The distance between Middle High German and modern German: A new approach about Fuzzy Search

Yuwei Nan
Freie Universität Berlin

Abstract

This paper presents a novel approach to quantifying the lexical similarity between Middle High German (MHD) and Modern German. By introducing a novel similarity metric, we aim to provide a more accurate assessment of the readability of MHD texts for contemporary German speakers. Our research analyzes dialect variations within MHD and the genre of the texts, contributing to a deeper understanding of the challenges and potential rewards of engaging with historical language.

1 Introduction

The increasing accessibility of historical texts has renewed interest in understanding Middle High German (MHD). However, the lexical gap between MHD and Modern German poses significant challenges for contemporary readers. This paper introduces a novel approach to quantifying lexical similarity, addressing the limitations of traditional methods. By considering dialect variations within MHD and the genre of the texts, our research aims to provide a more comprehensive understanding of the readability of MHD texts.

2 About the Corpus

This study utilizes a carefully compiled corpus of Middle High German texts from the 11th to 14th centuries, serving as the primary material for analysis. The corpus comprises approximately 100 articles, each with an average length of 1,630 lines and an average word count of 10,297. It was constructed following a detailed methodology to ensure both representativeness and accuracy. Key aspects of this methodology include the following:
Manuscript Fidelity: Original manuscripts or high-quality facsimiles were used to preserve the linguistic authenticity of the texts. Structured Organization: The texts were categorized into specific time periods (1050-1150, 1150-1200, 1200-1250, 1250-1300, and 1300-1350) and dialectal regions (Bavarian, Bavarian-Alemannic, Alemannic, Middle Franconian, Rhine Franconian-Hessian, East Middle German, and East Franconian).
Textual Diversity: The corpus includes a variety of text types, such as verse, prose, and legal documents, ensuring a comprehensive representation of the linguistic richness of the era.
To maintain relevance and adaptability for ongoing research, a dynamic approach allows for the incorporation of new texts as they become available. This ensures the corpus remains up-to-date and continues to support diverse analytical needs.

3 Methodology

For individuals unfamiliar with Middle High German (MHD), they often attempt to read MHD words as though they were modern German, disregarding grammatical context. Their approach typically involves deciphering each word individually, without regard to word order, and then piecing together a sentence from the resulting fragments. This intuitive, approximate method resembles what is known in database systems as "fuzzy search," which is why I refer to this process as such. This „fuzzy search“ method is a localized adaptation of the approach described in the paper "Measuring Language Divergence by Intra-Lexical Comparison."Ellison and Kirby, 2006
However, since we lack data on word pronunci-

ation and instead have direct access to a corpus, our comparison focuses solely on the Hamming distance between words.

In the "fuzzy search" approach, readers tend to identify modern German words that have the smallest Hamming distance from the MHD words they are trying to guess. Among words with the same Hamming distance, they often favor the word form that most closely resembles the original MHD word. For example, when encountering the MHD word „unte", they are more likely to think of "unter" or "unten" rather than "Ente," even though "Ente" has the same Hamming distance from "unte." Our research aims to measure the success rate of this "fuzzy search" process. Specifically, we seek to answer: how successful is this method across different genres, time periods, or dialects? Which factors lead to the highest or lowest rates of success?

4 Implementation

The implementation process consists of several key steps, beginning with extracting the lemma form of Middle High German (MHD) texts from a html-based corpus. This extraction focuses on isolating words while ignoring grammar and other contextual markers, and tags each text according to its genre and time period. For each document, word frequencies are counted to analyze usage patterns.

4.1 sec:Data Cleaning

The first step in data cleaning involves removing arrows and other symbols used for grammatical annotation in the original corpus. This ensures that only the lemma themselves are retained. The MHD text is then duplicated, producing a clean version for comparison with modern German. For instance, the MHD word "zîtic" contains the character "î," which is unavailable in modern German and unrecognized by dictionaries like Langenscheidt. Therefore, "zîtic" is modified to "zitic" before being sent for further dictionary searches. This mirrors how a typical reader might mentally adapt unfamiliar characters to their modern counterparts.

4.2 Duplicate Cleaning

The duplicate cleaning step follows a principle: if the core part of a word is unrecognizable or unreadable, then the entire word is considered unreadable. This also applies to new words formed by adding prefixes or suffixes. For example, "zîtic" in MHD translates to "zeitig" in modern German. If a reader cannot recognize "zîtic," they will likely struggle with "unzîtic." However, once "zîtic" is understood, recognizing "unzîtic" becomes easier.

4.3 Manual Correction

Due to the complexity of MHD dialects, manual corrections are required to account for regional variations. For instance, "zîtic" may appear as "zîtec" or "zîtig" depending on the dialect, but only "zîtic" is listed in standard dictionaries. In cases where the corpus uses "zîtig," manual intervention aligns this variant with dictionary entries.

4.4 Word Meaning Verification

A web crawler is employed to verify word meanings in both MHD and modern German. MHD words are checked against the Wörterbuchnetz dictionaryLexer, 2023, while the cleaned modern German versions are checked using Langenscheidt. The results from both dictionaries are compared. If the MHD word and its modern German counterpart have the same meaning, the word is considered "directly readable." If there is a significant difference in meaning, it is classified as "not directly readable." Words that cannot be found in either dictionary, even after manual cleaning, are marked as "unreadable" and excluded from further analysis. For instance, the web crawler would send "zitic" to Langenscheidt and "zîtic" to the MHD dictionary, saving both results for comparison.

4.5 Fuzzy Search Success Rate Calculation

The fuzzy search success rate is calculated based on the frequency of words in the corpus. This analysis determines how successfully MHD words can be interpreted using modern German, providing insights into the readability of historical texts across different genres, time periods, and dialects. For example, if Langenscheidt recognizes "zitic" and provides a meaning similar to the MHD dictionary entry for "zîtic" (such as "zeitig"), this would be considered a successful fuzzy search attempt.

$$\text{Fuzzy Search Success Rate} = \frac{\text{Successfully Found Words}}{\text{Total Words} - \text{Unreadable Words}}$$

5 Results

The fuzzy search success rate for all words in the corpus falls within the range of 0.76 to 0.58, indicating a relatively narrow distribution exerpt a few corpus. This suggests that there are no significant differences in fuzzy search performance across the corpus in general. The text with the highest success rate is *Die religiösen Dichtungen des 11. und 12. Jahrhunderts* in the Alemannic dialect, while the text with the lowest success rate is *Williram von Ebersberg: 'Hoheliedkommentar* from the 14th century. Interestingly, neither the time period (epoch) nor the dialect plays a dominant role in determining the fuzzy search success rate. Instead, the genre of the text shows a more noticeable impact. Religious texts consistently achieve higher fuzzy search success rates compared to other genres.

Table 1 shows the fuzzy search success rate for each corpus document. The document's name are only related to their epoch and language, but not about genre, but this information could be found in the htm file.

6 Interpretation

The higher success rate observed in religious texts can be attributed to their frequent use of words with Latin origins, which tend to remain stable over time. Such words, due to their roots in ecclesiastical and scholarly tradition, have undergone fewer changes in both form and meaning. In contrast, everyday words—those used in non-religious contexts—tend to evolve more rapidly, leading to lower success rates in the fuzzy search process. This linguistic stability in religious texts likely explains their greater readability when applying fuzzy search techniques, as compared to legal or administrative documents, where word usage may vary significantly across epochs.

This study presents several areas for improvement.

First, just because a word is found in the Langenscheidt dictionary does not necessarily mean that a modern German speaker would immediately recognize or understand it. For instance, the word "töricht" means "irrsinnig" (foolish), but without a dictionary at hand, a speaker might interpret it incorrectly, perhaps reading "töricht" as "tor-rich" and associating it with the word "Tor" (goal), imagining a football team scoring in a recent match. This demonstrates that lexical recognition alone is not always sufficient for comprehension.

Second, the study does not account for the varying importance, or "weight," of individual words within a sentence. Recognizing simple words like "in" or "uf" (modern "auf") might not provide meaningful understanding if critical components like the main verb are unfamiliar. If a reader fails to grasp the core elements of a sentence, such as the subject or verb, the entire sentence becomes incomprehensible, rendering the fuzzy search less useful.

Third, words that could not be found in the MHD dictionary were disregarded in the final calculations, which may introduce significant issues. Ignoring these words does not eliminate their presence in the text, and their meanings remain elusive. The inability to account for these words could distort the overall success rate, as readers are still left without a full understanding of the text.

Moreover, the manual corrections made during the implementation phase might have also influenced the results. While some words were corrected to facilitate the search process, the researcher's familiarity with Middle High German (MHD) may have unintentionally biased the study. A tester with no prior knowledge of MHD would likely find these texts far more unfamiliar and challenging than the researcher, who can read and interpret MHD. As a result, the text may appear less comprehensible to others than it does to the researcher.

7 Furture Work

Future research could address these limitations by using a broader range of MHD dictionaries to improve word recognition and developing a more nuanced method of assigning weights to individual words. For instance, scoring each sentence based on whether key components like the main verb or noun are understood could offer a more accurate reflection of the overall comprehension level. If the core elements of a sentence are not recognized, the sentence could be marked as "unintelligible," which would provide a more precise measure of the fuzzy search success rate. Additionally, incorporating a wider range of test subjects who have no prior exposure to MHD could yield results that better represent the experience of the general public when interpreting these texts.

8 Code and Github address

This is the code space for this paper: https://github.com/Halphas-Kreuz/MHD-Analyse
The htm folder contains the original corpus data in html form, which are all from the Müller and Klein, 2020
The cleanedLine folder contains the data after Data cleaning process. The program which I used are all stored in local_file_processing folder. The scrappy folder contain the web crawler code in python, while the wordCount folder contains the word and frequency data for each corpus file.

References

T Mark Ellison and Simon Kirby. 2006. Measuring language divergence by intra-lexical comparison. In *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics*, pages 273–280.

Matthias Lexer. 2023. Mittelhochdeutsches handwörterbuch. Digitalisierte Fassung im Wörterbuchnetz des Trier Center for Digital Humanities, Version 01/23. Abgerufen am 20.09.2024.

Stefan Müller and Thomas Klein. 2020. Mittelhochdeutsch-korpus. https://daten.badw.de/mhd-korpus/-/tree/auflage$_2$020. $ErarbeitetvomDFG-ProjektMittelhochdeutscheGrammatik.$

Table 1: Fuzzy Search Success Rate by Text File

File Name, Ratio	File Name, Ratio
12alem1.txt, 0.75989	12alem2.txt, 0.74879
13hess2.txt, 0.73639	12alem6.txt, 0.73366
12bair3.txt, 0.73361	14bair4.txt, 0.73360
14bair2.txt, 0.73341	13bair7.txt, 0.73306
13alem.txt, 0.72975	13bair11.txt, 0.72685
12alem3.txt, 0.72677	13frank1.txt, 0.72643
13bair2.txt, 0.72619	14frank2.txt, 0.72579
13alem2.txt, 0.72395	14ripu3.txt, 0.72238
13frank2.txt, 0.72142	14alem4.txt, 0.71613
12frank1.txt, 0.71569	13alem12.txt, 0.71565
12bair4.txt, 0.71477	13frank5.txt, 0.71144
14bair1.txt, 0.71104	13hess4.txt, 0.71063
14hess3.txt, 0.70981	13alem5.txt, 0.70922
13hess5.txt, 0.70907	13bair1.txt, 0.70891
14bair10.txt, 0.70838	14frank11.txt, 0.70792
14alem3.txt, 0.70791	12bair5.txt, 0.70787
13bair3.txt, 0.70604	12bair7.txt, 0.70601
13alem11.txt, 0.70589	14schw2.txt, 0.70560
14frank6.txt, 0.70270	13bair9.txt, 0.70257
12alem5.txt, 0.70245	14MD2.txt, 0.70216
13alem1.txt, 0.70203	13ripu6.txt, 0.70160
13ripu5.txt, 0.70023	14hess4.txt, 0.69985
13alem6.txt, 0.69972	14frank10.txt, 0.69919
13hess1.txt, 0.69840	14alem1.txt, 0.69766
13alem8.txt, 0.69665	14MD1.txt, 0.69656
13bair6.txt, 0.69581	14bair8.txt, 0.69533
14bair6.txt, 0.69511	12bair2.txt, 0.69502
13hess6.txt, 0.69443	13ripu2.txt, 0.69279
14MD4.txt, 0.69274	12bair9.txt, 0.69176
14frank1.txt, 0.69014	14MD3.txt, 0.69010
14bair3.txt, 0.68924	13thur1.txt, 0.68891
14bair5.txt, 0.68796	13ripu1.txt, 0.68767
14frank3.txt, 0.68762	14ripu1.txt, 0.68674
12frank2.txt, 0.68570	14hess2.txt, 0.68467
13alem4.txt, 0.68380	13alem7.txt, 0.68316
13hess3.txt, 0.68310	13frank3.txt, 0.68243
13bair10.txt, 0.68241	13alem3.txt, 0.68240
14frank5.txt, 0.68133	13frank4.txt, 0.68061
13bair5.txt, 0.67941	14hess1.txt, 0.67756
13bair4.txt, 0.67271	13bair12.txt, 0.67117
13ripu3.txt, 0.67078	13bair8.txt, 0.66911
14bair7.txt, 0.66891	14frank4.txt, 0.66734
14frank8.txt, 0.66541	14frank7.txt, 0.66516
13ripu4.txt, 0.66393	14alem2.txt, 0.66301
12bair6.txt, 0.66241	14hess5.txt, 0.66192
13alem10.txt, 0.66109	12alem4.txt, 0.65703
13alem9.txt, 0.65578	14bair9.txt, 0.65525
14schw1.txt, 0.65081	13thur2.txt, 0.64396
12bair1.txt, 0.63801	11bair.txt, 0.62735
11frank.txt, 0.58175	