualize them in a 2D plot. Figure 5a and 5b show that while comparable clusterings are created, their visualization is less structured, hence less practical in our use-case.

4.3 Metaphoric Words in Literal Contexts (Exp. 3)

To operationalize an evaluation technique that helps us to judge whether and how clusters formed by linguistic metaphor can represent metaphorical concepts, we investigate linguistic metaphors (word types) used in both meanings, metaphorical and literal. We measure how both meanings’ translation distributions differ from each other.

To measure the conceptual divergence between words that are used in both contexts, metaphorical and literal, we consider all translations from each token, used in both contexts separately, and sum the embeddings representations of the translations for

⁹1-word clusters and “garbage” clusters; ca. 25% for German and ca. 33% for English

Table 2: Clusters of EN lexeltis; high saturation indicates clean clusters; gray indicates noisy clusters

no.	lexeltis
1	'circulation'
2	'flow', 'appetites', 'ceiling', 'competition', 'Race'
3	'appetite', 'parasite', 'cure', 'cured', 'remedy', 'decreased', 'bulwark', 'guardian', 'shield', 'addiction', 'inhibiting', 'corrupt', 'infection', 'evil', 'destructive', 'casualties', 'collateral damage', 'damage', 'killed', 'feeding', 'loss', 'disaster', 'disintegration', 'abyss', 'advanced', 'hole', 'hurdles', 'bondage', 'vehicle', 'burden', 'trajectory', 'sails', 'bludgeon', 'sword', 'challenged', 'competitive', 'conflicts', 'struggle', 'resist', 'defeated', 'tragedy', 'won', 'played', 'wrestle'
4	'feed', 'tastes', 'alleviate', 'relieve', 'addict', 'addicted', 'overdoses', 'escape', 'corrupts', 'cripples', 'damages', 'harms', 'hurt', 'hurts', 'weakens', 'eroding', 'harm', 'Hurt', 'impacts', 'fly out', 'fallen', 'encumbers', 'overcame', 'decreasing', 'trump', 'competes', 'fail', 'failed', 'gives up', 'Lose', 'give up', 'oppose', 'wins', 'occupy'
5	'feeds', 'decrease', 'inhibit', 'reduce', 'Reduce', 'impairs', 'shrink', 'compromises', 'impact', 'threaten'
6	'infests', 'protect', 'preserves', 'protected', 'Protecting', 'secure', 'Destroy', 'destroy', 'defeat', 'guard', 'protecting', 'protects', 'tackle', 'opposing'
7	'alleviates', 'relieves', 'restrains', 'handicaps', 'greedy', 'injuries', 'threatens', 'attacked', 'attacks', 'prevents', 'advances', 'hinders', 'burdened', 'burdens', 'trumps'
8	'cures', 'heals', 'harbors', 'retreated', 'travels', 'retreats'
9	'relief', 'Guardian', 'protection', 'alleviation', 'Alleviation', 'devastation', 'destruction', 'collapse', 'Poisoning', 'diagnosis', 'annihilation', 'degradation', 'defender', 'safeguarded', 'Surrender', 'surrender', 'advancement', 'progression', 'quest', 'chasm', 'retreat', 'liberation', 'tackling', 'revolution', 'combat', 'downfall', 'overthrow', 'overcome', 'overcoming', 'Tackling'
10	'pill', 'haven', 'addicts', 'fever', 'Kills', 'cliff', 'sojourn', 'Trends', 'block', 'obstructive', 'walls', 'door', 'route', 'baseball', 'card', 'conflict', 'crisis', 'debate', 'issues', 'wars', 'team', 'community', 'blitzkrieg'
11	'Protect', 'kill', 'attack', 'fight', 'Attacking', 'fighting', 'war'
12	'overdose'
13	'kills', 'reduced', 'reduces', 'destroyed', 'destroying', 'destroys', 'recovery', 'lost', 'win', 'loses', 'winning'
14	'assault', 'invasion', 'death', 'attacking', 'evolution', 'mountain', 'engine', 'ball', 'opponent', 'foes'
15	'dangerous', 'endanger', 'endangers', 'jeopardizes'
16	'erode', 'challenge', 'safeguard', 'enforce', 'advance', 'restrain', 'eradicate', 'play'
17	'killing', 'Besieges', 'Killing', 'fought', 'Kill', 'Fighting', 'Fights', 'combating', 'vie'
18	'threatened', 'overdosed', 'barrier', 'wall', 'compete', 'resists', 'battles'
19	'preserving', 'preserve'
20	'defend', 'Defend', 'defended', 'defending', 'opposes', 'defends'
21	'growth', 'progress'
22	'journey', 'path', 'pathway'
23	'pit'
24	'barriers'
25	'Chasm'
26	'recedes', 'decreases'
27	'trail', 'season'
28	'football', 'contest', 'race', 'races', 'victory', 'duel', 'battle', 'War', 'triumph'
29	'resisted', 'opposed', 'Opposing', 'faced'
30	'resisting'
31	'failure'
32	'scored'
33	'defeating'

Table 3: Clusters of DE lexels; high saturation indicates clean clusters; gray indicates noisy clusters

no. lexels	
1	'Kreislaufl'
2	'Fluss', 'Strom', 'Waffenfluss', 'Steuerfluss', 'Geldfluss', 'Einkommensflusses'
3	'Appetit', 'schmeckt', 'geheilt', 'Geldkuren', 'Wahlschutz', 'Überdosis', 'Infektion', 'gierig', 'böse', 'böser', 'gefährlich', 'Kills', 'ernährt', 'verletzt', 'angreift', 'anzugreifen', 'angreifen', 'Barrieren', 'Steuerlast', 'übertrumpfen', 'konkurrieren', 'greife', 'scheitern', 'scheiterte', 'versagt', 'Versagen', 'stellen', 'anzupacken'
4	'ernähren', 'verringert', 'verringern', 'verringerten', 'entlasten', 'reduzieren', 'reduziert', 'reduzierte', 'geschadet', 'zurückgegangen', 'nimmt', 'abnehmen', 'sinken'
5	'einfießt', 'befällt', 'lindert', 'heilt', 'mildert', 'entlastet', 'bewahrt', 'überdosiert', 'hindern', 'hemmend', 'zurückhält', 'korrumpiert', 'lähmt', 'schadet', 'gefährden', 'gefährdet', 'beeinträchtigt', 'trägt', 'untergräbt', 'Erhaltung', 'reist', 'belasteten', 'belastet', 'lastet', 'überwunden', 'geht', 'übertrumpft', 'konkurriert', 'nehmen'
6	'Steuerhunger', 'Umsatzsteuersucht', 'Lohnsteuergrenze', 'Steuerreit'
7	'Parasit', 'Heilmittel', 'Demokratieheilmittel', 'Mittel', 'Hort', 'Bollwerk', 'Hüter', 'Übel', 'Frontalangriff', 'Hindernis', 'Last'
8	'lindern', 'heilen', 'Waffenkontrollpille', 'Schutz', 'schützen', 'geschützt', 'sichern', 'Waffensucht', 'korrumpieren', 'vernichtet', 'zerstört', 'zerstören', 'schaden', 'gefährliche', 'auszuhöhlen', 'töten', 'getötet', 'Vergiftung', 'zunichte', 'besiegen', 'Zerstörer', 'schützt', 'verletzen', 'auswirkt', 'bedrohen', 'Hüterin', 'Wächter', 'Protect', 'beschützt', 'durchzusetzen', 'belasten', 'einzuschränken', 'voranzutreiben', 'Töten', 'Widerstands', 'beseitigen', 'Überwindung', 'überwinden'
9	'Doppelbesteuerungserleichterungen', 'Erleichterung', 'Steuerentlastung', 'Grundsteuererleichterungen', 'Linderung', 'Vermögenssteuererleichterungen', 'Steuererleichterungen', 'Doppelbesteuerungsentlastung'
10	'Waffenpille', 'Armutsipille', 'Armutsaufenthalt'
11	'Steuerparadies', 'Hafen', 'Waffenkontrollhafen'
12	'Entlastung', 'Grundsteuersucht', 'Armutsbekämpfung', 'bösen', 'Opfer', 'Tod', 'gesicherter', 'Abgrund', 'Waffenrechtsförderung', 'vorankommen', 'Waffenbesitzrends', 'Hürden', 'Zollmauer', 'Knechtschaft', 'Berg', 'vorangekommen', 'verloren', 'Win-Win-Antwort', 'Knüppel', 'wettbewerbsintensiver', 'Rennen', 'Wahlverlust', 'Konflikte', 'Kampf', 'Waffenproblemen', 'Krieg', 'Schuldenkrieg', 'Armutskriegs', 'Steuer- und Abzugskrieg', 'Zollkriegen', 'verliert', 'erzielt', 'Tragödie', 'gewinnen', 'gewinnt', 'gewonnen', 'hineinspielen', 'Pro-Gun-Teams', 'Schusswaffen-Community', 'verteidigt', 'Waffenkontrolle', 'ringen', 'Waffenkontroll-Blitzkriegs', 'konfrontiert', 'kämpfen'
13	'Guardian'
14	'Waffenschützer', 'Angriff', 'herauszufordern', 'verteidigen', 'Verteidigerin', 'verteidigt', 'Verteidigung', 'greifen', 'greift', 'fördert', 'Gegner', 'widersetzen', 'Ablehnung', 'widersetzten', 'widersetzt', 'lehne', 'Gegners', 'ablehnt', 'lehnen', 'widersetzt', 'stärken'
15	'Schild', 'Opfern', 'Billionen-Dollar-Regierungsberg', 'Vehikel', 'Motor', 'Segeln', 'Vermögensträger', 'Schutzschild', 'Regierungsschild', 'Schwertes'
16	'Waffenkontrollsucht', 'suchtig', 'Sucht', 'Wahlsucht', 'Ausgabensucht', 'Steuersucht', 'Waffenkontrollsucht', 'Waffenstichtige', 'vorgeschossen'
17	'entgangen', 'Ansturm', 'Zerstörung', 'Vernögensvernichtung', 'Verwüstung', 'Schaden', 'Invasion', 'Zusammenbruch', 'Verlust', 'Diagnose', 'Ende', 'Verschlechterung', 'Bürokratiezerfall', 'Entwicklung', 'Wachstum', 'Weiterentwicklung', 'Fortschritt', 'Rückzug', 'Befreiung', 'Förderung', 'Armutskonflikt', 'Revolution', 'Untergang', 'Sturz', 'Kapitulation', 'Übergabe'
18	'hindert', 'behindert', 'bedroht', 'droht', 'schwächt', 'schädigt', 'bedrohen', 'treibt', 'bringt'
19	'schränkt', 'belagert', 'Schützen', 'aufzugeben', 'verlieren'
20	'tötet', 'Kollateralschaden', 'MoneySkills', 'aufgibt', 'fliegen', 'aufgeben', 'Niederlage'
21	'Waffenkontrollfiebers', 'Grube', 'Wiederaufbau', 'Waffenkontrolltür', 'Ball', 'Waffenkontroll-Baseball', 'Waffenkontrollkarte', 'Fußball', 'Waffenkontrollrennen', 'Waffenkrise', 'Waffendebatte', 'Waffenkontrollkampf', 'Waffenrechtskampfes', 'Wahlrevolution', 'Waffenkontrollkrieges', 'Waffenkrieg', 'Anti-Waffen-Sieg', 'Waffenteam', 'Waffenrechten', 'Waffenkontrollteam', 'Waffenrechtsgemeinschaft', 'Waffenkontrollgemeinschaft', 'Waffenkontrollangriffen'
22	'Verringerung'
23	'zerstörerischsten', 'zerstörerische'
24	'gefährlichen', 'korruptes', 'korrupten', 'korrupte', 'Steuerkonflikt'
25	'drohten', 'Vermögenskatastrophe', 'Waffenkontrollklippe', 'Geldwachstum', 'Weg', 'Suche', 'Armutsuche', 'Vermögenskluft', 'Loch', 'Armutsloch', 'Geldloch', 'Straßensperre', 'Hochzollmauern', 'Belastung', 'Waffentür', 'Waffenroute', 'Mindeststeuerpfad', 'Wahlsur', 'Waffenhasser-Karte', 'Armutskarte', 'Steuerwettbewerb', 'Steuerkriegs'
26	'Vernichtung', 'abgenommen'
27	'beeinträchtigen', 'auswirken'
28	'bekämpfen', 'kämpfen', 'Bekämpfung', 'kämpft', 'bekämpft', 'Bekämpfen', 'griff', 'bestreiten', 'antreten', 'bestritten', 'bekämpfte', 'Kampfes', 'Ungleichheiten', 'Armutsbekämpfungsindex', 'besiegte', 'besiegt', 'gehalten', 'gekämpft', 'Schwierigkeiten', 'vorgeht'
29	'verteidigte', 'verteidigten'
30	'angriffen', 'angefochten', 'forderten', 'stellten', 'herausgefordert'
31	'zurückging', 'zog', 'zieht'
32	'voranschreiten'
33	'hinausgekommen', 'weiterentwickelt', 'verfallen'
34	'Fortschritte', 'voranschreitet'
35	'obstruktive'
36	'mithalten', 'bestehen'
37	'Wahlkampfes', 'Steuerrennens', 'Wahlrennen', 'Hauptrennen', 'Wahlsieg', 'Sieg', 'Präsidentschaftswahlduell', 'Wahlsaison', 'Regierungskampf', 'Wahlkampf'
38	'US-Waffenkontrollkampfes'

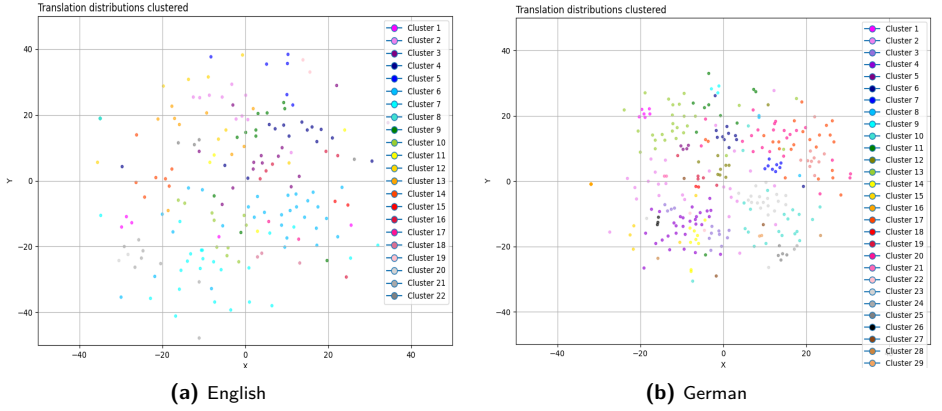


Figure 4: Translation distributions of metaphoric words using k-means after pca dimension reduction

both contexts. In this experiment, we only consider those words that have at least one literal as well as at least one metaphorical translation in our dataset. We apply several similarity statistics based on these two vectors. Appendix B Table 6 and 7 show the cosine similarity, the Z-test value and the significance level (α) based on the Kolmogorov–Smirnov test (KS-test) (Kolmogorov, 1933) under a given p-value.

The cosine value determines how similar both vectors are. That means a cosine value of 1.0 indicates that both contexts, literal and metaphorical, use the same translations. For matters of visualization, we color code higher cosine similarity values using higher color saturation. The Z-test value is a measure that is based on the mean error and the dispersion between two distributions, and, hence, measures whether two distribution are the same or different. A high absolute Z-value means that the distributions are different, a low absolute Z-value that they are similar. For matters of visualization, we color code higher Z-values using higher color saturation. Finally, the KS-test¹⁰ shows the statistical evidence (α) to which degree the null hypothesis can be rejected under the displayed p-value, which means that a certain variable, or distribution, does not follow a given distribution. If the p-value is smaller than the significance level, both distributions cannot be considered the same or similar, hence the null hypothesis can be rejected. In other words, the higher α the more likely it is that both contexts are different from each other. While a very high α correlates with very low p-values. We can see from the table that many metaphor distributions show a clear difference between literal and metaphorical translations at least in some of the similarity metrics.¹¹ In the German data, this evidence appears even more frequently.

¹⁰We use the Python implementation as delivered with the Scipy package.

¹¹The number of dimensions we work with in this experiment could be considered somewhat high (300). Hence, the significance of the KS-test might be impacted over-proportionally.

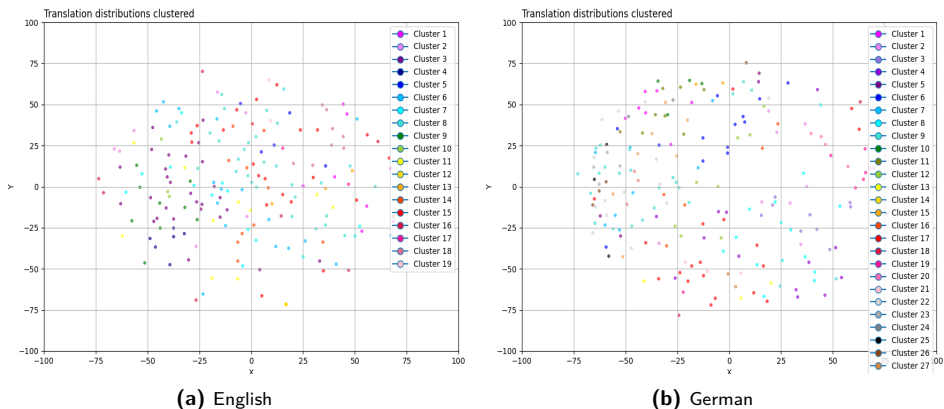


Figure 5: Translation distributions of metaphoric words using k-means after mds dimension reduction

Looking at sample evaluations, we find that, when a strong difference in the distribution of literal and figurative meaning is observed, often this is caused by the very rare use of the word in the “literal” context or meaning. We look at samples from English and German words according to Tab. 6 and 7. We distinguish their usage in a literal or metaphorical context and evaluate them according to very low and very high statistically estimated difference (visualized by a low versus high saturation of blue respectively). We count how many lexels and instances thereof from the, as we presume, literal usages are actually used in a literal sense. When the majority of all instances of a lexelt is used metaphorically and the p-value indicates strong correlation, cf. similar contextual usage, then the manual evaluation is scored +, otherwise -. In the case, the p-value indicated the opposite, we score the exact negation.

Strong statistical correlation indicates similar contexts:

“play”, -:

“play” was once annotated as metaphor and translated to the German “hineinspielen”. The unannotated appearances of “play” are aligned with “beitragen” (*contribute*) in the context “[...] dass höhere Strafen [...] dazu beitragen könnten [...] den Waffenfluss von Händlern zu Kriminellen einzudämmen” ([...] *that increased criminal penalties [...] could help play a role in stanching the flow of guns from traffickers to criminals*). However, in this context “beitragen” is used literally. Twice “play” translates to “spielen”: Once, in the context “[...] dass er einen großen Jackpot im Bingo gewinnen würde, da er es liebt, es zu spielen!” ([...] *that he would win big jackpot in bingo as he love to play it!*), also used in the literal context. Once in: “[...] eine führende Rolle [...] spielen” ([...] *could also play a leading role [...]*). In this context, “spielen” is again used metaphorically. For “play”, the statistics indicate comparable contexts between metaphorical and literal cases, but the evaluation shows different usage.

“revolution”, –:

“revolution” was used three times figuratively (‘revolution’ (1), ‘wahlrevolution’ (2)), and 4 times literally. Twice in “revolution by the working class”, and twice in the context of the colonists’ resistance against Britain, hence, all four times literally. Again, while statistics indicates similar contexts, manual evaluation show a different tendency in used contexts.

Weak statistical correlation - indicates different contexts:

“hole”, –:

“hole” is used four times metaphorically (translations: ‘loch’ (1), ‘armutsloch’ (2), ‘geldloch’ (1)), and three times “literally” (“hole in the social safety net”), which actually, also represents a metaphorical usage even if it forms a literal relationship to “net”.¹² Again, the distinguished contexts still turn out to belong to the same, metaphorical, sense.

“progress”, +:

“progress” twice translated to “Fortschritt” in the German version. Both appearances are used in a personification context (*“progress of their guns”*), hence it was annotated as metaphor. In the “literal” contexts, it was twice translated to “Fortschritte” out of which, once its meaning was literal: “[...] hat das FBI erhebliche Fortschritte [...] (*the FBI has made substantial progress*)”. It was further once translated to “in Bearbeitung” (“[...]die Anzahl der verabschiedeten Landesgesetze und der noch in Bearbeitung befindlichen Gesetzentwürfe [...]”), again used literally. For “progress” our statistical estimate predicts correctly that it is used in a different context, when it was not annotated for metaphor in the first place.

Strong statistical correlation indicates similar contexts:

“böse”, +:

“böse” appeared 11 times in the metaphorical context translated from “evil”. It was also used five times in the “literal” context, again translated from “evil”, always used in the metaphorical meaning (“evil gun pill”, “evil wealth”, “evil NRA” (2), “evil tyranny of concentrated wealth”). Statistic correlation here indicated similar context and manual evaluation proves it to be true.

“Weg”, –:

“Weg” translates in the metaphoric context twice to “journey”, three times to “path”, once to “pathway”, and once to “trajectory”. In the “literal” context it translates 8 times to “way”. In six of these contexts it is used literally within a lexical phrase (“to find a way to”, “the [best] way to ...” (5)); twice metaphorically (“have lost the way”, “along their way”). Once to “journey” (also literally, “[...] Jesus on his journey to the cross”), and once to “path” (used metaphorically, “what ’s down the path”). Again statics indicate similarity, but actual usage shows distinguishing contexts.

¹²According to collinsdictionary.com “safety net” describes (1) “a net used in a circus to catch high-wire and trapeze artistes if they fall”, or (2) “any means of protection from hardship or loss” like a social insurance can provide. Because there is a more basic meaning to “safety net”, “net” and “hole” are considered metaphorical in this sense according to Steen, Dorst, Herrmann, Kaal, and Krennmayr (2010).

Weak statistical correlation - indicates different contexts:

“Ende”, +:

“Ende” was once metaphorically annotated in the English version “death”. In the un-annotated (tendentiously literal) appearances “Ende” was five times translated from “end”, out of which four times it was used in a literal sense (“at the end of President Obama’s first term”, “election [...] to an end”, “defend my gun rights [...] to the end” (2)), and once metaphorically in a personification (“the end of [...] gun control efforts”). Here, again statistics indicate different usage context, and also the evaluation of the appearances in the corpus prove a tendency to different usage between literal and metaphorical senses.

“gewinnen”, –:

“gewinnen” was used three times figuratively and six times not annotated, hence assumed to be used in a literal context (translated: ‘win’ (3), ‘winning’ (3)). Evaluation showed all six appearances are also in a figurative context.

We can cautiously claim that positive correlation between the contextual embedding vectors of a lexelt’s translation and other metaphorical instances of a given lexelt can be a way to detect more metaphorical uses especially in German, even when they are not annotated. Positive correlation (low saturation of blue in Table 6) means that the word is translated with the same words in both contexts, which indicates that indeed the word is used in only one meaning.

Negative correlation can help to detect literal contexts of “metaphorical” words (i.e., contexts, in which a lexelt from the words annotated as metaphor is used literally) when translation variation is far off. However, translation variation does not ensure different contextual usages. This applies for example for compounds. We find that metaphorically used words translated with compounds in German, most probably, are metaphors. We present a comprehensive evaluation of computationally clearer cases in Tab. 4.

4.4 Translations of Compounds (Exp. 4)

Figure 6 shows the distribution in terms of frequency of translations of a word that is a simplicium or a compound in our dataset. Our dataset contains 41 compound-simplicium couples. Whereas 21 translations (tokens) are found for the simplicia, 37 translation tokens are found in total for the compounds. We can see that compounds and simplicia are translated slightly differently.

Table 5 shows that the cosine measures are almost always near 1.0. The German compound’s determinant usually represents the metaphor’s TD while the base word is the metaphor’s SD. Because we work with alignment data that associate a metaphor’s target, there is not necessarily a difference between the translation from German compounds/simple words to English pendants.

Figure 7 confirms that some of the compounds’ translations distributions differ from those of their simplex pendants, even though the vast majority shares the same translations between compound and simplicia (cf. Figure 7).

Table 4: Sample evaluation of English lexelts (top) and German lexelts (bottom); samples showing either significant difference or no difference in metaphorical vs. literal contexts according to p-value; manu(al) evaluation summarizes whether the actually usage of lexelts matches this prediction (Exp. 3)

lexelt	p-val	figurative use	literal use (actually literally)	manu.
p-value indicates strong correlation - similar contexts:				
burden	.73	5x"last", 1x"steuerlast" 1x"belasten", 1x"belastung"	1x"steuerlast" (0), 1x"entlastung" (0), 1x"belastung" (0)	+
feeding	1.0	1x"ernährt"	1x"stillen" (0)	+
hurdles	.34	1x"Hürden"	1x"Hürden" (0)	+
killed	.85	1x"getötet", 1x"töten" 1x"zunichte gemacht"	3x"getötet" (3), 1x "tötete" (1)	-
oppose	.97	4x"lehne(n) ab", 2x"widersetzen" 1x"gegen stellen"	3x"widersetzen" (2) 2x"entgegenzustellen" (0)	-
played	.96	2x"zum Narren gehalten" (.. rubes for fools)	3x"gespielt" (2), 1x"spielten" (0)	o
play	.88	1x"hineinspielen"	1x"beitragen" (1), 2x"spielen" (1)	-
protected	.48	1x"geschützte", 1x"schützt", 1x"schützen"	1x"geschütztes" (0) , 2x"geschützt" (0)	+
reduce	.73	4x"reduzieren", 4x"verringern"	1x"reduzieren" (0), 1x"verringern" (0)	+
revolution	.40	1x"revolution", 2x"wahlrevolution"	4x"revolution" (4)	-
p-value indicates no correlation - different contexts:				
block	.052	2x"Straßensperre"	1x"Barrieren" (0)	-
bulwark	.025	3x"Bollwerk"	5x"Bollwerk" (0)	-
defended	.001	3x"verteidigte", 3x"verteidigten"	1x"verteidigen" (0)	-
feeds	.001	1x"einfließt"	1x"ernährt" (0), 1x"nährt" (0)	-
hole	.013	1x"loch", 2x"armutsloch", 1x"geldloch"	3x"loch" (0)	-
protects	.000	6x"schützt"	2x"schützt" (0)	-
progress	.005	2x"Fortschritt"	2x"Fortschritte" (1) 1x"in Bearbeitung" (1)	+
race	.011	1x"waffenkontrollrennen", 1x"wahlkampfs" 1x"steuerrennens", 1x"hauptrennen" 1x"rennen"	1x"rasse" (1) 3x"rennen" (0)	-
win	.000	1x"Win-Win-Antwort" 2x"gewinnen", 1x"gewonnen"	3x"gewinnen" (0), Heller-Sieg (1)	+
winning	.113	1x"gewinnen"	3x"gewinnen" (0)	-
p-value indicates strong correlation - similar contexts:				
ablehnt	.811	1x"opposes"	1x"rejecting" (1)	-
böse	.730	11x"evil"	5x"evil" (0)	+
Belastung	.811	1x"burden"	1x"burden" (0)	+
erzielte	.848	1x"scored"	1x"earned" (1)	-
Hürden	.984	1x"hurdles"	1x"hurdles" (0)	+
konfrontiert	.771	1x"faced"	1x"faced" (0)	+
konkurrieren	.882	2x"compete"	3x"compete" (0)	+
Opfern	.848	1x"casualties"	1x"victims" (0)	+
Revolution	.730	1x"revolution"	4x"revolution" (4)	-
Weg	.811	2x"journey", 3x"path", 1x"pathway" 1x"trajectory"	8x"way" (6), 1x"journey" (1) 1x"path" (0)	-
p-value indicates no correlation - different contexts:				
bringt	.000	1x"advances" (.. voran)	1x"expresses" (.. zum Ausdruck) (0) 2x"puts further" (.. voran) (0) 2x"(doesn't) do (much)" (2) 1x"was (a quiet one)" (0) 1x"attract" (0)	-
Ende	.000	1x"death"	5x"end" (4)	+
Entwicklung	.000	1x"evolution"	2x"Development" (2), 4x"development" (4) 3x"developing" (3)	+
Gegner	.001	6x"opponent", 1x"foes"	1x"opponent" (1)	+
gewinnen	.001	1x"winning", 2x"win"	3x"winning" (0), 3x"win" (0)	-
gewonnen	.001	1x"won", 1x"(did) win"	2x"won" (0)	-
kämpfen	.000	6x"fight", 4x"fighting", 1x"vie"	2x"fight" (0), 1x"campaigning" (0) 1x"struggled" (0)	-
Sucht	.000	4x"addiction"	1x"addiction" (0)	-
töten	.000	9x"kill", 1x"killed", 1x"kills"	1x"kill" (1)	+
Tod	.000	4x"death"	1x"deaths" (1), 1x"death" (1)	+

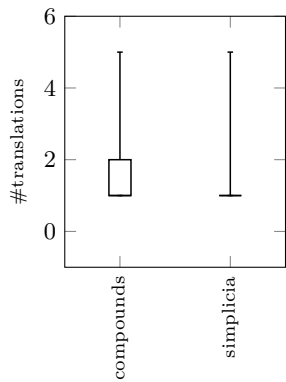


Figure 6: DE compounds' and simplex' translation distr. (En) accord. to vocabulary diversity (Exp. 4)

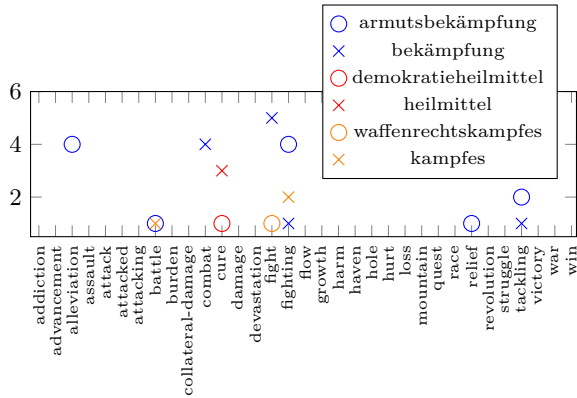


Figure 7: Distribution of samples of German compounds/simplicia with respect to their English translations and number of occurrence (Exp. 4)

5 Discussion

In Semino and Demjén (2016)’s overview of features of conceptual metaphor theory, Kövecses (2016) lists the following characteristics of conceptual metaphor that we discuss here with respect to how our experiments utilize them.

1. Concrete conceptual source domain & abstract conceptual target domain: A conceptual metaphor is a systematic set of correspondences between two domains of experience; the mapping helps to resolve a metaphor, in other words, the source domain conceptualizes the target domain, as in “fire vs anger”. Our metaphor data stem from contemporary ordinary news texts and interviews that contain terms, for example, from the fighting domain, such as “assault”, “attack”, “battle”, “fight”, “hurt”. We do not analyze concreteness here, but (J. Liu, O’Hara, Rubin, Draelos, & Rudin, 2020) show a clear correlation between stability and abstractness across many languages. Our experiments on **stability and depth-2 entropy** give us a numerical understanding of how stable translations of metaphor are, and which words rather tend to be translated back to the same source word, and whether stable or volatile words are more frequent in our corpus. We find evidence that depth-2 *entropy* plays a role in distinguishing literal and figurative context of a metaphor. Further, metaphors are very stable when they are translated with German compounds. However, our metaphors often use political vocabulary, which is not especially concrete or abstract, such as “protect”, “war”, “decrease”, “secure”, “struggle”.

2. Metaphor resides in thought: Kövecses (2016) writes, “When we conceptualize an intangible domain as a more tangible one, we can view a domain from different

Table 5: Corpus-specific cosine similarity between German compounds' and simplicia' translations in English according to vocab-diversity (Exp. 4)

compound-simplicia	cosine similartiy
waffenfluss-fluss	1.0
steuerfluss-fluss	1.0
geldfluss-fluss	1.0
demokratieheilmittel-heilmittel	1.0
steuerentlastung-entlastung	1.0
wahlsucht-sucht	1.0
ausgabensucht-sucht	1.0
waffensucht-sucht	1.0
grundsteuersucht-sucht	1.0
umsatzsteuersucht-sucht	1.0
steuersucht-sucht	1.0
armutsbekämpfung-bekämpfung	0.148
doppelbesteuerungsentlastung-entlastung	1.0
kollateralschaden-schaden	0.0
frontalangriff-angriff	0.120
waffenkontrollhafen-hafen	1.0
geldwachstum-wachstum	1.0
waffenrechtsförderung-förderung	1.0
armutssuche-suche	1.0
armutsloch-loch	1.0
geldloch-loch	1.0
billionen-dollar-regierungsberg-berg	1.0
steuerlast-last	1.0
waffenkontrollrennen-rennen	1.0
wahlverlust-verlust	1.0
hauptrennen-rennen	1.0
wahlsieg-sieg	0.914
waffenkontrollkampf-kampf	0.729
waffenrechtskampfes-kampfes	0.0
regierungskampf-kampf	0.724
us-waffenkontrollkampfes-kampf	0.724
wahlrevolution-revolution	1.0
waffenkrieg-krieg	1.0
schuldenkrieg-krieg	1.0
steuerkriegs-krieg	1.0
steuer-und-abnutzungskrieg-krieg	1.0
waffenkontrollangriffen-angriffen	0.0
steuersieg-sieg	0.914

perspectives”, in metaphorical words, different sources. It is even possible to create a non existing TD by adding more sources. **Bilingual embeddings** enable us to form a cognitive space that is supposed to map our conceptual perception represented by massive lexicalization. In our experiments we make use of these representations.

3. Conceptual metaphors are grounded: “Why are certain sources paired with certain targets?” Kövecses (2016) asks. Following the author, often, their ob-

jective/subjective similarity, such as “life” = “gambling game”, “human life cycle” = “plant life cycle”, lies in certain peculiarities the TD shares with the source. In static word embedding models, the representation of a word in the embedding space does not differentiate figurative and literal meaning nor do embeddings encode whether a word is supposed to be a metaphor’s SD or TD. The model represents those contexts of a word that are typical for a given domain. Say, the context of “fire” from a political news corpus would find it most often to be used metaphorically only, hence the representation reflects that. Our experiments on **literal & figurative contexts of metaphors** make use of contextual embeddings representing a word’s translation set in both meanings across two languages to address ambiguity. Distribution significance measures further can help to differentiate contexts.

4. Provenance of sources: Provenance of sources means that metaphor sources come from comparable areas and physical domains such as body parts, or other concrete objects. They are often supposed to be culturally and language independent, and instead are driven by the context (Semino & Demjén, 2016). We encode our linguistic metaphor sources by word vectors from contextualized representations. Hence, we derive a language-agnostic representation of them that can be used in several experimental setups, such as **concept clustering** (Exp. 1) and **literal versus figurative context deviation** (Exp. 2).

6 Discussion of Limitations

While our study presents an empirical approach towards linking linguistic and conceptual metaphor, we acknowledge the following potential limitations of our methodology.

6.1 Human versus Automated Translation

Some may raise concerns that our machine-translated corpus is not well suited to investigate behavior in metaphor translation. We do not think this is critical. First of all, automated approaches are frequently applied in retrieving and analyzing metaphor cross-lingually. The reader may be referred to related work, such as Maudslay et al. (2020); Shutova (2010); Tsvetkov et al. (2014), see also literature in Sec.1.2. Further, neural machine translation in many domains approaches human translation quality (Wu et al., 2016), and translation quality in particular does not necessarily affect the metaphoric expressions. In addition, we ensured the quality of the translation of our parallel corpus by manual evaluation (Berger, 2022).

6.2 Power-law and Asymmetry in the Corpus

Natural language text follows a power law distribution (Zipf, 2016). Our experiments on stability, depth-2 *entropy*, as well as the translation distributions in the TL that our methodology builds upon, are affected by such distributions. The natural power-law distribution of words in a textual context in particular caused that we often see many

data points near the zero, one or .5 y-axis (Exp. 1) or that we find a few bigger, a moderate number of medium-sized, and many quite small clusters (Exp. 2). However, we do not want to remove these many data points as they show empirical facts specific to the data sample we use.

On the other hand, as we do work with a one-to-one translation of one single corpus, we encounter a form of asymmetry in the metaphors detected, which might have an effect on the translation distribution of a given lexelt in our dataset.

6.3 Non-universal Conceptual Metaphors

Conceptual metaphor does not necessarily need to be universal. Concepts may be modified or lost, in particular across cultural differences. As such, Kövecses (2018) writes that the universality of conceptual metaphors can vary from culture to culture, since they experience different contextual influences and are elaborated in different ways, which also can be influenced by the given context. Also, Yan et al. (2010) investigate cross-lingual metaphor transfer asking how a “TL culture” affects the lexicalization of a metaphor and how that lexicalization, again, affects the concept of the metaphor to diverge from the SL. Yan et al. (2010) find that the way a metaphor concept is modified in a TL relies strongly on the entrenchment rank. This means that a metaphor either remains preserved, or, when not shared in the TL’s system, it might be replaced by another metaphor of a higher entrenchment rank.

7 Conclusion

Our paper addresses conceptual metaphor via its expression in parallel linguistic data. We assume, following the literature, that a conceptual metaphor can be observed via its different linguistic instances. Conversely, we hypothesize that we can gain insight into conceptual metaphor by observing several linguistic metaphors in a language that belong together, as operationalized by their common translations in a parallel corpus. We test this hypothesis in four experiments carried out in a parallel news corpus annotated for linguistic metaphors, using multilingual BERT-based embeddings.

We find that a shared vocabulary space offers an opportunity to investigate how a cross-lingual phenomenon can be measured. Utilizing word distributions, we find that, first, many metaphors remain stable when they are translated, which is confirmed by measures of bilingual translation stability as well as by numeric clustering. Second, we found that words used metaphorically and literally differ in terms of their translation distributions, and that metaphor sources with the same translation set tend to be used only in one sense, the figurative one. Third, we observe that the translations of metaphors expressed in German compound nouns behave very similarly to the translations of simplicia, but our dataset containing a meaningful number of German compounds has an impact on the stability and depth-2 *entropy* of English linguistic metaphors. In the future, we plan to build on these empirical findings by designing a discrimination technique for literal and figurative contexts in parallel data.

References

- Berger, M. (2022). Transfer learning parallel metaphor using bilingual embeddings. In *Proceedings of the 3rd workshop on figurative language processing (flp)* (pp. 13–23). (<https://aclanthology.org/2022.flp-1.3.pdf>)
- Berger, M., Kiwitt, N., & Reimann, S. (2024, May). Applying transfer learning to German metaphor prediction. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 1383–1392). Torino, Italia: ELRA and ICCL. Retrieved from <https://aclanthology.org/2024.lrec-main.123/>
- Bisang, W., Hock, H. H., Winter, W., Stefanowitsch, A., & Gries, S. T. (2006). *Corpus-based approaches to metaphor and metonymy*. Mouton de Gruyter.
- Bizzoni, Y., & Ghanimifard, M. (2018). Bigrams and bilstms two neural networks for sequential metaphor detection. In *Proceedings of the workshop on figurative language processing* (pp. 91–101). (<https://aclanthology.org/W18-0911.pdf>)
- Borbely, A. F. (1998). A psychoanalytic concept of metaphor. *The International journal of psycho-analysis*, 79(5), 923. (<https://www.proquest.com/docview/1298185260>)
- Casasanto, D. (2009). When is a linguistic metaphor a conceptual metaphor. *New directions in cognitive linguistics*, 24, 127–146. (<https://www.degruyterbrill.com/document/doi/10.1075/hcp.24.11cas/html>)
- Cettolo, M., Girardi, C., & Federico, M. (2012). Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the conference of european association for machine translation (eamt)* (pp. 261–268). (<https://cris.fbk.eu/bitstream/11582/104409/1/WIT3-EAMT2012.pdf>)
- Choi, M., Lee, S., Choi, E., Park, H., Lee, J., Lee, D., & Lee, J. (2021). Melbert: Metaphor detection via contextualized late interaction using metaphorical identification theories. *arXiv preprint arXiv:2104.13615*. (<https://arxiv.org/abs/2104.13615>)
- Clausen, Y., & Nastase, V. (2019). Metaphors in text simplification: To change or not to change, that is the question. In *Proceedings of the fourteenth workshop on innovative use of nlp for building educational applications* (pp. 423–434). (<https://aclanthology.org/W19-4444.pdf>)
- Dorst, A. G. (2015). More or different metaphors in fiction? a quantitative cross-register comparison. *Language and Literature*, 24(1), 3–22. (<https://journals.sagepub.com/doi/pdf/10.1177/0963947014560486>)
- Even-Zohar, I. (1971). *Introduction to a theory of literary translation* (Unpublished doctoral dissertation). Ph. D. Thesis, Mimeo. Tel Aviv University (Hebrew, with extensive English summary).
- Fellbaum, C. (2010). Wordnet. In *Theory and applications of ontology: computer applications* (pp. 231–243). Springer. (<https://link.springer.com/book/10.1007/978-90-481-8847-5>)

- Gao, G., Choi, E., Choi, Y., & Zettlemoyer, L. (2018). Neural metaphor detection in context. *arXiv preprint arXiv:1808.09653*. (<https://arxiv.org/abs/1808.09653>)
- Gordon, J., Hobbs, J. R., May, J., Mohler, M., Morbini, F., Rink, B., ... Wertheim, S. (2015). A corpus of rich metaphor annotation. In *Proceedings of the third workshop on metaphor in nlp* (pp. 56–66). (<https://aclanthology.org/W15-1407.pdf>)
- Gray, R. M. (2011). *Entropy and information theory*. Springer Science & Business Media.
- Holmes, J. S. (2021). *Translated!: Papers on literary translation and translation studies. with an introduction by raymond van den broeck* (Vol. 7). Brill. (<https://brill.com/display/title/27760>)
- Holyoak, K. J., & Stamenković, D. (2018). Metaphor comprehension: A critical review of theories and evidence. *Psychological bulletin*, 144(6), 641. (http://www.cogsci.bme.hu/~ktkuser/KURZUSOK/BMETE47MC15/2019_2020_1/Cikkek/HolyoakStamenkovic2018.pdf)
- Jolliffe, I. (2005). Principal component analysis. *Encyclopedia of statistics in behavioral science*.
- Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic bulletin & review*, 7(2), 257–266. (<https://link.springer.com/content/pdf/10.3758/bf03212981.pdf>)
- Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers* (pp. 79–86). (<https://aclanthology.org/2005.mtsummit-papers.11.pdf>)
- Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giorn Dell'inst Ital Degli Att*, 4, 89–91.
- Köper, M., & Schulte im Walde, S. (2017). Improving verb metaphor detection by propagating abstractness to words, phrases and individual senses. In *Proceedings of the 1st workshop on sense, concept and entity representations and their applications* (pp. 24–30). (<https://aclanthology.org/W17-1903.pdf>)
- Kövecses, Z. (2018). Metaphor universals in literature. *Antares: letras e humanidades*, 10(20), 154–168.
- Kövecses, Z. (2000). The scope of metaphor. In A. Barcelona (Ed.), *Metaphor and metonymy at the crossroads* (pp. 79–92). Berlin, Germany: Mouton. (<https://www.degruyter.com/document/doi/10.1515/9783110894677/pdf>)
- Kövecses, Z. (2016). Conceptual metaphor theory. In *The routledge handbook of metaphor and language*. Taylor & Francis. (<https://www.jbe-platform.com/content/journals/10.1075/arc1.6.08kov>)
- Lakoff, G. (2009). The neural theory of metaphor. Available at SSRN 1437794. (<https://escholarship.org/content/qt9n6745m6/qt9n6745m6.pdf>)
- Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press. (https://www.academia.edu/download/50816110/Metaphor_we_live_by.pdf)
- Landauer, T. K. (2007). Lsa as a theory of meaning. *Handbook of latent semantic analysis*, 3, 32. (<http://cognaction.org/cogs105/readings/LSA.pdf>)

- Langacker, R. W. (1987). *Foundations of cognitive grammar: Volume i: Theoretical prerequisites* (Vol. 1). Stanford university press.
- Lin, Z., Ma, Q., Yan, J., & Chen, J. (2021). Cate: A contrastive pre-trained model for metaphor detection with semi-supervised learning. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 3888–3898). (<https://aclanthology.org/anthology-files/anthology-files/pdf/emnlp/2021.emnlp-main.316.pdf>)
- Liu, J., O'Hara, N., Rubin, A., Draelos, R., & Rudin, C. (2020). Metaphor detection using contextual word embeddings from transformers. In *Proceedings of the second workshop on figurative language processing* (pp. 250–255). (<https://aclanthology.org/2020.figlang-1.34.pdf>)
- Liu, Y., Ye, H., Weissweiler, L., Wicke, P., Pei, R., Zangenfeind, R., & Schütze, H. (2023). A crosslingual investigation of conceptualization in 1335 languages. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 12969–13000). Toronto, Canada: Association for Computational Linguistics. (<https://aclanthology.org/2023.acl-long.726/>)
- Luong, M.-T., Pham, H., & Manning, C. D. (2015). Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st workshop on vector space modeling for natural language processing* (pp. 151–159). (<https://aclanthology.org/W15-1521.pdf>)
- Maudslay, R. H., Pimentel, T., Cotterell, R., & Teufel, S. (2020). Metaphor detection using context and concreteness. In *Proceedings of the second workshop on figurative language processing* (pp. 221–226). (<https://aclanthology.org/2020.figlang-1.pdf>)
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26. (<https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf>)
- Mohler, M., Tomlinson, M. T., & Rink, B. (2015). Cross-lingual semantic generalization for the detection of metaphor. *Int. J. Comput. Linguistics Appl.*, 6(2), 117-140. (<https://www.gelbukh.com/ijcla/2015-2/IJCLA-2015-2-pp-117-140-Cross-lingual-Semantic-Generalization.pdf>)
- Papakyriakopoulos, O., Hegelich, S., Serrano, J. C. M., & Marco, F. (2020). Bias in word embeddings. In (p. 446–457). Association for Computing Machinery. (<https://dl.acm.org/doi/pdf/10.1145/3351095.3372843>)
- Pellauer, D., & Dauenhauer, B. (2002–2022). *Paul Ricoeur*. online; Stanford Encyclopedia of Philosophy. (<https://plato.stanford.edu/entries/ricoeur/>)
- Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). (<https://aclanthology.org/D14-1162.pdf>)
- Piccioni, S. (2013). What can metaphor tell us about the language of translation? *Procedia - Social and Behavioral Sciences*, 95, 354-362.

- Pragglejaz Group. (2007). MIP: A method for identifying metaphorically used words in discourse. *Metaphor and symbol*, 22(1), 1–39. (<https://www.tandfonline.com/doi/abs/10.1080/10926480709336752>)
- Ricoeur, P. (1978). *The rule of metaphor: Multi-disciplinary studies of the creation of meaning in language*. Routledge and Kegan Paul. (trans. Robert Czerny with Kathleen McLaughlin and John Costello, S. J., <https://muse.jhu.edu/pub/50/monograph/book/105427>)
- Sanchez-Bayona, E., & Agerri, R. (2022). Leveraging a new spanish corpus for multilingual and crosslingual metaphor detection. *arXiv e-prints*, arXiv–2210.
- Schäffner, C. (2004). Metaphor and translation: some implications of a cognitive approach. *Journal of pragmatics*, 36(7), 1253–1269. (<https://www.sciencedirect.com/science/article/pii/S0378216604000244?>)
- Schuster, J., & Markert, K. (2023). Nut-cracking sledgehammers: Prioritizing target language data over bigger language models for cross-lingual metaphor detection. In E. Breitholtz, S. Lappin, S. Loaiciga, N. Ilinykh, & S. Dobnik (Eds.), *Proceedings of the 2023 clasp conference on learning with small data (lsd)* (pp. 98–106). Gothenburg, Sweden: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2023.clasp-1.12>
- Semino, E., & Demjén, Z. (2016). *The routledge handbook of metaphor and language*. Taylor & Francis. (<https://www.academia.edu/download/79134201/9781315672953.pdf>)
- Shutova, E. (2010). Automatic metaphor interpretation as a paraphrasing task. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics* (pp. 1029–1037). (<https://aclanthology.org/N10-1147.pdf>)
- Shutova, E., & Teufel, S. (2010). Metaphor corpus annotated for source-target domain mappings. In *Lrec* (Vol. 2, pp. 2–2). (<https://openreview.net/pdf?id=8TVq60hx1bB>)
- Song, W., Zhou, S., Fu, R., Liu, T., & Liu, L. (2021). Verb metaphor detection via contextual relation learning. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 4240–4251). Online: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.acl-long.327> doi: 10.18653/v1/2021.acl-long.327
- Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., & Krennmayr, T. (2010). Metaphor in usage. *Cognitive Linguistics*, vol. 21, no. 4. (<https://www.degruyter.com/document/doi/10.1515/cogl.2010.024/html>)
- Teich, E. (2003). *Cross-linguistic variation in system and text: A methodology for the investigation of translations and comparable texts*. de Gruyter.
- Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., & Dyer, C. (2014). Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)*

- (pp. 248–258). (<https://aclanthology.org/P14-1024.pdf>)
- Tsvetkov, Y., Mukomel, E., & Gershman, A. (2013). Cross-lingual metaphor detection using common semantic features. In *Proceedings of the first workshop on metaphor in nlp* (pp. 45–51). (<https://aclanthology.org/W13-0906.pdf>)
- Upadhyay, S., Faruqui, M., Dyer, C., & Roth, D. (2016). Cross-lingual models of word embeddings: An empirical comparison. *arXiv preprint arXiv:1604.00425*. (<https://arxiv.org/pdf/1604.00425>)
- Van den Broeck, R. (1981). The limits of translatability exemplified by metaphor translation. *Poetics today*, 2(4), 73–87. (<https://www.jstor.org/stable/pdf/1772487.pdf>)
- Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.
- Veale, T., Shutova, E., & Klebanov, B. B. (2016). Metaphor: A computational perspective. *Synthesis Lectures on Human Language Technologies*, 9(1), 1–160. (https://openreview.net/pdf?id=8dwgVhag_wk)
- Vukovic, N. T. (2007). Corpus-based approaches to metaphor and metonymy. *Suvremena Lingvistika*, 33(63), 133–134. (https://www.researchgate.net/publication/292744446_Corpus-based_approaches_to_metaphor_and_metonymy)
- Vulić, I., & Moens, M.-F. (2015). Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th international acm sigir conference on research and development in information retrieval* (pp. 363–372). (<https://dl.acm.org/doi/pdf/10.1145/2766462.2767752>)
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... others (2016). Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*. (<https://arxiv.org/abs/1609.08144>)
- Yan, D., Noël, D., & Wolf, H.-G. (2010). Patterns in metaphor translation: a corpus-based case study of the translation of fear metaphors between english and chinese. *Using Corpora in Contrastive and Translation Studies*. Newcastle: Cambridge Scholars Publishing, 40–61. (http://www.lancaster.ac.uk/fass/projects/corpus/UCCTS2008Proceedings/papers/Ding_et_al.pdf)
- Zipf, G. K. (2016). *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio books.

Appendix A: Translation Graphs (Exp. 1)

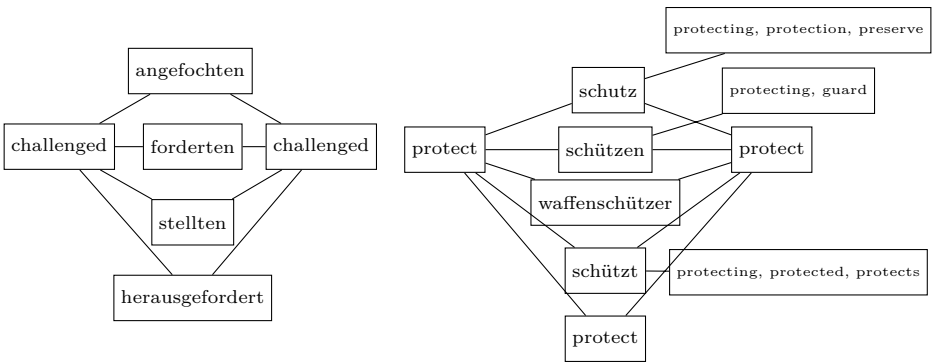


Figure 8: Translation graphs of “challenged” and “protect”

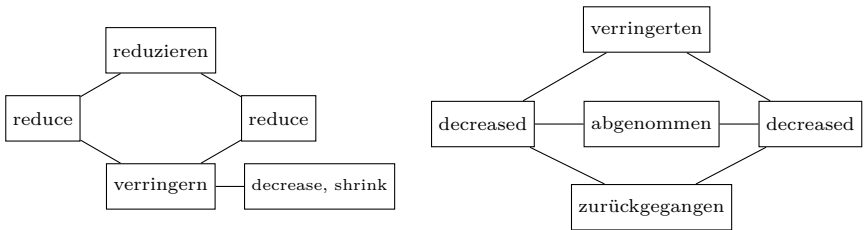


Figure 9: Translation graphs of “reduce” and “decreased”

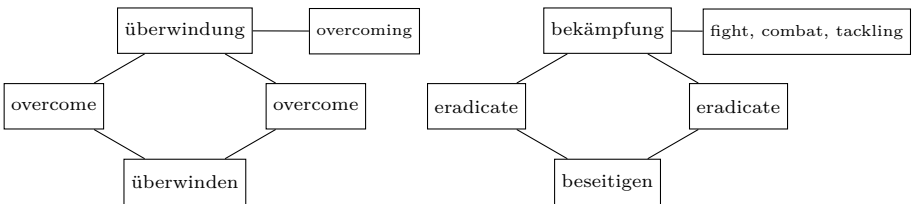


Figure 10: Translation graphs of “overcome” and “eradicate”

Appendix B: Significance Statistics (Exp. 3)

Table 6: Cos(ine) sim(ilarity), Z-test, significance level and p-value of the KS-test, stability and entropy; strong saturation indicates significant difference

lexelt	cos sim	Z-val	α	p-val
feeds	0.61	0.05	0.10	0.001
relief	0.68	0.07	0.16	0.000
decrease	0.56	0.06	0.10	0.002
bulwark	0.74	0.01	0.08	0.025
protected	0.78	0.01	0.04	0.478
escape	0.59	0.01	0.03	0.771
assault	0.69	0.02	0.06	0.180
reduce	0.73	0.03	0.04	0.730
evil	0.86	0.01	0.04	0.518
damage	0.86	0.01	0.05	0.224
killed	0.86	0.02	0.03	0.848
threatens	0.76	0.02	0.09	0.003
feeding	0.61	0.02	0.02	0.992
challenge	0.71	0.02	0.04	0.644
death	0.85	0.01	0.05	0.304
impact	0.67	0.01	0.06	0.100
defend	0.81	0.03	0.04	0.559
defended	0.49	0.02	0.10	0.001
guard	0.61	0.01	0.05	0.275
preserve	0.67	0.03	0.12	0.000
protects	0.63	0.05	0.11	0.000
safeguard	0.51	0.02	0.05	0.368
growth	0.91	0.01	0.05	0.224
advance	0.65	0.04	0.11	0.000
hole	0.79	0.01	0.08	0.013
block	0.56	0.04	0.07	0.052
hurdles	0.86	0.01	0.05	0.335
burden	0.75	0.05	0.04	0.730
progress	0.53	0.00	0.09	0.005
path	0.56	0.12	0.06	0.143
win	0.71	0.03	0.11	0.000
compete	0.76	0.02	0.06	0.113
competition	0.68	0.03	0.11	0.000
competitive	0.73	0.02	0.10	0.002
race	0.77	0.04	0.08	0.011
season	0.65	0.06	0.12	0.000
debate	0.77	0.00	0.04	0.602
revolution	0.83	0.02	0.05	0.403
oppose	0.80	0.04	0.02	0.973
winning	0.50	0.05	0.06	0.113
won	0.73	0.02	0.04	0.687
play	0.61	0.01	0.03	0.882
played	0.58	0.03	0.03	0.957
overcome	0.60	0.06	0.07	0.068
overcoming	0.68	0.04	0.04	0.439

Table 7: Cos(ine) sim(ilarity), Z-test result, statistical significance for given p-value of the Kolmogorov-Smirnov test, strong saturation indicates significant difference

lexelt	cos sim	Z-val	α	p-val
Erleichterung	0.62	0.05	0.05	0.368
verringert	0.57	0.01	0.06	0.180
Bollwerk	0.45	0.06	0.08	0.016
Sucht	0.67	0.01	0.11	0.000
behindert	0.66	0.04	0.08	0.016
Verringerung	0.49	0.04	0.06	0.100
böse	0.86	0.02	0.04	0.730
zerstört	0.86	0.02	0.07	0.039
Steuererleichterungen	0.56	0.06	0.02	0.973
Opfer	0.65	0.02	0.08	0.010
töten	0.75	0.02	0.11	0.000
getötet	0.78	0.01	0.10	0.001
droht	0.77	0.00	0.04	0.518
Opfern	0.61	0.07	0.03	0.848
ernährt	0.61	0.01	0.04	0.730
Zusammenbruch	0.68	0.05	0.03	0.771
Verlust	0.68	0.02	0.04	0.478
besiegen	0.69	0.05	0.09	0.006
Tod	0.70	0.08	0.14	0.000
Ende	0.70	0.01	0.11	0.000
schützt	0.71	0.03	0.07	0.052
verletzt	0.53	0.05	0.04	0.559
verteidigen	0.83	0.02	0.07	0.068
Schützen	0.83	0.00	0.04	0.602
kämpfen	0.87	0.03	0.12	0.000
greift	0.57	0.01	0.06	0.180
Entwicklung	0.66	0.04	0.11	0.000
Wachstum	0.80	0.02	0.08	0.010
fördert	0.53	0.01	0.08	0.016
bringt	0.77	0.05	0.11	0.000
Fortschritt	0.69	0.02	0.10	0.002
Weg	0.79	0.02	0.03	0.811
Suche	0.52	0.04	0.04	0.518
Loch	0.66	0.03	0.07	0.052
Hindernis	0.58	0.04	0.04	0.559
Hürden	0.73	0.03	0.02	0.984
Belastung	0.69	0.04	0.03	0.811
Förderung	0.65	0.03	0.06	0.100
konkurrieren	0.72	0.00	0.03	0.882
Steuerwettbewerb	0.65	0.06	0.06	0.113
Kampf	0.84	0.00	0.04	0.518
Revolution	0.85	0.01	0.04	0.730
Krieg	0.77	0.01	0.07	0.045
Gegner	0.61	0.03	0.10	0.001
widersetzt	0.61	0.03	0.05	0.224
verlieren	0.54	0.03	0.07	0.077
erzielte	0.55	0.00	0.03	0.848
Waffenrechten	0.77	0.01	0.10	0.000
Tragödie	0.74	0.01	0.09	0.007
gewinnen	0.55	0.05	0.10	0.001
gewonnen	0.67	0.03	0.10	0.001
ablehnt	0.68	0.01	0.03	0.811
Überwindung	0.66	0.07	0.05	0.275
konfrontiert	0.52	0.01	0.03	0.771
stärken	0.63	0.03	0.04	0.559
Schwierigkeiten	0.54	0.06	0.03	0.911

Correspondence

Maria Berger 

Ruhr University Bochum
German Department
Bochum, Germany
maria.berger-a2l@rub.de

Tatjana Scheffler 

Ruhr University Bochum
German Department
Bochum, Germany
tatjana.scheffler@rub.de

Enhancing ASR-based error analysis for atypical speech: post hoc integration of aphasic and dialectal phonetic features

Abstract

Despite current progress, automatic speech recognition (ASR) often struggles with non-standard speech, for example, influenced by dialectal or pathological features. (Re)training ASR models to accommodate these variations is not always possible due to limited data. This paper proposes applying the knowledge about non-standard (aphasic and dialectal) phonetic features to the ASR transcription post hoc. Using speech data from German speakers with aphasia who speak the Thuringian-Upper Saxon dialect, this study evaluates the impact of these modifications on an ASR-based error analysis pipeline. The approach helps to reduce automatic error rates on the recordings manually labelled as error-free. The performance of the pipeline also improves both in general acceptance or rejection of the responses and error attribution. General acceptance/rejection accuracy reaches the mean of 83.3%, which is considered sufficient to be used in a digital application for speech and language therapy support.

1 Introduction

Digital transformation has led to significant advancements in healthcare, particularly through the integration of artificial intelligence (AI). In the field of speech and language therapy (SLT), AI-driven tools are increasingly used for diagnosing and supporting people with aphasia (PWA) (Adikari, Hernandez, Alahakoon, Rose, & Pierce, 2024; Azevedo et al., 2024; Pottinger & Kearns, 2024). This paper is part of a broader project, *aphaDIGITAL*¹, aimed at developing a mobile application for German-speaking PWA. The project is regional and focuses on Southern Saxony-Anhalt and bordering regions of Saxony and Thuringia.

The app is to provide detailed, personalised feedback in oral exercises using automatic speech recognition (ASR) and subsequent automatic text processing, united into an error analysis pipeline (see Subsection 2.1). Currently, this pipeline includes ASR as well as phonemic/phonetic and semantic analysis components. This paper presents our work on the ASR component of the pipeline, specifically a solution to mitigate the effects of aphasic and dialectal features in the users speech on its automatic recognition, which is performed with selected open-source models.

1.1 ASR in aphasia speech therapy

Aphasia is an acquired language disorder caused by focal brain damage. Although there are several characteristic clinical profiles of the disorder, some linguistic symptoms are particularly salient

¹<https://aphadigital.sprechwiss.uni-halle.de/>

and widespread among PWA. Anomia, or difficulty in word retrieval, is one of the most common deficits Benson (1988), and it is typically treated with naming-oriented semantic exercises.

Naming-oriented semantic exercises have been automated with the help of ASR for Portuguese (Abad et al., 2013; Pompili et al., 2011), Australian English (Ballard, Etter, Shen, Monroe, & Tien Tan, 2019), British English (Barbera et al., 2021), and German (Heide et al., 2023; Hirsch, Tiggelkamp, Neumann, Frieg, & Knecht, 2025; Lin et al., 2022). In all these applications, the feedback is binary – correct or incorrect – without deeper analysis of the users speech errors. Reported accuracy rates of accepting or rejecting PWA's responses in an oral naming exercise are above the recommended accuracy threshold of 80% for use in SLT (McKechnie et al., 2018) – except for the case described by Ballard et al. (2019). They test the application with the participants suffering from apraxia of speech, comorbid to aphasia, and the application reaches only 75% accuracy. Heide et al. (2023) do not report results on aphasic speech, although the use of ASR is mentioned on the app's website². The development of the ASR technology for the Neolexon Aphasia-App (Lin et al., 2022) is thoroughly described in Klumpp (2024), but according to the information on the app's website³, the technology has not yet been implemented. The system developed by Hirsch et al. (2025) is still in the research stage.

Modern ASR systems present distinctive results on typical speech. For example, a word error rate (WER) for English reaches 0.05 (Protalinski, 2017; Xiong et al., 2017) and lower in certain cases (Graham & Roll, 2024; Picovoice, 2025). When tested on a range of datasets, German ASR solutions demonstrate similar results (Picovoice, 2025, see also Wirth & Peinl, 2022). However, their performance on impaired speech in general, and PWA's speech in particular, remains unreliable (Green et al., 2021; Rykova & Walther, 2025). Aphasic speech is characterised by linguistic searching behaviour and self-corrections, hesitation phenomena with or without vocal utterances, phonemic structure distortions, imprecise articulation or hyperarticulation. The irregularities are often inconsistent and unpredictable, which makes them hard to model – in distinction to motor speech disorders (e.g. dysarthria; see Caballero Morales & Cox, 2009; Mulfari, Carnevale, & Villari, 2023). For the current application, these challenges are compounded by dialectal variations (see Fischer & Jäck, 2023; Pompili et al., 2011), which already present difficulties for ASR systems even with German speakers without speech disorders (Költzsch, 2024). The present paper addresses this challenge in the context of the aforementioned project.

1.2 ASR of dialects

Non-standard language varieties pose significant challenges for the development of specialised ASR systems due to the absence of a standardised written form, internal variation, and the limited availability of training data. WER values of 0.2 and higher are common in dialectal ASR. Evaluation of dialectal speech recognition systems is also not trivial due to the absence of a common benchmark (Hinsvark et al., 2021).

Research has been carried out in several languages, including the dialects of American English (Harris, Mgbahurike, Kumar, & Yang, 2024), Arabic (for a review, see Alsayadi, Abdelhamid, Hegazy, Alotaibi, & Fayed, 2022), Chinese (for a review, see Li, Mai, Wang, & Ma, 2024),

²<https://lingo-lab.de>

³<https://neolexon.de/>

Thai (Suwanbandit, Naowarat, Sangpetch, & Chuangsuwanich, 2023), and German varieties and dialects (e.g., Bystrich, 2023). Hinsvark et al. (2021) discuss promising approaches to dialectal speech recognition, focusing on the challenge of phonetic variation within a language (which the authors refer to as "accented speech"). Phonetic variation consistently leads to degraded ASR performance, even in the absence of dialect-specific vocabulary (Harris et al., 2024).

To improve ASR performance on dialectal speech, it is common to train the corresponding models from scratch with deep learning or using more traditional methods like Gaussian mixture models and hidden Markov models, which can achieve a WER below 0.19 (see Nigmatulina, Kew, & Samardžić, 2020). Another solution involves applying transfer learning or fine-tuning to models pretrained on a standard variety (Bystrich, 2023; Elmahdy, Hasegawa-Johnson, & Mustafawi, 2014; Harris et al., 2024; Suwanbandit et al., 2023). However, both methods require a considerable amount of audio data and corresponding transcripts.

One possible way to circumvent the obstacle of limited audio data is to extend the pronunciation dictionary of the ASR system, which maps orthographic words to their pronunciation(s). Masmoudi, Khmekhem, Estève, Belguith, and Habash (2014) formulate rules for Tunisian Arabic dialect patterns and apply these rules to generate such extensions, achieving a WER of 0.23 with a trained-from-scratch ASR model. Bystrich (2023) explores phonetic mapping rules to supplement the phonetic lexicon of a Standard German ASR system with new Bavarian dialectal pronunciations for existing words and transcribe the spoken dialect into orthographic Standard German (cf. Nigmatulina et al., 2020 and Ulasik et al., 2021). Bystrich (2023) finds that the mapping of some phones is more effective than others, while some mappings have a negative impact.

Ali and colleagues proposes mining non-standard spellings of Egyptian Arabic from social media and applying them to dialect ASR in two ways. One method involves including the non-standard spellings into the language model, which reduced the WER of a trained-from-scratch Egyptian Arabic ASR model from 0.6 to 0.45 (Ali, Mubarak, & Vogel, 2014). The other method uses the non-standard spellings as targets in the evaluation of ASR performance (Ali, Nakov, Bell, & Renals, 2017). The authors call the resulting metric WER for dialects (WERd) and argue that WERd is a more appropriate metric for evaluating dialect ASR. Evaluated on a corpus of Egyptian dialect, an Arabic ASR system achieves a mean WERd of 0.36 compared to a WER of 0.46.

Swiss German

Swiss German – a continuum of Swiss German dialects, to be more precise – differs significantly from Standard German (of Germany). There are variations in phonology, lexicon, morphology and syntax (Scherrer & Rambow, 2010). Furthermore, the dialects differ from each other, primarily in phonology and vocabulary. Swiss German dialects are mainly used for oral and informal written communication and do not have a standardised orthography, while Standard German is used in more rather written contexts (Plüss et al., 2023). This makes automatic speech processing for Swiss German especially challenging. Additionally, publicly available annotated data remain scarce. STT4SG-350, a large corpus of Swiss German speech recordings transcribed into Standard German and balanced across dialectal regions was released only recently (Plüss et al., 2023).

There are other corpora, smaller and less balanced, from which some are transcribed in Swiss German (e.g., Samardžić, Scherrer, & Glaser, 2016; see also Plüss et al., 2023).

Researchers have explored different approaches to ASR for Swiss German dialects. Nigmatulina et al. (2020) develop the first multi-dialectal Swiss German ASR framework based on ArchiMob corpus (Samardžić et al., 2016) and experiment with two types of output: a dialectal phonemic spelling and a normalised transcription closely resembling Standard German. For ASR systems with dialectal output, the authors also introduce an alternative evaluation metric, FlexWER. It generally follows the idea of WERd (Ali et al., 2017), but makes use of the word-level mapping between the dialectal and the normalised representations: an output word is considered correct if it maps to the same normalised transcription as the corresponding word in the ground truth. The lowest WER score of 0.29 is achieved by an ASR system with normalised output. However, a corresponding ASR system with dialectal output achieves a lower FlexWER score (0.21) and also performs slightly better in terms of character error rate (CER): the lowest score is 0.146, with a difference of up to 0.015.

Ulasik et al. (2021) focus on the transcription of Swiss German speech into text in Standard German, treating the task as speech translation (ST). ST combines ASR and machine translation: spoken utterances in a source language are translated into a target language, either as text or speech. The translation can be performed stepwise – first ASR in the source language, then automatic translation into the target language (cascaded approach) – or with the input in the source language mapped directly to the output in the target language (end-to-end approach). Since the authors use the test corpus of Swiss German transcribed into Standard German (Plüss, Neukom, & Vogel, 2021), they employ an end-to-end ST approach for their experiments. BLEU (Bilingual Evaluation Understudy), a standard metric for automatic evaluation of machine translation, is used to assess model performance. BLEU scores range from 0 to 100, with higher scores indicating better performance.

Ulasik et al. (2021) train three types of models using various methods and corpora, select the best model of each type, and combine them into a single system, applying an ensemble method of majority voting both with and without perplexity weighting. The three models ensembled with simple majority voting achieve the highest BLEU score of 38.7 (cf. mean BLEU score of 75 achieved by the model fine-tuned with a large STT4SG-350 corpus in Plüss et al., 2023). Ulasik et al. (2021) experiment with a supervised spelling-correction model that identifies likely errors in the automatic transcriptions and suggests corrections. However, the four tested spelling-correction algorithms decrease overall system performance.

Finally, Kresic and Abbas (2024) present a method to normalise Swiss German written speech to Standard German, fine-tuning a multilingual large language model (mT5). The authors suggest integrating this method with speech recognition for increasing the effectiveness of the latter.

1.3 Research questions

In the absence of adequate data for (re)training ASR models, the current research is based on applying knowledge about non-standard (aphasic and dialectal) phonetic features post hoc to the ASR output. The implementation of the dialectal features adapts the method of using spelling alternatives for standard ASR evaluation introduced by Ali et al. (2017). In the present study,

dialectal pronunciation alternatives are generated based on non-standard pronunciation rules, similarly to Masmoudi et al. (2014) and Bystrich (2023). The main research questions are as follows:

- *Can a post hoc implementation of non-standard speech features improve the performance of an ASR-based error analysis pipeline?*
- *What are the ways to improve automatic error attribution?*

2 Materials & Methods

This work introduces modifications to the initial error analysis pipeline described in Rykova and Walther (2024a). The modified pipeline, which includes a module for post hoc phonetic features, is tested using existing speech recordings from PWA.

2.1 Error analysis pipeline

The original error analysis pipeline for naming exercises was presented in Rykova and Walther (2024a). In a naming exercise, the user is presented with a picture or text stimulus and is expected to utter a target word or phrase as a response. In the presented app, the user's response is transcribed with the ASR. The ASR output is then compared to the target transcription, and an error rate (ER) threshold is applied. If $ER = 0$, the response is error-free. The error is considered phonemic/phonetic if $0 < ER < ER \text{ threshold}$, and semantic if $ER \geq ER \text{ threshold}$.

ER is defined as the minimum of the character error rate (CER) between the orthographic forms of the ASR output and the target, and the phoneme error rate (PER) between their respective phonemic transcriptions. These transcriptions are generated automatically using the grapheme-to-phoneme (g2p) converter *bootphon espeak phonemizer* (Bernard & Titeux, 2021). Both CER and PER are normalised to range from 0 to 1. Based on previous findings (Rykova & Walther, 2025) and discussions with speech therapists, the ER threshold was set to 0.5.

For example:

- Target: *Gürtel* /gyrtəl/ ('belt')
ASR output: "schnalle" /ʃnalə/ ('buckle')
CER = 0.875, PER = 0.834 → ER = 0.834
ER > 0.5 → **semantic error**
- Target: *Gürtel* /gyrtəl/ ('belt')
ASR output: "gürtelm" /gyrtəlm/
CER = 0.143, PER = 0.286 → ER = 0.143
ER < 0.5 → **phonemic/phonetic error**

Four ASR models (Fleck, 2022; Grosman, 2021; Guhr, 2022; NVIDIA, 2022) previously shown to be suitable for aphasic speech (Rykova & Walther, 2025) were used. In one-word recognition tasks involving atypical speech, the lowest mean CER among these models is 0.11, observed in speech from alcohol-intoxicated speakers.

The pipeline evaluates each ASR models output both as a whole and as separate words, removes duplicates, and selects the output with the lowest ER for further analysis and feedback (cf. ensembled use of ASR models in Ulasik et al., 2021). Phonemic/phonetic errors are analysed with a phonemic/phonetic analysis component, and semantic errors are analysed with a semantic analysis component. The user has three attempts to produce the target correctly. After the first two attempts, they receive detailed feedback based on the error type. After the third attempt, the pipeline only checks whether the answer is correct (error-free) or incorrect, regardless of the error type.

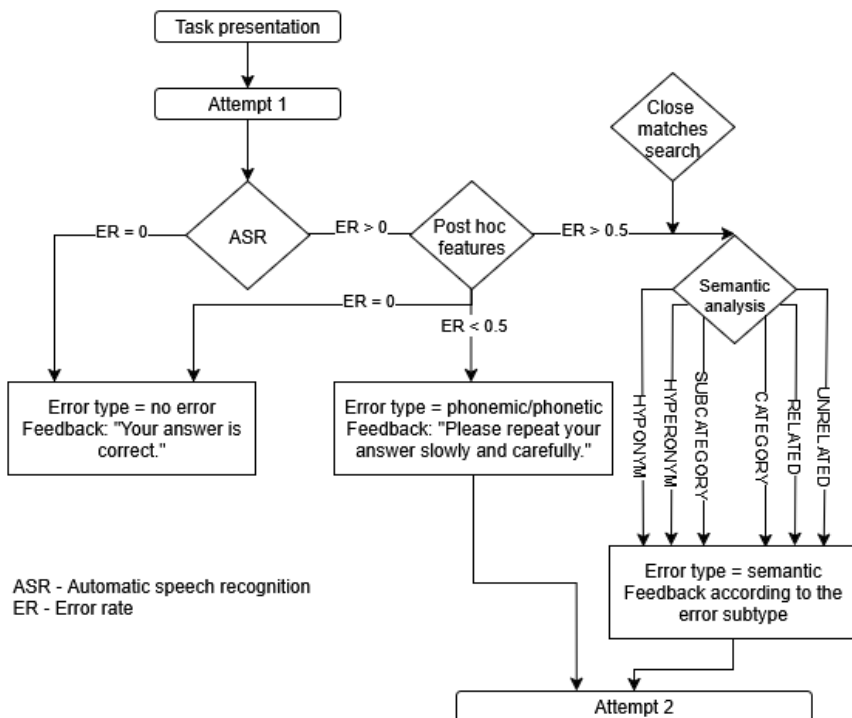


Figure 1: Modified error analysis pipeline section for attempt 1. Adapted from Rykova and Walther (2024a), p. 3387.

Upon internal discussion with a project team, the original Rykova and Walther's (2024a) pipeline has been modified:

- The ER threshold remains at 0.5, but if a half of the target is produced correctly, it is now considered a phonemic/phonetic error (previously – semantic error): phonemic/phonetic errors correspond to $0 < ER \leq 0.5$, and semantic errors correspond to $ER > 0.5$ (cf. Klumpp, 2024 and Redrovan-Reyes et al., 2019).
- Detailed analysis of phonemic/phonetic errors, previously executed after the first two attempts, is now removed after the first attempt. Phonemic/phonetic errors are analysed in details after the second attempt only.

This paper presents a new post hoc features component to be executed in every attempt after the ASR component when $ER > 0$. Figure 1 illustrates the modified pipeline for the first attempt. In a modified version, if the $ER > 0$ after the first comparison of the ASR output to the target (performed within the ASR component), non-standard features are introduced to the ASR output. The ER is then recalculated and, if it is greater than zero (the response is not error-free), compared to the threshold to determine the error type. If $0 < ER \leq 0.5$, the user is asked to repeat their response, thus initiating the second attempt. If $ER > 0.5$, the response is subject to semantic analysis with the corresponding pipeline component(s). For details on semantic analysis, please see Rykova and Walther (2023).

2.2 Post hoc features

For the research purposes, the analysis of test recordings was conducted in three steps, uniting aphasic features under step 2:

- **Step 1:** Original ASR output.
- **Step 2:** Implementation of aphasic speech features:
 - Syllabification and hyperarticulation: long pauses between syllables may cause them to be recognised as separate words. In such cases, spaces between ASR output segments are removed (as shown effective in Rykova & Walther, 2025);
 - Vowel prolongation: slow, careful speech may result in prolonged vowels. The vowel length symbol is removed, and vowel quality is adjusted (e.g. /i:/ → /ɪ/ – if such changes help to reduce PER).
- **Step 3:** Implementation of dialectal features: according to the project's geographic focus and available data, features from the Thuringian-Upper Saxon dialect group were selected (Rocholl, 2015; Siebenhaar, January, 2024; Wallraff, 2007). Based on these features, phonemic transformation rules were formulated. These rules of phonemic changes were applied to the target phonemic transcription in order to generate alternative targets and compare the phonemic transcription of the ASR output to them.

Table 1: Steps of non-standard phonetic features implementation.

Step	Description	Example
1	Separate words from ASR output	Target: <i>Buch</i> ASR outputs: ein buch, en woch, von wo Result: buch is accepted as an error-free answer
2	Aphasic features: Removing whitespaces	Target: <i>Schuhanzieher</i> ASR output: schuh anzieher, CER = 0.07 Output corrected: schuhanzieher, CER = 0 Result: the answer accepted as error-free
	Reducing vowel length (removing vowel length symbol)	Target: <i>Zigarre</i> /tsi:garə/ ASR output: zigare /tsi:garə/, PER = 0.125 Output corrected: /tsi:garə/, PER = 0 Result: the answer is accepted as error-free
3	Dialectal features: Generation of alternative target transcriptions based on phonemic transformation rules	Target: <i>Haustür</i> /haʊstʏ:r/ ASR output: haus dir /haʊs di:r/ Output corrected (step 2): hausdir /haʊsdi:r/ (Vowel length is not reduced at step 2 because it does not bring a reduction in PER) Alternative target generated based on rules $t \rightarrow d$ and $y \rightarrow i$: /haʊsdi:r/ Result: the answer is accepted as error-free

Due to technical reasons, in the app backend, removing whitespaces is implemented at the very first stage of the ASR output evaluation, corresponding to the ASR component in Figure 1 (see Rykova & Walther, 2025). Vowel prolongation and dialect features are addressed successively in the post hoc features component. Examples of the post hoc features implementation are provided in Table 1. When CER-based analysis suffices, only orthographic forms are shown. For PER-based analysis, automatic g2p transcriptions are included (please note that these transcriptions are intrinsically imperfect).

A full list of dialect-based phonemic transformation rules is provided in the Appendix. To address limitations of the g2p algorithm (Bernard & Titeux, 2021), which is constrained to Standard German phonemes, some dialectal features were approximated using existing phonemes or combinations (e.g., single consonants for long consonants, /o:/ for /ɔ:/). Additionally, all sounds that correspond to an orthographic *r* were unified as /r/ to avoid mismatches between the target and ASR output transcriptions caused by g2p inconsistencies.

2.3 Test recordings

A total of 412 audio recordings obtained from PWA during the Aachen Aphasia Test (AAT) (Huber, 1993) served as the test material. The recordings were provided by the Clinic of Cognitive

Neurology, University of Leipzig Medical Center. Access to the data was granted under an ethical agreement for research purposes. The dataset includes 219 recordings from the repetition task and 193 from the naming task. The naming task was performed by ten speakers (three female), seven of whom (two female) also completed the repetition task. All speakers were from Saxony and Saxony-Anhalt. No additional information about the participants was available.

All recordings were converted from mp3 to WAV format and resampled to 16 kHz to meet the requirements of the ASR models. The test material was extracted from longer recordings of the test sessions. The following (sub)tasks were selected for analysis:

- Naming task:
 - objects – one-stem nouns (e.g., *Gürtel* 'belt', *Zigarre* 'cigar')
 - objects – multi-stem nouns (e.g., *Staubsauger* 'vacuum cleaner', *Kühlschrank* 'refrigerator')
 - colours – adjectives (e.g., *rot* 'red', *orange* 'orange')
 - situations – sentences (e.g., *Hund* 'dog', *Pfeife* 'pipe')
- Repetition task:
 - monosyllabic words (e.g., *Ast* 'branch', *Glas* 'glass')
 - loan- and foreign words (e.g., *Telefon* 'phone', *Schokolade* 'chocolate')
 - compound words (including compounds with lexical morphemes)
(e.g., *Kraftfahrzeugschein* 'vehicle license', *Verantwortungslosigkeit* 'irresponsibility')

For the naming subtask 'situations – sentences', several individual words were selected from the expected multi-word responses. Although the other subtasks required single-word responses, speakers occasionally produced multi-word utterances (e.g., inserting an article before the target word).

The AAT naming task is used to assess word retrieval and verbal expression. It consists of words with one to five syllables long. The repetition task is used to assess auditory processing and speech motor control. It contains words of complex phonemic structures, one to eight syllables long. It also includes less frequent words than the naming task.

The extracted recordings were orthographically transcribed by the second author of the paper. Based on these transcriptions, the number of syllables in each of the utterances was determined. Using this number of syllables and the duration of the utterance, speech rate of a particular participant was calculated in syllables per second (syl/s).

2.4 Pipeline's performance analysis

The recordings were analysed auditorily by the first two authors of this paper and labelled as containing either no errors, phonemic/phonetic errors, or semantic errors. The second author, an SLT specialist, applied principles of SLT and dialect knowledge, while the first author, a clinical linguist, took into account ASR performance and technical considerations (cf. Klumpp, 2024). The final label was determined by consensus among the annotators (cf. Saz et al., 2009).

The label no errors was given to those utterances that would be fully accepted by an SLT specialist. These utterances might still contain false starts, self-corrections, syllabification, and dialectal features. The label "phonemic/phonetic error" refers to deviations in pronunciation that still render the word recognisable. The label "semantic error" refers to responses that are either unrelated to the target (including incorrect words or non-words) or are incomprehensible. It is important to note that the repetition task recordings contained very few semantic errors, as most participants were able to repeat the target words to some extent.

For the purpose of evaluating the pipeline's performance in terms of error rates, only the 249 recordings (120 from the naming task and 129 from the repetition task) that were labelled as error-free were used. All 412 recordings were used for analysing error attribution.

2.5 Statistical analysis

All statistical analyses were conducted in R (R Core Team, 2023) at a 95% confidence level. Error rates were analysed using the Students t-test, or the Wilcoxon Rank Sum Test when the sample size was too small for parametric testing.

Significance in the pipeline's accuracy changes was tested with the help of McNemar's Chi-squared test, designed for paired nominal data (McNemar, 1947). An exact binomial test (Hollander, Wolfe, & Chicken, 2014) was used to verify whether the results were statistically above the chance level.

3 Results & Discussion

This section presents the results of testing the error analysis pipeline using recordings from people with aphasia. It discusses the automatic error rates (ERs) on error-free recordings, the general accuracy of the pipeline in accepting or rejecting responses, its performance in attributing errors and disadvantages of different error misclassification scenarios. It also examines individual differences observed in the selected participants' data.

3.1 Error rates

A series of t-tests revealed no statistically significant differences in ERs between the naming and repetition tasks when analysing error-free recordings:

- step 1: $t(232) = 0.63$, $p = 0.532$
- step 2: $t(230) = 0.77$, $p = 0.442$
- step 3: $t(233) = 0.86$, $p = 0.388$

As a result, the two subsets were merged for further analysis. The ER decreased progressively across the three steps, as shown in Figure 2. Paired t-tests confirmed that these reductions were statistically significant:

- step 1 vs. step 2: $t(248) = 4.77$, $p < 0.01$

- step 2 vs. step 3: $t(248) = 3.76, p < 0.01$
- step 1 vs. step 3: $t(248) = 5.87, p < 0.01$

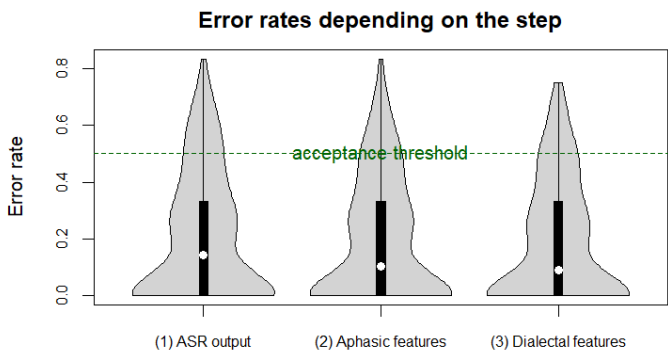


Figure 2: Decrease of the error rates on error-free recordings after non-standard features implementation.

Table 2 presents the changes in the central tendency and data dispersion of the ERs occurred at each step of the analysis.

Table 2: Changes in the error rates’ central tendency and data dispersion after non-standard features implementation.

	measure	step 1	step 2	step 3
	mean	0.2	0.192	0.184
	median	0.143	0.105	0.091
	standard deviation	0.225	0.218	0.211
	range	0-0.83	0-0.83	0-0.75
	interquartile range	0-0.33	0-0.33	0-0.33
	percentage of the values equal to zero	44.18%	44.98%	45.38%

While the improvements in ERs on error-free recordings were statistically significant, the absolute differences in average ER values remained modest. The post hoc implementation of aphasic features contributed more to reducing ERs to zero than dialectal features. However, both types of features contributed equally to reducing the mean ER values and their spread in relation to the mean (measured with standard deviation).

It is worth noting that the quality of the original recordings had a substantial impact on ASR performance. The data had been collected for different purposes and stored in mp3 format. In some cases, poor audio quality rendered the post hoc corrections ineffective (e.g., reducing an ER from

0.83 to 0.75 still results in rejecting the response as semantically incorrect). The proposed method shows potential in reducing the impact of aphasia and dialectal variations on ASR performance, but improving the ASR models and the quality of input audio remains essential.

3.2 Pipeline's general accuracy

Evaluated solely on whether users' responses are semantically accepted ($ER \leq 0.5$) or rejected ($ER > 0.5$) – without considering phonemic/phonetic errors (cf. Klumpp, 2024) – the pipeline's accuracy improved at each processing step: 78.2% $\rightarrow$ 81.1% $\rightarrow$ 83.3%. Results of the McNemar's Chi-squared test demonstrated statistical significance of the improvements:

- step 1 vs. step 2: $\chi^2(1) = 10.08$, $p = 0.001$
- step 2 vs. step 3: $\chi^2(1) = 7.11$, $p < 0.001$
- step 1 vs. step 3: $\chi^2(1) = 19.05$, $p = 0.008$

Precision for both accepted and rejected responses, as well as recall for accepted responses, increased from step to step. The recall for rejected responses remained unchanged. At the final step, the pipeline correctly identified 83.9% of semantically correct responses and 77.8% of semantically incorrect responses.

3.3 Pipeline's error attribution

Overall pipeline's accuracy in error attribution increased from 50.5% after step 1 to 53.9% after step 3, which is statistically significant according to the McNemar's Chi-squared test: $\chi^2(1) = 12.07$, $p < 0.001$. The accuracy of 53.9% is above chance level of 33.3%: $p < 0.001$ in an exact binomial test.

Due to the nature of the tasks, semantic errors are rare in the repetition task, while phonemic/phonetic errors occur more than twice as often in the naming task. According to the manual labelling of the given data, there were 6 semantic errors and 84 phonemic/phonetic errors in the repetition task, and 39 semantic errors and 34 phonemic/phonetic in the naming task. Because of these differing error distributions, the pipeline's error attribution was analysed in detail separately for each task – see Figure 3.

In the error attribution, the pipeline achieved accuracy score of 58% in the naming task and 50.2% in the repetition task, which are both above the chance level (p -values < 0.001 in an exact binomial test). There are not enough discordant pairs in each task to verify the statistical significance of the differences in accuracy scores between the steps.

In the confusion matrices in Figure 3, one can notice that there are not that many individual changes from step to step, and general picture remains similar. Numerically the greatest changes occur in the recall of phonemic/phonetic errors in the naming task: it increases from 55.9% after step 1 to 67.6% after step 3. The increase in the recall of phonemic/phonetic errors in the repetition task from 52.4% after step 1 to 60.7% after step 3, and the increase in the precision of identifying semantic errors in the naming task from 50.8% after step 1 to 58.5% after step 3 can be also

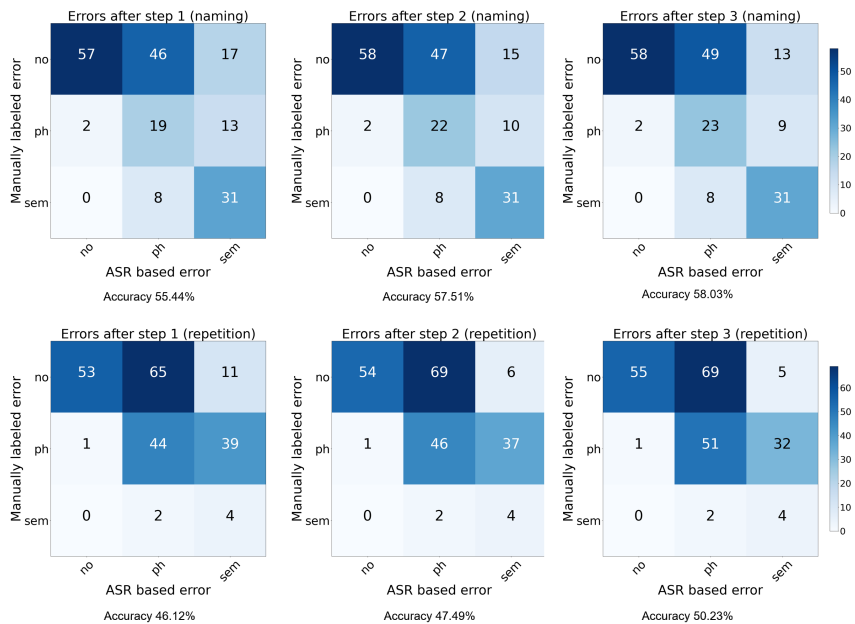


Figure 3: Confusion matrices for the pipeline’s error attribution in naming and repetition task at each step.

viewed as numerically considerable. Otherwise, the increase in precision and recall of the errors did not exceed 3%. In the rest of this subsection, only the values after step 3 are discussed.

The pipeline demonstrates similar precision in identifying error-free responses across both tasks: 96.7% in the naming task, and 98.2% in the repetition task. Only three recordings together from both tasks were misclassified as error-free, while containing phonemic/phonetic errors according to manual labelling. For example, user’s response *peife* was recognised as *pfeife* with target *Pfeife* pipe. Such misclassifications do not depend on the non-standard features and happen due to the ASR systems reliance on language and pronunciation models, which may favour plausible word forms.

Overseeing a phonemic/phonetic error is not viewed as a critical scenario because the application’s primary goal is to train vocabulary and not pronunciation. This type of pipeline’s errors could be addressed by training ASR models on pseudowords or actual mispronunciations with corresponding transcriptions.

The recall of error-free responses is, however, below 50% in both tasks: 48.3% in the naming task, and 42.6% in the repetition task. Due to the imperfections of ASR, error-free responses are frequently erroneously misclassified as phonemic/phonetic errors (and less frequently – as semantic errors), which makes the precision in identifying true phonemic/phonetic errors

relatively low: 28.8% in the naming task, and 41.8% in the repetition task. For the same reason, phonemic/phonetic errors are misclassified as semantic errors, which is most noticeably reflected in precision in identifying semantic errors in the repetition task: 9.8%. Probably, as soon as the target word was that difficult for the speaker that it caused them commit phonemic/phonetic errors, it became extremely difficult for the ASR. Nevertheless, the recall of phonemic/phonetic errors in both tasks and the precision in identifying semantic errors in the naming task are the areas that benefit the most from implementing the non-standard features (see above).

Misclassifying an error-free response as a phonemic/phonetic error is considered less disadvantageous for the user than misclassifying it as a semantic error. The feedback on phonemic/phonetic errors suggests that the response is semantically correct and encourages practising pronunciation, while the feedback on a semantic error would assume a semantically wrong response and the hints would confuse the user. Following the same logic, misclassifying phonemic/phonetic errors as semantic errors, which frequently occurred in the repetition task, is rather a harmful scenario.

The recall of semantic errors is relatively high and did not change with implementation of the non-standard features: 79.5% in the naming task, and 66.7% in the repetition task. In the erroneous scenarios, semantic errors were misclassified as phonemic/phonetic errors. It occurred when a semantically incorrect user's response was orthographically or acoustically similar to the target. The response could be a real word: *zigarette* cigarette vs. target *Zigarre* cigar, or a non-word: *schirrach* /ʃɪrɐx/ vs. target *Spruch* /ʃprʊx/.

Misclassifying a semantic error as a phonemic/phonetic error is not considered a harmful scenario because the user would still receive adequate, although not as tailored to the error type, feedback from the pipeline. Thus, after the first attempt, the pipeline would prompt the user to repeat the word slowly and clearly. After the second attempt, it would provide a phonemic/phonetic hint (see Figure 1 and Rykova & Walther, 2024a). Incorporating lexical analysis earlier in the pipeline could help reduce the acceptance of such semantically incorrect responses. Additionally, analysing syllabic structure could further aid in distinguishing between phonemic/phonetic and semantic errors, complementing the existing ER threshold.

3.4 Interspeaker variability

Given the variability in recording quality, speaker diagnoses, and individual speech characteristics, it seems reasonable to examine the results on a per-speaker basis. For this analysis, data from those seven speakers who completed both the naming and repetition tasks were used.

Table 3 presents the following information for each speaker:

- N: total number of recordings;
- N+R: number of recordings from the naming and repetition tasks;
- E-F: percentage of the recordings manually labelled as error-free;
- M dur: mean duration of recordings (s);
- M sp.r.: mean speech rate (syl/s);

- ER:M mean of the error rate (ER) on error-free items;
- ER:SD standard deviation of the ER on error-free items;
- Acc: general acceptance ($ER \leq 0.5$)/rejection ($ER > 0.5$) accuracy;
- AccN: error attribution accuracy in the naming task;
- AccR: error attribution accuracy in the repetition task.

Additionally, the SLT specialist noted that P4 exhibited laborious articulation, P6 showed signs of dysarthria, and P7 showed signs of apraxia of speech.

Table 3: Results presented per speaker

Speaker information								
		P1	P2	P3	P4	P5	P6	P7
	N items	57	72	61	39	35	67	49
	N+R	32+25	42+30	30+31	9+30	4+31	16+61	29+20
	E-F	56%	74%	89%	67%	94%	15%	39%
	M dur	1.71	0.87	1.54	2	0.82	1.15	2.5
	M sp.r.	1.85	3.15	2.12	1.54	3.45	1.86	0.85
Automatic analysis								
step 1	ER:M	0.08	0.15	0.17	0.24	0.25	0.3	0.54
	ER:SD	0.16	0.2	0.19	0.2	0.23	0.2	0.19
	Acc	86%	87.5%	93.4%	89.7%	85.7%	58.2%	38.8%
	AccN	78.1%	57.1%	41.4%	77.8%	25%	68.8%	13.8%
	AccR	64%	53.3%	63.6%	33.3%	38.7%	41.2%	26.3%
step 2	ER:M	0.08	0.14	0.16*	0.24*	0.24*	0.29	0.53
	ER:SD	0.16	0.2	0.18	0.2	0.22	0.19	0.19
	Acc	89.5%	87.5%	98.4%	92.3%	91.4%	61.2%	42.9%
	AccN	78.1%	59.5%	41.4%	77.8%	25%	75%	20.7%
	AccR	72%	56.7%	63.6%	33.3%	38.7%	41.2%	26.3%
step 3	ER:M	0.07	0.14*	0.16*	0.23*	0.23**	0.27	0.5**
	ER:SD	0.16	0.2	0.18	0.19	0.21	0.18	0.16
	Acc	89.5%	88.9%	98.4%	92.3%	94.3%	64.2%	53.1%
	AccN	78.1%	59.5%	41.4%	77.8%	25%	75%	24.1%
	AccR	76%	60%	63.6%	33.3%	38.7%	45.1%	36.8%
* - statistically significant difference from step 1								
** - statistically significant difference from step 1 and step 2								

A closer look at the results confirms the heterogeneity of speech difficulties among PWA. For four of the seven speakers (P4P7), the mean ER on error-free samples was above the mean of the

whole dataset at each step (0.2, 0.192, and 0.184, respectively). Nevertheless, mean ERs showed a statistically significant decrease following the implementation of non-standard features for five speakers. From the remaining two, P1 already had the lowest mean ER ($M = 0.08$ after step 1, and $M = 0.07$ after step 3), while P6 had very few error-free recordings ($n = 10$), limiting the reliability of their results.

General acceptance/rejection accuracy of the pipeline increased with the implementation of the non-standard features. After step 3, the pipeline achieved above 88% general acceptance/rejection accuracy for five speakers (P1P5), with a maximum of 98.4% for P3. For P6, the results were above chance level (cf. minimum accuracy of 65.1% in Ballard et al., 2019), and for P7, the results were at chance level. P7's speech rate is extremely low (0.85 syl/s), which is known to present difficulties for one-word ASR (Rykova & Walther, 2024b).

The three speakers with higher ERs and low error attribution accuracy in the repetition task (P4, P6, P7) were also those identified by the SLT specialist as having articulation difficulties. Motor speech disorders, comorbid to aphasia, challenge ASR systems even more. P6 and P7 supposedly suffer from motor speech disorders, comorbid to aphasia, which challenges ASR even more. comparably, the application described by Ballard et al. (2019), reaches the lowest mean acceptance/rejection accuracy than the other similar ones, including the one presented in this paper, probably because the participants of the study have apraxia of speech. However, dysarthric speech patterns are often predictable and can be modelled (Caballero Morales & Cox, 2009; Mulhari et al., 2023). Other articulatory difficulties may also be modelled upon further analysis.

Surprisingly, for P5, who produced 94% of their responses without errors, the mean ER was relatively high, and the error attribution accuracy was relatively low. This may be due to an imbalanced distribution of recordings across the tasks and very short recordings with P5's utterances. The latter aligns with findings from previous research (Rykova & Walther, 2024b).

4 Conclusions

This study presents a post hoc approach for incorporating aphasic and dialectal phonetic features into an ASR-based error analysis pipeline for people with aphasia. The proposed method is evaluated using recordings from the German Aachen Aphasia Test, specifically from naming and repetition tasks. Post hoc correction of ASR output based on knowledge of input speech characteristics shows promising results and might be applied as an alternative to spelling correction models trained on ASR errors, which do not necessarily improve ASR performance (Ulasik et al., 2021). The idea of using alternative reference targets in ASR performance evaluation (Ali et al., 2017; Nigmatulina et al., 2020) has also proven effective.

The results demonstrate that subsequent integration of non-standard phonetic features – first aphasic, then dialectal – leads to a statistically significant reduction in error rates on error-free recordings. Although the absolute improvements are modest (cf. results for speech under alcoholic intoxication), the enhancements are consistent across speakers and tasks, and the results are comparable to CER scores of an ASR system for dialectal speech with no disorders (Nigmatulina et al., 2020).

The general acceptance/rejection accuracy of the pipeline increases by 5%, reaching 83.3% (and up to 98.4% on an individual level). Such accuracy score is comparable to other SLT applications

that use ASR in automation of naming tasks and above the recommended threshold for SLT purposes (McKechnie et al., 2018).

In addition to improved acceptance/rejection accuracy, the pipeline also shows better performance in error attribution, reaching the overall accuracy of 53.9% and 58% accuracy in the naming task – both above the chance level. Since this is the first known application to offer automated error classification for PWA's speech, there are no directly comparable benchmarks.

Pipeline errors seem to stem mainly from limitations in the ASR component itself, which handles dysarthric or otherwise atypical articulatory patterns inadequately. This could be addressed with retraining, speaker-adaptation, and additional modelling of dysarthric and other articulatory difficulties.

To improve the recognition of phonemic/phonetic errors, (re)training or adapting ASR models using pseudowords or actual mispronunciations could be of help. Incorporating syllabic structure and lexical analysis earlier in the pipeline should strengthen differentiating phonemic/phonetic and semantic errors and decrease misclassification of semantically incorrect but orthographically/acoustically similar responses.

On an individual level, besides qualitative articulatory difficulties, the pipeline demonstrates sensitivity to speech rate and utterance duration. Extremely slow speech and short utterance durations can lead to poorer performance of the pipeline (in line with Rykova & Walther, 2024b).

5 Limitations and future work

The most significant difficulty faced during this study – and the broader aphaDIGITAL project – was the scarcity and poor quality of relevant data. The pipeline was tested using existing recordings of aphasic speech, which were collected under conditions different from those intended for the application. The dataset includes only ten speakers and covers a limited number of target words (cf. low numbers of participants in Abad et al., 2013; Ballard et al., 2019; Barbera et al., 2021). These constraints in both data quality and quantity mean that the current evaluation should be considered a feasibility study rather than a full validation. Nevertheless, the results demonstrate the viability of the proposed method and are comparable to those reported by other researchers working with pathological or dialectal speech.

Although the pipelines overall acceptance/rejection accuracy is similar to that of ASR-assisted SLT applications in other languages, its performance in error attribution remains insufficient for deployment in real-world therapeutic contexts. The majority of misclassifications appear to stem from limitations in the ASR component, which must be addressed upon collection of relevant data.

The current pipeline is designed for German-speaking PWA, particularly those who speak dialects from the Thuringian-Upper Saxon group. However, the pipeline's modularity allows for relatively easy substitution of dialectal features upon formulating the corresponding rules of phonemic changes, making it adaptable to other regional varieties or languages. With appropriate modifications, the pipeline could be extended to other therapeutic or language training contexts.

To address the limitations of this study, future work should focus on:

- live testing of the application with PWA in relevant scenarios;
- systematic data collection aligned with the apps intended use;

- expanding the dataset to include more speakers, a broader range of speech tasks, and greater dialectal and clinical diversity - and further analysing the impact of these factors.

References

- Abad, A., Pompili, A., Costa, A., Trancoso, I., Fonseca, J., Leal, G., ... Martins, I. P. (2013). Automatic word naming recognition for an on-line aphasia treatment system. *Computer Speech & Language*, 27(6), 1235–1248.
- Adikari, A., Hernandez, N., Alahakoon, D., Rose, M. L., & Pierce, J. E. (2024). From concept to practice: a scoping review of the application of AI to aphasia diagnosis and management. *Disability and Rehabilitation*, 46(7), 1288–1297.
- Ali, A., Mubarak, H., & Vogel, S. (2014). Advances in dialectal Arabic speech recognition: a study using Twitter to improve Egyptian ASR. In *Proceedings of the 11th International Workshop on Spoken Language Translation: Papers* (pp. 156–162).
- Ali, A., Nakov, P., Bell, P., & Renals, S. (2017). WERd: Using social text spelling variants for evaluating dialectal speech recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 141–148).
- Alsayadi, H. A., Abdelhamid, A. A., Hegazy, I., Alotaibi, B., & Fayed, Z. T. (2022). Deep investigation of the recent advances in dialectal Arabic speech recognition. *IEEE Access*, 10, 57063–57079.
- Azevedo, N., Kehayia, E., Jarema, G., Le Dorze, G., Beaujard, C., & Yvon, M. (2024). How artificial intelligence (AI) is used in aphasia rehabilitation: a scoping review. *Aphasiology*, 38(2), 305–336.
- Ballard, K. J., Etter, N. M., Shen, S., Monroe, P., & Tien Tan, C. (2019). Feasibility of automatic speech recognition for providing feedback during tablet-based treatment for apraxia of speech plus aphasia. *American journal of speech-language pathology*, 28(2S), 818–834.
- Barbera, D. S., Huckvale, M., Fleming, V., Upton, E., Coley-Fisher, H., Doogan, C., ... Crinion, J. (2021). NUVA: a naming utterance verifier for aphasia treatment. *Computer Speech & Language*, 69, 101221.
- Benson, D. F. (1988). Anomia in aphasia. *Aphasiology*, 2(3-4), 229–235.
- Bernard, M., & Titeux, H. (2021). Phonemizer: text to phones transcription for multiple languages in python. *Journal of Open Source Software*, 6(68), 3958.
- Bystrich, T. (2023). *Data-driven and rule-based approaches to improving Bavarian speech recognition* [Bachelor's Thesis]. Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS. (Available at <https://publica.fraunhofer.de/entities/publication/05257697-0aca-4ffc-aa1d-c953d56de790>)
- Caballero Morales, S. O., & Cox, S. J. (2009). Modelling errors in automatic speech recognition for dysarthric speakers. *EURASIP Journal on Advances in Signal Processing*, 2009, 1–14.
- Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.
- Elmahdy, M., Hasegawa-Johnson, M., & Mustafawi, E. (2014). Development of a TV broadcasts speech recognition system for qatari Arabic. In N. Calzolari et al. (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 3057–3061). European Language Resources Association (ELRA).
- Fischer, M., & Jäck, A. (2023). Einsatz KI-gestützter Diagnostik von Sprach- und Sprechstörungen bei neurodegenerativen Erkrankungen [Use of AI-supported diagnostics of speech and language disorders in neurodegenerative diseases]. *Nervenheilkunde*, 42(09), 626–634.

- Fleck, M. (2022). *XLS-R-300 Wav2Vec2 German with language model*. <https://huggingface.co/mfleck/wav2vec2-large-xls-r-300m-german-with-lm>. (Accessed on January 24, 2024)
- Graham, C., & Roll, N. (2024). Evaluating Openai's Whisper ASR: Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4(2), 025206.
- Green, J. R., MacDonald, R. L., Jiang, P.-P., Cattiau, J., Heywood, R., Cave, R., ... Tomanek, K. (2021). Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases. In *Proceedings of Interspeech 2021* (pp. 4778–4782).
- Grosman, J. (2021). *Fine-tuned XLSR-53 large model for speech recognition in German*. <https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-german>. (Accessed on January 24, 2024)
- Guhr, O. (2022). *wav2vec2-large-xlsr-53-german-cv9*. <https://huggingface.co/oliverguhr/wav2vec2-large-xlsr-53-german-cv9>. (Accessed on January 24, 2024)
- Harris, C., Mgbahurike, C., Kumar, N., & Yang, D. (2024). Modeling gender and dialect bias in automatic speech recognition. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the association for computational linguistics: Emnlp 2024* (pp. 15166–15184). Miami, Florida, USA: Association for Computational Linguistics.
- Heide, J., Netzebandt, J., Ahrens, S., Brüsch, J., Saalfrank, T., & Schmitz-Antonischki, D. (2023). Improving lexical retrieval with LingoTalk: an app-based, self-administered treatment for clients with aphasia. *Frontiers in Communication*, 8, 1210193.
- Hinsvark, A., Delworth, N., Rio, M., McNamara, Q., Dong, J., Westerman, R., ... Jette, M. (2021). Accented speech recognition: A survey. *ArXiv, abs/2104.10747*. Retrieved from <https://api.semanticscholar.org/CorpusID:233347110>
- Hirsch, H.-G., Tiggelkamp, Y., Neumann, C., Frieg, H., & Knecht, S. (2025). Evaluating the user interface of the rehalingo speech training system with aphasic patients. In S. Gravunder (Ed.), *Studentexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2025* (pp. 61–68). TUDpress, Dresden.
- Hollander, M., Wolfe, D. A., & Chicken, E. (2014). *Nonparametric statistical methods*. John Wiley & Sons.
- Huber, W. (1993). *Aachener Aphasie Test (AAT)*. Göttingen, Zürich: Verlag für Psychologie Hogrefe.
- Klump, P. (2024). *Phonetic transfer learning from healthy references for the analysis of pathological speech* (Doctoral dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany). Open FAU. (Available at <https://open.fau.de/items/d0c6b800-e217-4049-ab1f-a746fc9b3966>)
- Kresic, M., & Abbas, N. (2024). Normalizing swiss german dialects with the power of large language models. *Procedia Computer Science*, 244, 287–295. (6th International Conference on AI in Computational Linguistics)
- Költzsch, T. (2024, January). *Sprachcomputer der Sparkasse Nürnberg kann kein Fränkisch*. Retrieved from <https://www.golem.de/news/kommunikationspanne-sprachcomputer-der-sparkasse-nuernberg-kann-kein-fraenkisch>

-2401-181689.html

- Li, Q., Mai, Q., Wang, M., & Ma, M. (2024). Chinese dialect speech recognition: a comprehensive survey. *Artificial Intelligence Review*, 57(2), 25.
- Lin, Y., Klumpp, P., Pfab, J., Abdelioua, A., Gebray, D., & Späth, M. (2022). Eine automatische Sprachbewertung für die neolexon Aphasie-App mithilfe Künstlicher Intelligenz [Automatic language assessment with artificial intelligence for the neolexon aphasia app]. In *Sprachtherapie aktuell: Forschung-Wissen-Transfer 9 (1): XXXIV, Workshop Klinische Linguistik e2022-11*.
- Masmoudi, A., Khmekhem, M. E., Estève, Y., Belguith, L. H., & Habash, N. (2014). A corpus and phonetic dictionary for Tunisian Arabic speech recognition. In N. Calzolari et al. (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 306–310). European Language Resources Association (ELRA).
- McKechnie, J., Ahmed, B., Gutierrez-Osuna, R., Monroe, P., McCabe, P., & Ballard, K. J. (2018). Automated speech analysis tools for childrens speech production: a systematic literature review. *International journal of speech-language pathology*, 20(6), 583–598.
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157.
- Mulfari, D., Carnevale, L., & Villari, M. (2023). Toward a lightweight ASR solution for atypical speech on the edge. *Future Generation Computer Systems*, 149, 455–463.
- Nigmatulina, I., Kew, T., & Samardžić, T. (2020). ASR for non-standardised languages with dialectal variation: the case of Swiss German. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects* (pp. 15–24).
- NVIDIA. (2022). *NVIDIA Conformer-Transducer Large (de)*. https://huggingface.co/nvidia/stt_de_conformer_transducer_large. (Accessed on January 24, 2024)
- Picovoice. (2025, February). *Speech-to-text benchmark*. Retrieved from <https://github.com/Picovoice/speech-to-text-benchmark/blob/master/README.md>
- Plüss, M., Deriu, J., Schraner, Y., Paonessa, C., Hartmann, J., Schmidt, L., ... Cieliebak, M. (2023). STT4SG-350: A speech corpus for all Swiss German dialect regions. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 1763–1772). Toronto, Canada: Association for Computational Linguistics.
- Plüss, M., Neukom, L., & Vogel, M. (2021). Swisstext 2021 task 3: Swiss german speech to standard german text. In *Proceedings of the swiss text analytics conference* (Vol. 2021).
- Pompili, A., Abad, A., Trancoso, I., Fonseca, J., Martins, I. P., Leal, G., & Farrajota, L. (2011). An on-line system for remote treatment of aphasia. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies* (pp. 1–10).
- Pottinger, G., & Kearns, Á. (2024). Big data and artificial intelligence in post-stroke aphasia: a mapping review. *Advances in Communication and Swallowing*, 27(1), 41–55.
- Protalinski, E. (2017, May). *Googles speech recognition technology now has a 4.9% word error rate*. Retrieved from <https://venturebeat.com/2017/05/17/googles-speech-recognition-technology-now-has-a-4-9-word>

-error-rate/

- R Core Team. (2023). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Redrovan-Reyes, E., Chalco-Bermeo, J., Robles-Bykbaev, V., Carrera-Hidalgo, P., Contreras-Alvarado, C., Leon-Pesantez, A., ... Olivo-Deleg, J. (2019). An educational platform based on expert systems, speech recognition, and ludic activities to support the lexical and semantic development in children from 2 to 3 years. In *2019 IEEE Colombian Conference on Communications and Computing (COLCOM)* (pp. 1–6).
- Rocholl, M. J. (2015). *Ostmitteldeutsch—eine moderne Regionalsprache? Eine Untersuchung zu Konstanz und Wandel im thüringisch-obersächsischen Sprachraum [East-Central German a modern regional language? An investigation into constancy and change in the Thuringian-Upper Saxon language area]*. Georg Olms Verlag.
- Rykova, E., & Walther, M. (2023). Concept for semantic error analysis in a mobile application for speech and language therapy support. In C. Draxler (Ed.), *Studentexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2023* (pp. 127–133). TUDpress, Dresden.
- Rykova, E., & Walther, M. (2024a). AphaDIGITAL – digital speech therapy solution for aphasia patients with automatic feedback provided by a virtual assistant. In *Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS)* (pp. 3385–3394). University of Hawai'i at Mnoa.
- Rykova, E., & Walther, M. (2024b). Linguistic and extralinguistic factors in automatic speech recognition of German atypical speech. In P. H. Luz de Araujo, A. Baumann, D. Gromann, B. Krenn, B. Roth, & M. Wiegand (Eds.), *Proceedings of the 20th conference on natural language processing (KONVENS 2024)* (pp. 358–367). Association for Computational Linguistics.
- Rykova, E., & Walther, M. (2025). Evaluation of German ASR solutions in the context of speech and language therapy support of people with aphasia. *Loquens*, 12, e116.
- Samardžić, T., Scherrer, Y., & Glaser, E. (2016). ArchiMob - a corpus of spoken Swiss German. In N. Calzolari et al. (Eds.), *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (pp. 4061–4066). European Language Resources Association (ELRA).
- Saz, O., Yin, S.-C., Lleida, E., Rose, R., Vaquero, C., & Rodríguez, W. R. (2009). Tools and technologies for computer-aided speech and language therapy. *Speech communication*, 51(10), 948–967.
- Scherrer, Y., & Rambow, O. (2010). Natural language processing for the Swiss German dialect area. In *Semantic approaches in natural language processing – proceedings conference on natural language processing 2010 (konvens)*.
- Siebenhaar, B. (January, 2024). Personal communication.
- Suwanbandit, A., Naowarat, B., Sangpetch, O., & Chuangsuwanich, E. (2023). Thai dialect corpus and transfer-based curriculum learning investigation for dialect automatic speech recognition. In *Proceedings of Interspeech 2023* (pp. 4069–4073).
- Ulasik, M. A., Hürlimann, M., Dubel, B., Kaufmann, Y., Rudolf, S., Deriu, J., ... Cieliebak, M. (2021). ZHAW-CAI: ensemble method for Swiss German speech to standard german text.

In *Swiss text analytics conference–swisstext 2021*, online, 14–16 june 2021.

- Wallraff, U. (2007). *Ausgewählte phonetische Analysen zur Umgangssprache der Stadt Halle an der Saale [Selected phonetic analyses of the colloquial language of the city of Halle an der Saale]* (Doctoral dissertation, Martin-Luther-Universität Halle-Wittenberg, Halle, Germany). Universitäts- und Landesbibliothek Sachsen-Anhalt. (Available at <https://opendata.uni-halle.de/handle/1981185920/9620>)
- Wirth, J., & Peinl, R. (2022). Automatic speech recognition in German: a detailed error analysis. In *2022 IEEE International Conference on Omni-layer Intelligent Systems (COINS)* (pp. 1–8).
- Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M. L., Stolcke, A., ... Zweig, G. (2017). Toward human parity in conversational speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12), 2410–2423. doi: 10.1109/TASLP.2017.2756440

Appendix

Rules of phonemic changes based on the characteristics of the Thuringian-Upper Saxon dialect group

The presentation of the context roughly follows the notations used for phonological rules (Chomsky & Halle, 1968), where C stands for consonant, V stands for vowel, # stands for 'no sound' and is also used to indicate the boundaries of words. The transcriptions were generated automatically using the grapheme-to-phoneme converter *bootphon espeak phonemizer* (Bernard & Titeux, 2021) with some modifications (see Section 2.2).

Schwa (ə) deletion and following assimilation of place (if applicable) could be a feature of informal speech or a dialectal pronunciation. The discussion of the reasons for this phenomenon is outside the scope of the current work, and the corresponding cases are presented as dialectal features following Wallraff (2007).

phonemic change	context (if applicable)	example: standard transcription → alternative transcription
a → ɔ	___ r	Markt /markt/ → /mɔrkt/
a: → o:	___ r	Nahrung /na:rʊŋ/ → /no:rʊŋ/
ɪ → ʏ	___ r, ___ m, ___ ʃ, r ___	irgendwie /ɪrgəndvi:/ → /ʏrgəndvi:/
i: → ɪ	___ C(ə)l, ___ C(ə)r, ___ C(ə)n	Zwiebel /tsvi:bəl/ → /tsvɪbəl/
ʏ → ɪ		Brücke /brykə/ → /brikə/
y: → i:		Blüte /bly:tə/ → /bli:tə/
ʊ → ɔ	___ r	Turm /tʊrm/ → /tɔrm/
œ → ɛ		Knöchel /knœçəl/ → /knɛçəl/
ø: → e:		Vögel /fø:gəl/ → /fe:gəl/
e: → ɛ:		Meer /me:r/ → /mɛ:r/
aʊ → o:		auch /aʊx/ → /o:x/
aɪ → e:		Fleisch /flaɪʃ/ → /fle:ʃ/
ɔɪ → e:		Leute /lɔɪtə/ → /le:tə/
ə → #	___ n, ___ l, ___ m	besen /be:zən/ → /be:zn/
p(ə)n → bm	___ #	umstülpen /ʊmʃtʏlpən/ → /ʊmʃtʏlbm/
b(ə)n → bm	___ #	leben /le:bən/ → /le:bm/
k(ə)n → gŋ	___ #	schlucken /ʃlʊkən/ → /ʃlʊgŋ/
nən → n	___ #	brennen /brɛnən/ → /brɛn/
mən → m	___ #	kommen /kɔmən/ → /kɔm/
ŋən → ŋ	___ #	fangen /faŋən/ → /faŋ/
pf(ə)n → pfm	___ #	hüpfen /hʏpfən/ → /hʏpfm/
p → b		Papier /papi:r/ → /babi:r/

phonemic change	context (if applicable)	example: standard transcription → alternative transcription
t → d		Teufel /tʊøfəl/ → /dʊøfəl/
k → g		Kalt /kalt/ → /galt/
b → p	___ r, ___ l	Blüt /bly:t/ → /ply:t/
d → t	___ r, ___ l	Droge /dro:gə/ → /tro:gə/
g → k	___ r, ___ l	Glocke /glɔkə/ → /klɔkə/
nd → n		finden /findən/ → /finən/
k → ç	V [+front] ___, l ___, r___	Krieg /kri:k/ → /kri:ç/
k → x	V [-front] ___ C [+alveolar]	Jagd /ja:kt/ → /ja:xt/
g → j	# ___ V, V ___ V	genau /gənaʊ/ → /jənaʊ/
g → ç	V [+front] ___ V	Ziege /tsi:gə/ → /tsi:çə/
g → ʃ	V [+front] ___ V	Ziege /tsi:gə/ → /tsi:ʃə/
g → x	V [-front] ___ V	Beluga /belu:ga/ → /belu:xa/
ç → ʃ		Schicht /ʃɪçt/ → /ʃɪft/
pf → f	# ___, m___ #	Pfeife /pfaiɸə/ → /faiɸə/
pf → b	___ V	Kampf /kampɸ/ → /kamɸ/
pf → p	___ #	Apfel /apɸəl/ → /abəl/
ts → s	___ #	Zopf /tsɔpɸ/ → /tsɔp/
		Holz /hɔlts/ → /hɔls/

Correspondence

Eugenia Rykova 

Catholic University of Eichstätt-Ingolstadt &
University of Eastern Finland
eugenia.rykova@ku.de

Elisabeth Zeuner

Martin Luther University Halle-Wittenberg

Susanne Voigt-Zimmermann 

Martin Luther University Halle-Wittenberg

Mathias Walther 

Technical University of Applied Sciences TH Wildau