Adapting Measures of Literality for Use with Historical Language Data

Adam Roussel
Department of Linguistics
Ruhr University Bochum
roussel@linguistics.rub.de

Abstract

This paper concerns the adaptation of two existing computational measures relating to the estimation of the literality of expressions to enable their use in scenarios where data is scarce, as is usually the case with historical language data. Being able to determine an expression's literality via statistical means could support a range of linguistic annotation tasks, such as those relating to metaphor, metonymy, and idiomatic expressions, however making this judgment is especially difficult for modern annotators of historical and ancient texts. Therefore we re-implement these measures using smaller corpora and count-based vectors more suited to these amounts of training data. The adapted measures are evaluated against an existing data set of particle verbs annotated with degrees of literality. The results were inconclusive, yielding low correlations between 0.05 and 0.10 (Spearman's ρ). Further work is needed to determine which measures and types of data correspond to which aspects of literality.

1 Introduction

Though it is usually taken as a given that an expression's 'literal' meaning is readily identifiable, the notion is more ambiguous and complex than it initially seems. As Lakoff (1986) explains, there are several different senses that may apply when we describe something as 'literal'. When an expression is used in a literal sense, we might mean that it is used to mean the thing it *usually* means, whether relative to the broader language community or to a narrower domain-specific lect. Alternatively, we might mean that the meaning of the whole expression corresponds to a systematic combination of its constituent parts, i.e. that the expression's meaning is compositional. Or we might be saying something about what an expression is not doing: It's not an instance of irony, metonymy, metaphor, or any other sort of context-dependent extension of its *minimal* meaning. Yet generally, in linguistic research, we tend to regard these as being equivalent.

It makes a difference which definition you choose and how you regard literality: In Gibbs Jr. et al. (1993) annotators were presented with annotation guidelines highlighting one of Lakoff's definitions of literality and asked to judge the literality of a range of expressions. The result was that some expressions were judged as having significantly different properties, depending on the literality definition in use.

There are a range of semantic annotation tasks which, implicitly or explicitly, require the annotator to make a judgment as to the literality of an expression in a given context: This could be the case in compositionality annotation of multiword expressions (MWEs) or the annotation of idioms or metaphors, which must be distinguished from what they are not: literal usages.

Yet, when annotators are working with historical language data, the ambiguity of these distinct variations of literality, which could otherwise be mitigated with well-written annotation guidelines, is coupled with a lack of linguistic intuition, which is what semantic annotation tasks usually rely on.

In the larger research project in which this paper is situated, it is our aim to develop statistical measures to estimate the degree of literality of particular usages of a given expression in order to support annotators by giving them tools to compensate the lack of linguistic intuitions for historical language varieties. To this end, here we:

- identify existing measures for estimating the degree of literal and non-literal language use,
- adapt the relevant measures for use with historical language data, and
- present the results of a comparison with the original formulation of these measures.

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 209–215
November 16, 2024.

2 Related Work

The theoretical descriptions of literality outlined above highlight multiple aspects which characterize it: The literal meaning is the one that is *conventional* or at least *typical* for that expression, or its the one that is involved when an expression is read compositionally, or its the *minimal* meaning, which an expression is thought to have devoid of any context, its context-free interpretation. Previous approaches have tended to focus either on an expression's compositionality or its conventionality when measuring the degree to which it is literal or non-literal.

Much of the computational work that deals with the measurement of degrees of compositionality and literality focuses on particular classes of MWEs. Most commonly these are nominal compounds, as in Schulte im Walde et al. (2013), Weeds et al. (2017), and Cordeiro et al. (2019), but others address compositionality in MWEs more broadly, as in Salehi et al. (2015) or Savary and Cordeiro (2017). In general, an expression is taken to be *compositional* to a greater degree when the semantics of the whole are more similar to a systematic combination of the expression's parts. The composition function is usually an additive model (Mitchell and Lapata, 2008), which assigns a weight to each vector before adding them together. This approach to compositionality requires a representation for both the expression as a whole as well as for the parts individually, so it tends to work best with a fixed inventory of expressions to be analysed, since you need to know which expressions to combine before training a semantic representation ahead of time. Since our data set doesn't concern a fixed set of expressions, we plan to address the aspect of compositionality in future work.

There are a range of studies which deal with the annotation and classification of figurative language which are also relevant, such as those concerning metaphor, metonymy, irony, and idioms. In all of these cases, 'literal' is defined negatively, as the class not sought, the normal and default assumption. As such, the features that would characterize literality are not modelled directly.

The set of studies that come closest to providing an account of literality itself are those that concern idiomatic expressions. Many of the features discussed in these studies reflect the notion that literal mentions are somehow 'typical' of the expression in question and non-literal mentions are thus 'atypical'. They differ in how this typicality is modelled. Sporleder and Li (2009) use an unsupervised approach that applies the notion of *lexical cohesion*, operationalizing this as the mean similarity of the terms within a particular window. The candidate expression is then removed and the cohesion is calculated without it. When the level of cohesion increases upon removal of the candidate expression, it suggests that this is an atypical context for that expression, and the instance is classified as idiomatic or non-literal. Ehren (2017) describes an extension of this approach, which replaces the "normalized Google distance" of the original, which has a very low degree of reproducibility, with similarities between word vectors. This approach forms the basis for the cohesion measure we employ below.

Socolof et al. (2022) also address the identification – or rather the *characterization* – of idioms: Their study is more of a characterization because they consider idioms to exist on a spectrum together with novel metaphors, collocations, and ordinary literal language. Thus it isn't a distinct class as a classification task would suggest. All of these expressions can be related to one another along two orthogonal axes of conventionality and contingency, where conventionality describes the extent to which words are used in their "usual" or "typical" sense, and contingency refers to the tendency for words to be used in a particular, fixed context. The dimension that broadly differentiates literal and non-literal usages is that of conventionality. The conventionality measure that we adapt in this work stems from this study.

3 Methods

3.1 Evaluation data set

In order to evaluate these measures with regard to how well they reflect our intuitions as to the literality of expressions in general, we compare their output to a modern data set of German particle verbs (grammatical constructions consisting of a verb and a separable particle) annotated with literality ratings on a scale from 0 'literal' to 5 'non-literal' (Köper and Schulte im Walde, 2016). We evaluate against this data set not because of any interest in particle verbs in particular, but because it is the only data set we are aware of that contains scalar ratings of literality rather than a binary classification. The data set consists of German sentences containing particle verbs, as in examples (1)–(3), with roughly 50 sentences for each one. Since some of the particle verbs were not so frequent in the corpus from which the sentences were extracted, some have fewer than 50 sentences, and we omit those instances with less than 5 sentences, since no reasonable comparison between the instances is possible in that case. Thus we have for this study 155 distinct particle verbs, with 6426 sentences total and 41.5 sentences per lemma on average.

The literality ratings for the examples (1)–(3) are also given below. We include a clearly literal instance (1), a clearly non-literal one (2), and one marginal one (3). Each instance was rated by three raters, with high correlation between them, raters 1 and 2 Pearson's $\rho = 0.959$, 2 and 3 $= 0.943$, and 1 and 3 $= 0.932$, though the corresponding agreement appears moderate with Fleiss' $\kappa = 0.35$. In the original study, these ratings were combined into two bins, literal and non-literal, with an agreement of Fleiss' $\kappa = 0.70$ for this classification-oriented setting. In examples (1)–(3) we include the ratings averaged across all three raters.

(1) Dazu untere Äste kräftig **abklopfen** und herabfallende Läuse auf einem Stück Papier oder Karton auffangen.
'To that end **pound** heavily on the lower branches and catch the lice that fall down on a piece of paper or cardboard.'
0 ⇒ literal

(2) Bin ein alter Bücherwurm und hab meine Spezialadressen **abgeklopft**.
'I am an old bookworm and have **checked** (lit. knocked on) my special addresses.'
5 ⇒ non-literal

(3) Kommerzielle Seiten werden nur in Ausnahmefällen **aufgenommen**.
'Commercial pages will only be **taken up** in exceptional cases.'
3.67 ⇒ non-literal

We observe that the ratings are strongly biased towards the extremes of the scale, with 0 and 5 being the most common ratings overall and only very rare instances of 2 or 3, suggesting that the intended use in a classification task was part of the instructions given to annotators.

The data for some lexemes contain very few non-literal instances, and others are more mixed.

3.2 Adapting literality measures

In this study, we adapt two measures: One, Ehren (2017), relates to lexical cohesion, is an embeddings-based version of Sporleder and Li (2009)'s original version, which relied on "normalized Google distances". The second, conventionality (Socolof et al., 2022), compares a single instance of an expression with a set of background instances, measuring the degree to which this instance deviates from the general tendency of the background set.

The original formulations rely on resources that are often available, indeed abundant, for modern languages, large corpora derived from collections of unstructured text scraped from the web. The word2vec vectors used in Ehren (2017) tend to require about 1 billion tokens before they are of usable quality (Sahlgren and Lenci, 2016), and the BERT model used in Socolof et al. (2022) would have been trained on 4 billion words from Wikipedia in addition to other sources.

While there are various strategies to be explored for working with historical language varieties and small data (see, e.g., Hedderich et al., 2021), two of these are the use of **unsupervised** approaches and the use of techniques that require **less data**. These are the two requirements that motivated the choice of cohesion and conventionality, as unsupervised measures, in order to model literality, and these requirements will also act as constraints on the adaptation of the two chosen measures.

Historical corpora are in general much smaller than modern corpora, yet they are often more richly annotated. What are often considered 'expensive' resources for modern languages, such as manually constructed lexica and corpora with rich linguistic annotations, are more attainable than large amounts of text. Crucially, for historical varieties, corpora do not tend to grow: the data there is is what we have. While corpora for language stages after the widespread adoption of the printing press, such as the DTA corpus (Berlin-Brandenburgischen Akademie der Wissenschaften, 2024), spanning the 17th to early 20th centuries, can reach similar sizes to modern data sets – the complete DTA corpus contains 370 million tokens – this is not the case for older data sets. While often the bottleneck is the transcription and digitization of the manuscript sources, in other cases, there are simply few extant manuscripts to be digitized.

As our target historical language variety we consider Middle High German, for which we consider two example corpora: the Reference Corpus of Middle High German (ReM, Roussel et al. (2024)) and the corpus of the Middle High German Conceptual Database (MHDBDB, Zeppezauer-Wachauer (1992)). The former encompasses just over 2 million tokens and the latter just over 9 million. However, our vector representations must be trained on modern German corpora in order to evaluate against the annotated data set and in order to enable a comparison against the original measures using pre-trained models. We therefore simulate the low-data setting of the abovementioned historical corpora by the use of similarly sized modern corpora. As a stand-in for ReM, we use the "dev" and "train" portions of the Hamburg Dependency Treebank (Foth et al., 2014) at about 2 million tokens, and for MHDBDB we use the "2011 mixed" corpus of 1 million sentences from the Leipzig Corpora Collection (Goldhahn et al., 2012), which contains about 7.6 million tokens.

For these amounts of data, neural embeddings do not tend to provide the best results. Sahlgren and Lenci (2016) compared a range of different models of distributional semantics on different sizes of training corpora, and their study suggests that a count-based model transformed using PPMI and SVD could provide the best results with the amounts of data we have available. Such a model has a further advantage in that its operation is more transparent than a prediction-based one. Though, as the authors note, none of the models do particularly well in this setting, so it remains to be seen whether the measures will remain effective with these inputs.

In order to model a word's use in a specific context (tokens), in addition to its distribution in the whole corpus (types), we adopt an approach to modelling specific usage contexts that is described in Geeraerts et al. (2023) and which ultimately goes back to Schütze (1998). Type vectors are constructed from word co-occurrences in the entire corpus (transformed with PPMI/SVD), then a token vector is constructed by adding together the type vectors for all of the context words that occur in a certain window around the target token to be represented. In effect, a token is represented as a set of second-order co-occurrences: Two tokens are similar when they co-occur with words that co-occur especially often.

We then implement cohesion and conventionality using either type vectors or token vectors as required. **Cohesion** (cf. Ehren, 2017) is defined here as follows: For each token instance w_i, we compare the type vector for w with the type vectors for all the words in the context of w_i, calculating the mean similarity between all pairs of these vectors both including and excluding the target expression w. If the mean similarity is greater without w, then this reflects lower lexical cohesion, and we expect it to correlate with less literality for this usage. Our adaptation of this measure differs from the original mainly in the embeddings used.

We take **conventionality** (cf. Socolof et al., 2022) to be defined as follows: A given word has a set of instances W, and conventionality is calculated for a single instance w_i by comparison with the other instances of this word $O = W \setminus w_i$. μ_O is the average token vector of the instances in O, and σ_O the component-wise standard deviation for these same instances. The conventionality is then:

$$\text{conv}(w_i) = \left\| \frac{w_i - \mu_O}{\sigma_O} \right\|_2 \tag{4}$$

This differs from the original formulation in that the original calculated the deviation of a particular phrase in which a word occurs versus all the other phrases in which the word also occurs, but we take a simplified approach. Since, in our low-data setting, the word in question is unlikely to have the same context more than once, we compare each instance against all of the other 49 instances for each lemma. Note also that the sign is reversed, since the scale in our evaluation data set uses higher numbers for less literal usages.[1]

4 Experimental Results

For the 155 lemmas in the annotated data set, we calculated cohesion and conventionality with each of the three implementations for each of the ≈ 41.5 annotated instances, resulting in 38,556 instances total. Of these, 1357 instances were omitted, either because the lemmas do not occur in the background corpus or because there wasn't sufficient context in the test sentence. For the remaining 37,199 combinations of sentence, lemma, measure, and implementation, we averaged the three raters' judgments together in order to compare them with the given value.

For the sake of comparison, we also evaluate each of these measures in a setting as close as pos-

[1] All of the code pertaining to these experiments is provided here under a free software license: https://gitlab.rub.de/ajroussel/nlp4dh2024.

sible to the original papers (*_orig). For cohesion, this means using pre-trained skip-gram word2vec vectors for the type vectors, and for conventionality, we use a pre-trained German BERT model to encode each sentence, from which we retrieve a contextualized representation for each target token instance.

The results of this comparison can be found in Table 1, and a visualization of the per-lemma correlations with average human judgments in Figure 1.

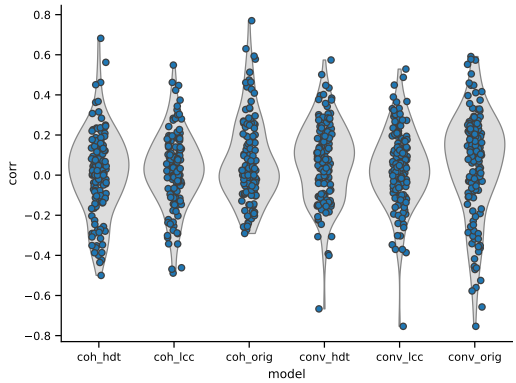

Figure 1: Correlations (Spearman's ρ) between models and human judgments. Each dot represents the correlation for a particular lemma.

Table 1: Overall correlations for each measure and implementation. Impl. = Implementation, Desc. = Description, Coh. = Coherence, Conv. = Conventionality. "hdt" or "lcc" indicates the corpus used to create the vectors, and "orig" are the pre-trained vector representations.

Impl.	Desc.	Coh.	Conv.
hdt	tSVD, 2M	0.072^{***}	0.073^{***}
lcc	tSVD, 10M	0.050^{***}	0.052^{***}
orig	w2v/BERT	0.098^{***}	0.010, ns

5 Discussion and Conclusion

Comparing the values of the two measures in their various implementations to the human judgments in the literal/non-literal data set does not appear to reveal any reliable patterns. Though in some cases the correlations are technically significant, the actual level of correlation is too low for either measure to be trusted in any particular case. The results suggest that these measures, implemented as described above, don't correspond, in general, to the notion of literality that the annotators had in mind. As a result, it is also impossible to say whether the adaptations of the measure for use with smaller corpora, such as for historical language varieties, were appropriate or whether they had any effect on the usefulness of the measures implemented.

An analysis of the correlations of the various measures with the judges' ratings on a per-lemma basis was likewise inconclusive. As is evident in Figure 1, the correlations for particular lemmas can vary quite widely between strongly negative and positive correlations. We haven't been able to find a clear reason for this; there are no apparent tendencies towards higher correlations when a lemma has a greater proportion of non-literal instances, for instance.

In future work, we plan to conduct more extensive annotation efforts specifically targeting literality in order to collect more fine-grained data to use in future experiments. Ideally, such a data set will cover not just particle verbs, but all open-class lexemes, and we plan to formulate detailed guidelines that will improve reproducibility and reusability of the data set.

6 Ethical considerations

Embeddings trained on corpus data scraped from the web, such as are employed in the comparison here, are known to have certain biases that could have had an effect on the outcomes of this study.

7 Limitations

We have characterized both of these measures as ones of 'literality' in general, but it's still unclear to what degree (a) each of these individually or in combination correspond to a recognizable and coherent concept of literality, and (b) whether the conception of literality captured in the annotated data set corresponds to the aspects of literality that the measures relate to, or whether any apparent correlation is spurious. This study in its current form isn't in a position to address these questions.

Acknowledgments

Thanks to the reviewers for their helpful feedback. Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – SFB 1475 – Projektnummer 441126958.

References

Berlin-Brandenburgischen Akademie der Wissenschaften. 2024. Deutsches Textarchiv: Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. https://www.deutschestextarchiv.de/. Accessed 2024-08-28.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

Rafael Ehren. 2017. Literal or idiomatic? identifying the reading of single occurrences of German multiword expressions using word embeddings. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 103–112, Valencia, Spain. Association for Computational Linguistics.

Kilian A. Foth, Arne Köhn, Niels Beuck, and Wolfgang Menzel. 2014. Because size does matter: The Hamburg Dependency Treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2326–2333, Reykjavik, Iceland. European Language Resources Association (ELRA).

Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang. 2023. *Lexical Variation and Change: A Distributional Semantic Approach*. Oxford University Press.

Raymond W. Gibbs Jr., Darin L. Buchalter, Jessica F. Moise, and William T. Farrar IV. 1993. Literal meaning and figurative language. *Discourse Processes*, 16(4):387–403.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Maximilian Köper and Sabine Schulte im Walde. 2016. Distinguishing literal and non-literal usage of German particle verbs. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 353–362, San Diego, California. Association for Computational Linguistics.

George Lakoff. 1986. The meanings of *literal*. *Metaphor and Symbolic Activity*.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio. Association for Computational Linguistics.

Adam Roussel, Thomas Klein, Stefanie Dipper, Klaus-Peter Wegera, and Claudia Wich-Reif. 2024. Referenzkorpus Mittelhochdeutsch (1050–1350). ISLRN 937-948-254-174-0.

Magnus Sahlgren and Alessandro Lenci. 2016. The effects of data size and frequency range on distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 975–980, Austin, Texas. Association for Computational Linguistics.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983, Denver, Colorado. Association for Computational Linguistics.

Agata Savary and Silvio Ricardo Cordeiro. 2017. Literal readings of multiword expressions: as scarce as hen's teeth. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 64–72, Prague, Czech Republic.

Sabine Schulte im Walde, Stefan Müller, and Stefan Roller. 2013. Exploring vector space models to predict the compositionality of German noun–noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 255–265, Atlanta, Georgia, USA. Association for Computational Linguistics.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Michaela Socolof, Jackie Cheung, Michael Wagner, and Timothy O'Donnell. 2022. Characterizing idioms: Conventionality and contingency. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4024–4037, Dublin, Ireland. Association for Computational Linguistics.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762, Athens, Greece. Association for Computational Linguistics.

Julie Weeds, Thomas Kober, Jeremy Reffin, and David Weir. 2017. When a red herring in not a red herring: Using compositional methods to detect non-compositional phrases. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short*

Papers, pages 529–534, Valencia, Spain. Association for Computational Linguistics.

Katharina Zeppezauer-Wachauer. 1992. Mittelhochdeutsche Begriffsdatenbank (MHDBDB). https://mhdbdb.plus.ac.at/. Universität Salzburg, accessed 2024-08-27.