Metaphor Mining in Historical German Novels: An Unsupervised Learning Approach.

Stefan Pernes
University of Würzburg
Würzburg, Germany
stefan.pernes@uni-wuerzburg.de

***Abstract*—This paper describes a work-in-progress to identify and categorize metaphorical language use in a large corpus of historical German novels. An unsupervised learning method is utilized to detect metaphorical expressions and underlying conceptual metaphors. Furthermore, an extension is proposed that allows for the analysis of diachronic developments of modeled metaphor types. A corpus ranging from the 16th to the 20th century serves to illustrate the challenges of this approach as well as its potential, not only as a tool for the analysis of stylistic variation, but also as a glimpse into the conceptual world views embedded in the texts under examination.**

***Index Terms*—computational metaphor identification; conceptual metaphor; diachronic analysis; literary history**

I. Introduction

With figurative language being a ubiquitous phenomenon - on average appearing in every third sentence of general-domain text [1] - the development of metaphor identification systems turns out to be an important component in many text mining use cases, even more so in the light of *smart data* approaches aiming for deeper semantic analysis and representations of text. The research project being described addresses the use of such a system in analyzing literary developments over a long period of time, as well as targeting epistemic realities, not only of the literary but the actual world-making in examined time periods. This view of metaphor, as not only a rhetorical device but as deeply ingrained in every-day language, is known as *Conceptual Metaphor Theory* [2] and comprehends the phenomenon as a genuinely cognitive mechanism that manifests itself in language in the form of surface metaphorical expressions (also: linguistic metaphors). Such expressions can be described as a systematic mapping, or rather, a projection of one domain of experience (the source, e.g. *war*) onto another (the target, e.g. *argument*). By now widely adopted, it is an empirically grounded approach [e.g. 3] that allows for a sensible aggregation and tracing of metaphorical expressions in a large corpus such as the one in question. Furthermore, using conceptual metaphor as a level of analysis should enable the uncovering of conceptual world views - e.g. cultural and moral models embedded in the texts -, ultimately leading to a empirically grounded cognitive-anthropological perspective on figurative language.

II. Dataset

The dataset [1] consists of about 1700 German literary works, ranging from the early 16th to the early 20th century. Given the large timespan covered, significant orthographic and lexical variation is to be expected until the advent of a widely adopted standardization in writing at the end of the 18th century. Apart from a normalization, what also has to be taken into account, is a balancing of the dataset suited for the metaphor identification task, which could be accomplished by including relevant historical encyclopedias and dictionaries.

III. Metaphor Identification using Hierarchical Clustering

The methodological approach adopted here is an unsupervised metaphor identification system using *Hierarchical Graph Factorization Clustering* (HGFC) as proposed by Shutova and Sun [4]. It is a bottom-up clustering approach primarily focused on conventionalized metaphor, but depending on the size and balancing of the dataset, also capable of identifying novel metaphor. The main intuition behind this approach is "to investigate how metaphor partitions the linguistic feature space" [1] on the basis of selectional preferences. E.g., in order to model the common verbal metaphor type as demonstrated here, a number of most frequent nouns are extracted from the corpus, alongside the verbs to which the nouns stand in a certain grammatical relation (subject, direct object, or indirect object relation). The algorithm aims to produce clusters of nouns according to the calculated feature distributions (verbs and their relation types), which can be ranked according to membership probabilities for a certain query (a concept/noun) - with literal usage being the top ranked cluster and figurative usages following thereafter. Soft clustering such as HGFC is especially suited for this approach to metaphor identification, as source-target domain mappings imply multiple connections between vertices which can be accounted for in this setting.

A. Preprocessing and Feature Extraction

Preprocessing is performed using a modular pipeline [2] including PoS-tagger, lemmatizer, and dependency parser components for

[1] http://www.germanistik.uni-wuerzburg.de/lehrstuehle/computerphilologie/forschung/projekte/digibib

[2] https://github.com/DARIAH-DE/DARIAH-DKPro-Wrapper

German text. Following this, the 2000 most frequent nouns in the corpus and their corresponding verbs (using the grammatical relations described above) are extracted. The resulting noun-verb feature matrix containing relative feature frequencies for each noun is then used to calculate a noun-noun similarity matrix using Jensen-Shannon Divergence on the co-occurence vectors.

B. Example Analysis

As a first test case, the baseline system [3] described in Shutova and Sun [4] is reproduced. A subset of 143 novels from the dataset, covering the full range of time periods, is used to generate the following examples. While this baseline displays shortcomings, it can be used to uncover some metaphorical relations in the test corpus. Also, quite a few nouns could be observed to cluster in synonymous, antonymous, and metonymous relations. Clustering levels were inspected manually and, of course, naming relevant clusters after their specific conceptual metaphors has to be done manually - using Lakoff's master metaphor list [5].

Examples of source-target mappings found in the corpus, with corresponding conceptual metaphors taken as labels:

IDEAS ARE FOOD + DESIRE IS HUNGER

- source: *hunger, apfel*
- target: *gemahl, erinnerung, bildung, idee, anspruch, weibe, wollust, neuigkeit*

HARM IS LACKING A NEEDED POSSESION

- source: *gift*
- target: *versuchung, besitz, göttin*

ARGUMENT IS WAR + STATES ARE LOCATIONS

- source: *gefecht*
- target: *politik, paradies, nebenzimmer, hintergrund, beginn*

BELIEFS ARE BEINGS WITH A LIFE CYCLE

- source: *korn, winter*
- target: *aufregung, revolution, oktober, schöpfung, freundschaft, zucht*

All examples are gathered by manually examining level 6 of the hierarchical clustering, which turned out to exhibit the optimal level of generality for this experiment. What can be observed here is that the clusters produced by agglomerative clustering display a lack of coherence. This is also an observation made by Shutova and Sun [4], stating that HGFC produces more pure and complete clusters. One "important reason AGG fails is that it by definition organizes all concepts into tree and optimizes its solution locally, taking into account a small number of clusters at a time. However, being able to discover connections between more distant domains and optimising globally over all concepts is crucial for metaphor identification" [ibid.]. What also becomes clear is that the conceptual metaphor mappings manually assigned using the master metaphor list are largely applicable, but by no means exhaustive, suggesting that other language material and/or a bottom-up computational approach like the one in question will expose additional conceptual metaphors, which are not included in the list. Another observation that can be made - looking at the underlying verb features not reproduced here - is that they as well lack semantic coherence. The metaphorical source-target mappings displayed here can thus be interpreted as an effect of the entire distributed representation. Viewed in this light, those are promising results in the need for a (substantially) larger dataset.

C. Next Steps

First and foremost, the next step is to implement the original hierarchical graph factorization clustering algorithm as described in [4]. Also, it is essential to prepare the entire corpus using the preprocessing methods described above. That in turn provides a reasonable basis for experimenting with various parameters and implementing a mechanism for modeling the diachronicity of the data in question.

- One obvious parameter to experiment with is the normalization of orthographic variation, which had been left as is for the initial data exploration in order to gather more information on the distribution of different writing norms in the corpus.
- Another apparent factor is modeling the different forms that metaphorical expressions can assume. An expansion towards adjective-noun constructions is straightforward and, together with the verbal expressions already included in the model, accounts for a large proportion of metaphorical expressions normally encountered. Other types that might be worth considering are similes and copula constructions.
- In order to establish a suitable method for the modeling of diachronic variation more experiments will be needed: One possibility is to generate partial models which correspond to pre-defined time periods, but this has the downside of requiring a large enough data set for each period seperately. Another approach is to introduce metadata into the model, e.g. building a second feature matrix that registers where in the corpus each noun-feature combination occurs, and to use this data to inform the feature lookup process after clustering.

[3] Agglomerative clustering using Ward's linkage as implemented in SciPy: https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html

IV. Research Potential

The main idea of this research is to employ an automatic metaphor identification technique - robust enough to process unstructured text from the literary domain as well as from different time periods - in order to gain insight into the development of figurative language and the cognitive and cultural models entailed by it. A number of questions lend themselves to gain from this approach. First of all in literary history, where found metaphorical mappings can act as features for genre identification and authorship attribution. This was already undertaken, although manually, by Lodge [6], who showed that the relationship between metaphor and metonymy can be applied to distinguish between literary genres, movements, authors, texts, as well as parts of texts. It seems natural to revisit this work in light of larger data sets and current analysis methods. Another area of application is the field of critical discourse analysis, media studies, and every other field that is practically concerned with the conditions of knowing (epistemology) and perceiving (aesthetics) and looking to uncover ideological subtexts and collective orientations by means of analyzing media products and works of art. Along the same line, Lakoff [7] did a contrastive analysis of conceptual metaphors in different cultures and arrived at far reaching conclusions regarding the pervasiveness of logical-positivist conceptions and their inherent western bias. Finally, another application and necessary byproduct of this research will be a cross-lingual alignment - at least in parts - between the inventory of conceptual metaphors as described by Lakoff et al. [5] and its application to a German language corpus. After all, metaphor seems to be universal for a number of primary formations, but otherwise culturally determined and ideologically situated [8]. A predefined set of categories might not be fully applicable in other languages or for use in exhaustive computational approaches such as the one described here [1].

V. Conclusions

This paper describes the application of a hierarchical clustering approach to metaphor identification in literary texts. First results are encouraging, showing that significant metaphorical mappings can be found in historical German novels, even using a non-specialized algorithm. However, it has also become clear that the amount of data is crucial for this approach and needs for normalization and balancing of the data have been identified.

VI. Acknowledgment

The author would like to thank everyone at the Department for Literary Computing, University of Würzburg and DARIAH-DE, Digital Research Infrastructure for the Arts and Humanities for the support.

References

[1] E. Shutova, "Design and evaluation of metaphor processing systems," *Computational Linguistics*, 2015.

[2] G. Lakoff and M. Johnson, *Metaphors we live by*. University of Chicago press, 1980.

[3] R. W. Gibbs, "Metaphor and thought. the state of the art," in *The Cambridge Handbook of Metaphor and Thought*, R. W. Gibbs, Ed. Cambridge University Press, 2008, pp. 3–14.

[4] E. Shutova and L. Sun, "Unsupervised metaphor identification using hierarchical graph factorization clustering." in *HLT-NAACL*, 2013, pp. 978–988.

[5] G. Lakoff, J. Espenson, and A. Schwartz, "The master metaphor list," University of California at Berkeley, Tech. Rep., 1991.

[6] D. Lodge, *The modes of modern writing: metaphor, metonymy, and the typology of modern literature*. University of Chicago Press, 1988.

[7] G. Lakoff, *Women, fire, and dangerous things: What categories reveal about the mind*, 1987.

[8] A. Deignan, "Corpus linguistics and metaphor," in *The Cambridge handbook of metaphor and thought*, R. W. Gibbs Jr, Ed., 2008, vol. 280, p. 290.