



Universiteit
Leiden

Master Computer Science

Analyzing language complexity and medical terminology use in Dutch medical consultations

Name: Annelotte Baarslag
Student ID: s2419874
Date: 23/06/2026
Specialisation: Data Science
1st supervisor: Suzan Verberne
2nd supervisor: Arwen Pieterse

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

This thesis investigates the differences in language between doctors and cancer patients in Dutch medical consultations. We use text difficulty metrics to analyze differences in language complexity, and biomedical entity linking to analyze differences in the use of medical terminology. Differences in language between the doctor and the patient make it more difficult for the patient to understand the information, which can make it harder to reach shared decision-making.

To investigate these differences, we analyze three datasets of transcribed Dutch cancer consultations. We first use a classifier to distinguish doctor and patient turns and investigate which words mainly contribute to this difference. We then use text difficulty metrics, such as the FRE score, to estimate differences in language complexity. Lastly, we use named entity recognition followed by biomedical entity linking to analyze the use of medical terminology by both the doctor and the patient.

We find that doctor and patient turns can be distinguished by a classifier, but this difference is mainly due to conversational roles, for example, the use of different subject pronouns. Across all datasets, doctors used more complex language and medical terminology than patients, and used many words that patients did not use. This shows that doctors potentially make it difficult for patients to understand the information provided. These findings suggest that there are substantial differences in language complexity and medical terminology between doctors and patients, and these differences can be quantitatively analyzed using NLP methods.

Contents

1	Introduction	5
2	Related work	7
2.1	Health literacy and shared decision making	7
2.2	Oral literacy and doctor–patient consultations	7
2.3	Dutch biomedical entity linking	8
3	Data	10
3.1	IBIS and IBIS-colon	10
3.1.1	Consult data	10
3.1.2	Patient data IBIS	10
3.2	ABEL	10
3.2.1	Consult data	10
3.2.2	Patient data ABEL	11
4	Methods	12
4.1	Preprocessing	12
4.1.1	IBIS and IBIS-colon	12
4.1.2	ABEL	13
4.2	Distinguishing doctor and patient turns with a classifier	13
4.3	Dutch Biomedical Entity Linking	14
4.3.1	Replicating paper	14
4.3.2	Named Entity Recognition	15
4.3.3	Expanding the WALVIS corpus	16
4.3.4	Validation set-up	17
4.3.5	Evaluating final performance	17
4.4	Language complexity	18
4.4.1	General language complexity	18
4.4.2	Correlation with patient education level	18
4.4.3	Language complexity adaptation	19
4.4.4	Dialogue pace	19
4.5	Medical terminology use	20
4.5.1	Amount of medical terminology used	20
4.5.2	Most used medical terms, categories, medical terminology differences, and differences in mentions	21
4.5.3	Correlation language complexity and medical terminology use	21
5	Experiments and results	22
5.1	Distinguishing doctor and patient turns with a classifier (RQ1)	22
5.2	Dutch Biomedical Entity Linking (RQ2)	23
5.2.1	Validation	23
5.2.2	Evaluation	24
5.3	Consult complexity (RQ3)	25
5.3.1	General language complexity	25
5.3.2	Correlation text difficulty metrics with patient education level	25
5.3.3	Language complexity adaptation	26

5.3.4	Dialogue pace	26
5.4	Medical terminology use (RQ4)	27
5.4.1	Amount of medical terminology used	27
5.4.2	Most used medical terminology	27
5.4.3	Most used categories	30
5.4.4	Medical terminology use differences	31
5.4.5	Mention differences	33
5.4.6	Correlation language complexity and medical terminology	33
6	Discussion	35
6.1	NLP perspective on evaluation of models	35
6.2	Differences in doctor and patient language	36
7	Conclusion	40
	Appendices	42
A	Excluded UMLS categories	42
B	Top features classifier	43

1 Introduction

Patients with cancer have to make difficult decisions in their treatment, for example between different treatment regimens such as surgery or chemotherapy. These decisions are not purely medical, but also depend on the preferences of the patient. Patient involvement in treatment is known as shared decision-making (SDM), which improves health outcomes and promotes equity [1].

Effective shared decision making remains challenging in practice [2]. One barrier is the mismatch in language use between clinicians and patients [3]. Communication problems between a clinician and their patient can lead to reduced patient autonomy and misunderstandings between the doctor and patient. Therefore, quantifying differences in language use between doctors and patients is important for improving clinical communication and designing more effective SDM. This paper investigates the differences in patient and doctor language in Dutch cancer consultations.

One of the aspects of the language difference between doctor and patient is the difference in language complexity, that is the lexical difficulty of their language. This complexity is often measured in written text using difficulty metrics such as sentence and word lengths, and readability scores such as the Flesch Reading Ease (FRE) score [4]. A second aspect of the language difference is the difference in the use of medical terms. In literature, this is often referred to as the use of medical jargon, medical terminology that non-clinicians may not understand. Doctors often use a lot of medical jargon, which can come across as cold and be more difficult to understand [5]. Especially for patients with low health literacy, medical terminology is difficult to understand, which can lead to worse health outcomes [6].

A method for identifying differences in medical language use is to compare the words that patients and doctors use for the same concept. This can be achieved using biomedical entity linking (BEL), the task of linking mentions of biomedical entities in free text to their corresponding canonical forms in a knowledge base. Hartendorp et al. [7] proposed the first evaluated biomedical entity linking model for the Dutch language, which can be used to analyze Dutch medical consultations.

In this paper, we use three datasets that contain Dutch consultations between doctors and cancer patients to analyze the differences between doctor and patient language. First, we train a classifier to distinguish between doctor and patient turns to assess whether language differences enable it to classify whether a text belongs to a doctor or a patient. Second, we use different text difficulty metrics to estimate differences in language complexity between the doctor and the patient. Third, we use Dutch biomedical entity linking to analyze differences in the use of medical terminology between doctors and patients.

The research question that is being answered in this thesis is as follows: **To what extent do doctors and patients differ in their language use in Dutch cancer consultations and how well can we measure these differences using text difficulty metrics and Dutch biomedical entity linking?** To address this question, the following sub-questions will be answered:

1. How well can a classifier distinguish doctor and patient turns, and which words mainly contribute to this difference?
2. What is the quality of a MedRoBERTa.nl model trained on weakly supervised Wikipedia data when applied to transcripts of Dutch medical consultations?

3. What are the differences in language complexity between doctors and patients, estimated with existing text difficulty metrics?
4. What are the similarities and differences in the use of medical terminology between doctors and patients, and how is the use of medical terminology related to language complexity?

The main contributions are as follows:

- We replicate the method for Dutch biomedical entity linking by Hartendorp et al. [7].
- We develop and publish a complete pipeline for extracting biomedical entities from Dutch clinical consultations using named entity recognition and biomedical entity linking.
- We provide a method for automatically analyzing the language complexity of the doctor and the patient in transcribed consultations using a classifier and text difficulty metrics.
- We analyze similarities and differences in the use of medical terminology between doctors and patients.
- Our analysis confirms earlier research stating that doctors use more complex language and more medical terminology than patients.

2 Related work

2.1 Health literacy and shared decision making

Health literacy (HL) is the capacity of a person to understand and use health information to make health decisions [8]. In 2015, the European Health Literacy Survey measured health literacy across eight European countries [9]. 47% of respondents had limited HL, indicating difficulty finding, understanding, and applying health information. In the Netherlands, 28.7% of the population had limited HL. A gap in health literacy between clinicians and patients can result in poor communication and health outcomes [6]. Patients with lower health literacy have difficulty understanding instructions from their doctor, completing medical forms, and determining whether they need to follow up on their visit [10]. Low health literacy is also associated with increased mortality and hospital readmission and decreased quality of life in patients with heart diseases [11], and health literacy seems to have a significant impact on the care of surgical patients [12]. Patients with lower health literacy can also have difficulty processing oral communication [13]. SDM is difficult to achieve for people with low HL, and clinicians should try to adapt how they present the information to the patient in order to reach SDM [14].

2.2 Oral literacy and doctor–patient consultations

Although much of the research on HL has focused on written text, a substantial part of communication between doctors and patients occurs in clinical consultations. Oral literacy is a part of health literacy that concerns the understanding and use of health information that is communicated verbally [15]. Oral communication between doctors and patients affects many factors, including patient engagement, motivation, and health.

Many articles studied oral literacy by clinicians in healthcare encounters, which were recently analyzed in a scoping review by Yen et al. [4]. Oral literacy is assessed mainly by analyzing transcripts of spoken consultations.

The following four factors are often studied when researching the use of spoken language by doctors:

- Use of medical terminology/jargon, which is often defined as medical words that would not likely be understood by someone who is not a clinician [16, 17, 18].
- Language complexity, which is estimated by text difficulty metrics such as the Flesch-Kincaid reading ease score (FRE), Flesch-Kincaid reading grade level (FKRL), SMOG, or text components such as words per sentence and percentage of passive sentences [19, 16, 20, 21, 22].
- Dialogue characteristics, such as the turn length and turn density [23, 24, 22].
- Relation between the oral literacy of the clinician and reported outcomes of patients or care-givers, such as the satisfaction of the patient with the doctor or the visit [24].

Although frameworks have been developed to measure oral literacy [25, 26], they do not specify how the different aspects of oral literacy should be measured. It is also noticeable that many studies focus only on the language complexity of the clinician and not that of the patient, or they only compare one of the aspects of oral literacy. Meza et al. [27] researched

the differences in language complexity by calculating the FKRL and FRE score in simulated English consultations, and they found that physicians spoke on average at an Eighth-Grade reading level, while patients spoke at a Fourth-Grade reading level. Koch-Weser et al. [28] compared the jargon use between doctors and patients and concluded that the use of medical jargon was more common among doctors than patients. Tran et al. [29] analyzed transcripts of discussions of the prospect of a surgery between physician and patient by manually labeling the medical jargon words from the transcripts. They found that, on average, physicians used more jargon, larger words, and longer sentences than patients.

In most papers, medical terminology is measured using a manually compiled list. This leads to differences in the results, because the list of medical terminology is different per research, and makes measuring medical terminology very time-consuming. There is limited work on the automated analysis of language complexity and the use of medical jargon in Dutch medical texts and clinical consultations. Some papers have been published about analyzing Dutch Electronic Health Records (EHRs). Afzal et al. [30] developed an algorithm, ContextD, a Dutch version of the ConText algorithm, to extract and identify contextual properties of medical terms, such as negation. Verkijk and Vossen [31] developed MedRoBERTa.nl, which is a language model that can be used to extract symptoms from Dutch clinical texts [32, 33]. Kim et al. [34] fine-tuned MedRoBERTa.nl to assign function levels (such as emotional status) to the EHRs of COVID-19 patients. However, these investigations are all about written text and do not compare the doctor and patient. To the best of our knowledge, there is no prior work on automatically analyzing the complexity of language and/or the use of medical terminology in Dutch medical consultations by doctors and patients.

2.3 Dutch biomedical entity linking

Biomedical entity linking (BEL) is the task of linking mentions of biomedical entities in text to normalized, unique identifiers within an ontology [35]. For example, in a text, someone can write “cystitis” or “blaasontsteking”, which are different words for the medical concept of bladder infection, both of which are linked in the UMLS to the CUI C0010692. BEL is an important task in medical natural language processing (NLP) because it helps extract useful medical information and provides context for other tasks, such as relationship extraction. BEL roughly consists of four steps: preprocessing, mention representation, candidate generation, and concept prediction. BEL has diverse challenges. One of them is the complex jargon in biomedical texts, which can also differ between biomedical sub-specialties. Another challenge is the creation of training data for BEL, which requires a lot of time and domain-specific knowledge [36].

Several BEL techniques have been developed. DNorm [37] was among the first papers to apply machine learning to BEL by training a linear classifier on TF-IDF representations of mentions and concepts. Later, deep learning techniques were also used for BEL. First, architectures such as CNN were used, employing TF-IDF and word embeddings for tokenization. Later, these embeddings were replaced by deep contextualized embeddings, such as ELMo [38] and BERT [39], mainly BERT encoders pre-trained on biomedical text, such as BioBERT [40].

More recent BEL models often finetune BERT encoders specifically for entity linking. BioSyn [41] uses BioBERT for learning biomedical entity representations only based on the synonyms of entities. SapBERT [42] further improved biomedical entity linking by using self-alignment pretraining on UMLS synonym pairs.

Recent work has also focused on using contextual information for entity linking, as the meaning of biomedical mentions often depends on the context. KRISSBERT [43] uses knowledge-rich self-supervision by using domain knowledge, and BioLinkerAI uses LLMs to capture knowledge for enhancing BEL.

Besides improvements in the BEL techniques, there also have been a large increase in the amount of BEL corpora. These datasets differ mainly in their text domain, the ontology used for linking, and the types of entities they include. The most common source is scientific literature, with many corpora covering a wide range of biomedical concepts linked to Medical Subject Headings (MESH)¹ or the Unified Medical Language System (UMLS)². Clinical datasets usually target practically relevant entities such as disorders, tests, and treatments, which are mapped to the UMLS or International Classification of Diseases (ICD) codes [35].

The performance of BEL models on non-English mentions is significantly lower than on English mentions. This is because ontologies such as the UMLS primarily cover English terminology, and annotated datasets in other languages are limited. Only a small number of publicly available tools exist for Dutch biomedical entity linking. Hartendorp et al. [7] proposed a method for automatically generating a weakly labeled BEL-dataset for each language by combining UMLS, Wikidata, and Wikipedia, and they introduced the first evaluated BEL model trained on Dutch. They used MedRoBERTa.nl as a base model, performed second-phase pretraining through self-alignment on a Dutch biomedical ontology, and then finetuned on the weakly labeled BEL-dataset. They evaluated their model on the Dutch portion of the Mantra GSC-corpus and achieved 54.7 % classification accuracy and 69.8% 1-distance accuracy. In this thesis, we contribute to the literature by first implementing and reproducing the method of Hartendorp et al., then expanding it, and then applying the BEL model and text difficulty metrics to Dutch medical consultations.

¹<https://www.nlm.nih.gov/mesh/meshhome.html>

²<https://www.nlm.nih.gov/research/umls/index.html>

3 Data

We used three different datasets, IBIS, IBIS colon, and ABEL, which all contain transcribed consultations of doctors discussing adjuvant treatment with a cancer patient and sometimes their companion.

3.1 IBIS and IBIS-colon

3.1.1 Consult data

The IBIS dataset contains transcripts of consultations of patients with breast cancer who have had surgery before and discuss with an oncologist whether they will undergo additional treatment, and if yes, what kind of treatment (chemotherapy and/ or hormone therapy). The dataset consists of MS Word (doc) files of consultations between 191 different patients and their doctors, one file per consultation. For the 12 patients with multiple consultations, the files are put in a folder per patient. The IBIS-colon dataset also contains transcripts of consultations to discuss whether or not the patient will undergo adjuvant treatment; however, these are patients with colon cancer instead of breast cancer. The dataset consists of MS Word (doc) files of consultations between 22 patients and their doctors, one file per consultation. In this dataset, there are no patients with multiple consultations.

In both datasets, all files are structured in the same way. At the beginning of the file, the consultation code, duration of the consultation, and the ID of the specialist are denoted. Each turn of the doctor starts with either "A:" or "A;", each turn of the patient starts with either "P:" or "P;". In some consultations, the patient is accompanied by someone whose turns begin with either "Pm:" or "Pm;". In-line mentions of noise, such as the patient laughing, are denoted in round brackets, and comments on actions, such as the doctor typing something or a privacy-related topic, are denoted in square brackets.

3.1.2 Patient data IBIS

For 49 of the IBIS consultations, we have information about the patient, saved in a .sav file. The dataset contains 52 columns, which provide information about the patient and the decision-making during the consultation. For this research, we are only interested in the *'opl_recode'* column of the dataset, which denotes the education level of the patient. A "1" indicates a low education, which means that the patient has had no education or the highest level of education is primary school. A "2" indicates intermediate education, meaning the patient completed lower general secondary education or vocational training. A "3" indicates high education, meaning the patient completed pre-university education, high vocational training, or university. In Figure 1, the distribution of the education levels is plotted.

3.2 ABEL

3.2.1 Consult data

The ABEL dataset contains transcripts of consultations of patients with rectal cancer who are meeting with the radiologist to discuss whether they will undergo radiotherapy before the surgery to remove the tumor. The datasets consist of MS Word (doc) files of consultations between 161 patients and their doctors, one file per consultation. In this dataset, there are

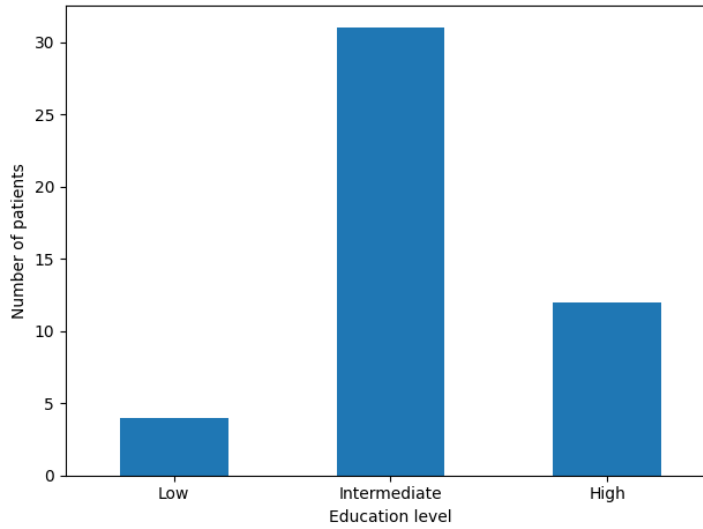


Figure 1: Number of patients for low, intermediate, and high education level from the available patient data for the IBIS dataset.

no patients with multiple consultations. Each consultation begins with its ID, followed by the consultation data. In each document is an explanation for the used symbols and fonts. All patient texts are denoted in italics, while the doctor texts are in non-italics. In most of the consultations, the patient is accompanied by someone else. These texts are also in italics, but contain a prefix such as “M:” (husband), “V:” (wife), or “E:” (partner). Comments (such as a period of silence) are written between round brackets and are bold. In some files, the consultation is entirely within a single table cell, whereas in other files the text is in regular paragraphs, with each paragraph corresponding to a turn.

3.2.2 Patient data ABEL

For 93 of the patients of the ABEL study, we have the education level as reported by the patient in a .sav file, which is scored as shown in Table 1. We have the belonging consultations of 73 of these patients.

Score	Education level
1.0	Primary school
2.0	Lower vocational education
3.0	Lower secondary education
4.0	Higher general secondary education
5.0	Pre-university education
6.0	Intermediate vocational education
7.0	Higher vocational education
8.0	University education
9.0	Other

Table 1: Education level options for ABEL

In Figure 2, the distribution of education levels among patients in the ABEL study is shown.

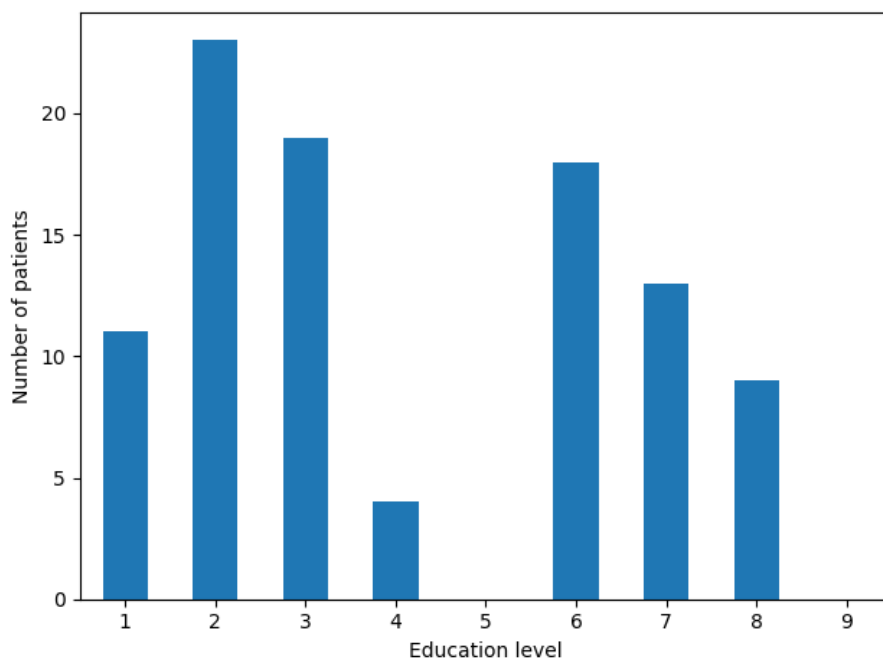


Figure 2: Number of patients for the different education levels from the available patient data for the ABEL dataset.

4 Methods

4.1 Preprocessing

We converted all .doc files to .docx, as these are easier to process in Python, and then used the Document class from *docx* to iterate over the paragraphs and tables in the documents. For consistency, the first letter of each sentence was capitalized, and all sentences were concluded with a full stop if a terminal punctuation mark was missing at the end of the sentence. As mentioned in Chapter 3, the document structure of the IBIS and IBIS-colon datasets is the same, but the structure of the ABEL dataset differs. In the following sections, we explain how we preprocessed the different files, resulting in per dataset:

1. One CSV file containing a pandas dataframe where the items are turns in the consultation, and the label is either Doctor or Patient, which is used for training a classifier to distinguish doctor and patient terms.
2. Per patient, a CSV file containing all doctor turns, and a CSV file containing all patient turns, which is used for estimating the literacy level of the doctor and the patient.

4.1.1 IBIS and IBIS-colon

For IBIS, we deleted overlapping files, for example, files called “part 1” when there is also a file containing the complete consultation. For the 12 patients with multiple consultations, we combine these three consultations into one file. We iterated all paragraphs and assigned all turns starting with “A:” or “A;” to the doctor and all turns starting with “P:” or “P;” to the patient. We ignore all other sentences in the document, such as explanation of the used symbols and turns of other doctors or company of the patient. We also delete all text within

the turns that are either between square brackets or round brackets, since this text indicates sounds of comments (such as “[hummen]”) and not spoken words. This results in a total of 26.499 turns over all consultations for IBIS, of which 14.113 from the doctor and 12.386 from the patient, and 3445 turns over all consultations for IBIS-colon, of which 1960 of the doctor and 1485 of the patient.

4.1.2 ABEL

Since the ABEL document does not contain explicit speaker prefixes to distinguish doctor and patient, we could not apply the same preprocessing steps for ABEL as we did for IBIS and IBIS-colon. Instead of including all text with a certain prefix, we excluded all text that does not belong to either doctor or patient. We excluded all turns that start with the prefix “Vrouw:”, “M:”, “V:”, “Vrouw2:”, “E:”, “Broer:”, “Man:”, “A2:”, “A:”, “V2:”, or “M2:”, since those are not belonging to the main doctor or the patient, but to either a companion of the patient or a second doctor. When one of these excluded prefixes occurred in a turn of either the patient or doctor, for example, when a companion interrupts the patient, we discarded all text after the prefix. We excluded the lines that explain the structure of the file, e.g., explaining that “V” denotes the wife of the patient. We also leave out text that is bold and text between parentheses, since these texts only contain comments or explanations about the file structure. After excluding all content that did not belong to the doctor or the patient, we iterated over the table cells and paragraphs, where we assigned all turns that are in italics to the patient and the remaining text to the doctor. This results in a total of 34.912 turns over all consultations, of which 18.545 turns are of the doctor, and 16.367 turns are of the patient.

4.2 Distinguishing doctor and patient turns with a classifier

Our first step was finding out whether or not there is a systematic difference in doctor and patient language by using a classifier to try to distinguish doctor and patient turns. For each dataset, we created a dataset that contains all labeled turns (as described in Section 4.1). The three datasets all contain consultations about adjuvant treatment, but the disease and possible treatments are different. In addition, the datasets are all of different sizes. Merging the three datasets would add noise and would not give a clear result in which words distinguish the doctor and the patient. Therefore, we chose not to merge these datasets and instead trained a separate model for each dataset. We split the data into train data and test data using an 80/20 split. This data was converted into numerical feature vectors using a TF-IDF Vectorizer³, with both unigrams and bigrams. We then trained a Linear Vector Support Classifier (LinearSVC⁴) on the vectorized training data and the labels, and then evaluated on the test data. The precision on the test data was used to inspect how well a classifier can distinguish the two classes.

To further explore these results, we inspected which words contribute most to the distinction between doctor and patient language. We did this by using the learned feature weights from the LinearSVC. Because LinearSVC is a linear model, each feature has a coefficient that reflects its contribution to the classification. We extracted the coefficient vector of the model and selected the ten features with the highest positive coefficients and the ten with the most

³https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

⁴<https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html>

negative coefficients. The highest positive coefficients are the words that were most distinct for the patient, and the most negative were most distinct for the doctor.

4.3 Dutch Biomedical Entity Linking

4.3.1 Replicating paper

To extract medical terminology from Dutch consultations, we used Dutch biomedical entity linking. Hartendorp et al [7] provided their GitHub repository for Dutch biomedical entity linking. Their process for creating a biomedical entity linking model for the Dutch language roughly consists of four steps: enhancing the UMLS, generating a WALVIS corpus, self-alignment pre-training of MedRoBERTa.nl, and finetuning the model with the WALVIS corpus. However, for replicating their process, we needed to write and rewrite a large part of their code.

First, we fixed syntax errors, most of which were presumably caused by the code being incompletely transferred from a Python file to Jupyter notebooks. In addition, some names, datasets, and methods were outdated, which were updated. One of the issues was the use of an outdated repository for Wikiextractor, which is used for extracting Wikipedia pages, which was solved by using an older version of Python (3.8) in a virtual environment. There was also a missing .env file in the repository, which was traced back and added. We published our GitHub repository, containing all the code that was used for this thesis⁵.

Secondly, we added a method for creating a custom concepts text file, which was used to retrieve the desired Wikidata data for the WALVIS corpus. We used <https://petscan.wmcloud.org/> for retrieving all medical categories from Wikipedia, by searching for the categories “Geneeskunde” (medicine), “Aandoening” (condition) and “Fysiologie” (physiology) with depth of 4.

Thirdly, we implemented and improved the method for creating a WALVIS corpus. The WALVIS corpus is created by retrieving Wikidata entries using a SPARQL query for retrieving all pages of the Dutch Wikipedia that contain a UMLS CUI, which is a Concept Unique Identifier. The SPARQL query is listed in Listing 1.

Listing 1: SPARQL query used to retrieve relevant Dutch Wikidata concepts

```
SELECT ?article ?concept ?cui WHERE {  
  ?article schema:isPartOf <https://nl.wikipedia.org/> .  
  ?article schema:about ?concept .  
  ?concept wdt:P2892 ?cui .  
}
```

We retrieved the SPARQL query in batches of 1000 results, timed out for 200 seconds and wrote the output to a jsonl file, which was repeated until all results are retrieved. This ensured that a large number of results can be retrieved without overloading the SPARQL endpoint or causing timeouts. We only included a result in our JSONL file if it is present in our enhanced ontology and does not belong to one of the excluded semantic categories, such as geographical areas or organizations. All excluded categories can be found in Appendix A.

We then selected the sentences from Wikipedia that contain a hyperlink to one of the collected pages from the Wikidata entries. We extracted the Wikipedia article titles from the URLs in the JSONL file and normalized them by decoding URL-encoded characters and replacing underscores with spaces. These normalized titles were used to make a dictionary where we

⁵https://github.com/annelottebaarslag/language_doctor_patient

map article titles to CUIs. We used *spacy* for segmenting each Wikipedia text dump document into sentences. For each sentence, we extracted the HTML hyperlinks using RegEx and we normalized their anchor text. If the anchor text existed in our CUI dictionary, we removed the HTML markup from the sentence and calculated the start and end index of the mention. We only kept sentences with at least one link that is present in our dictionary. We stored everything in a CSV file, where each row consisted of the biomedical mention, the CUI, the start and end index of the mention, and the sentence. This is our WALVIS CSV, which contains 21871 rows containing 3119 unique CUIs. We then dropped all duplicate mentions and all rows that contain CUIs that are not in our enhanced ontology, resulting in WALVIS*. Our WALVIS* has 6211 rows containing 3105 CUIs.

For creating the file for pretraining and the pretraining we used the provided code from Hartendorp et. al, resulting in a pretraining file in the format `CUI||term 1||term 2`. We performed hyperparameter tuning on MedRoBERTaNL using Optuna, with the following parameters:

- Learning rate: 10^{-5} , 10^{-4} , 10^{-3}
- Weight decay: 0.001, 0.01, 0.1
- Train batch size: 128, 256
- Miner margin: 0-0.2

We trained the final model with the optimal hyperparameters, which were a learning rate of 10^{-4} , weight decay of 0.01, train batch size of 128 and a miner margin of 0.003.

We rewrote the file to create the fine-tuning file, since the published code incorrectly created pairs in the form `CUI||mention 1||mention 2`, where both mentions are from the WALVIS corpus. Instead, we used the enhanced ontology to create pairs in the form of `CUI||mention||linked term`, where *mention* is a mention from the WALVIS* corpus and *linked term* the term in the enhanced ontology. Since some CUIs in our enhanced ontology have multiple names from different sources, we used the SNOMEDCT NL variant if it exists, otherwise we take the first name. We also performed hyperparameter optimization on the finetuned model, which resulted in a learning rate of 10^{-5} , weight decay of 0.001, train batch size of 256, and a miner margin of 0.023.

Like Hartendorp et al., we used the finetuned model to create embeddings for all concepts in the enhanced UMLS. Each concept name was tokenized, and the concept embeddings were stored in a matrix. We used Principal Component Analysis (PCA) to reduce the dimensionality to 256 embeddings, and all embeddings were L2-normalized. During inference, the mentions were tokenized, transformed to 256 dimensions, and L2 normalized. FAISS was used to compare the mention vectors to all indexed vectors from the ontology and returned the index of the closest concept and the cosine similarity. We only accepted predictions if the cosine similarity was above a threshold, which was empirically decided as described in Section 4.3.4.

Lastly, we also improved the existing code documentation and added multiple new documentation files to improve the process of replication in future research.

4.3.2 Named Entity Recognition

To apply BEL to the consultation files, we needed entity extraction to extract the medical terms. For this purpose, we fine-tuned MedRoBERTa.nl on a Dutch machine-translated version

of the MedMentions dataset.⁶ The dataset consists of sentences annotated with BIO labels (B, I, O), indicating the beginning, inside, or outside of a (medical) entity mention. We split each sentence into tokens and aligned the labels. In some cases, an I-label appeared without a B before it. To ensure correct BIO formatting, we converted these to B-labels. We used the MedRoBERTa.nl tokenizer, and labels were aligned to subword tokens by assigning the original label to the first subtoken and ignoring the following subtokens for the loss computation. We trained MedRoBERTa.nl for 5 epochs with a learning rate of 10^{-5} , a weight decay of 0.01, a warm-up ratio of 0.1, and a train and evaluation batch size of 16. The code was based on the Huggingface documentation for token classification.⁷

4.3.3 Expanding the WALVIS corpus

To enlarge the WALVIS corpus, we used English data in addition to Dutch. Hartendorp et al. suggested automatically translating the English Wikipedia pages to create the corpus. However, this is very expensive and inefficient, since essentially only the mentions from the corpus are used, not the full sentences. Instead, we chose to create an English WALVIS corpus, combined that with the enhanced UMLS to create positive pairs, and then only translated the English mentions, instead of the full corpus.

We followed the same steps as for the Dutch corpus, but used the English Wikipedia and English Wikidata dump instead. For our category file we used “Medicine”, “Medical condition” and “Physiology”, again with a depth of 4. We use Listing 2 for our SPARQL query.

Listing 2: SPARQL query used to retrieve relevant English Wikidata concepts

```
SELECT ?article ?concept ?cui WHERE {  
  ?article schema:isPartOf <https://en.wikipedia.org/> .  
  ?article schema:about ?concept .  
  ?concept wdt:P2892 ?cui .  
}
```

We also had to make a small change in the Wikiextractor source code, using “Category” instead of the Dutch “Categorie”. Our English WALVIS CSV contains 186,325 rows containing 9281 unique CUIs. The English WALVIS* has 29,545 rows containing 9240 CUIs.

We used the English WALVIS* corpus to generate positive pairs. The positive pairs look like CUI||English mention||linked term, where the linked term is the Dutch term from our ontology, which was linked in the same way as for our Dutch corpus.

For translation, we used *facebook/nllb-200-distilled-600M* from Huggingface⁸, and we only translated the second column – all English mentions – to Dutch. We used a batch size of 32 and a num_beams of 8. These values were chosen experimentally to get a high value for the number of beams and batch size without the GPU running out of memory. Since this way of translating resulted in some translation issues, we deleted all duplicate strings in the translations, removed all dots, and deleted all “ symbols.

We combined these translated mentions with our Dutch corpus and optimized and finetuned our pretrained model on this combined data. The optimal hyperparameters for the combined data were a learning rate of 10^{-4} , weight decay of 0.1, train batch size of 128 and a miner margin of 0.0004. This means that we ended up with two finetuned models: one model, which

⁶<https://github.com/mi-erasmusmc/DutchClinicalCorpora/tree/main/MedMentions>

⁷<https://huggingface.co/learn/llm-course/en/chapter7/2>

⁸<https://huggingface.co/facebook/nllb-200-distilled-600M>

was finetuned only on the original Dutch data, and one finetuned on the original Dutch data and the translated English data. Our Dutch and English WALVIS and WALVIS* csvs, together with our positive pair files for pretraining and finetuning, are uploaded to our GitHub repository.

4.3.4 Validation set-up

To select the best model configuration and decision threshold, we constructed a validation set consisting of 100 sentences. These sentences were randomly sampled by first choosing a random dataset, then choosing a random file within that dataset, followed by selecting all sentences longer than five words, and finally choosing a random sentence from those sentences. This process is repeated 100 times. Entities in these sentences were manually annotated using the following rules:

1. The entity represents a (bio-)medical concept.
2. The entity contains at least one noun.
3. The entity is not overly generic or non-specific in meaning (e.g., pain, healthy, appointment, risk, human).
4. The entity exists in our enhanced UMLS.

We wanted our NER model to extract all entities that follow rules 1 and 2, and our BEL model to link all entities that follow all 4 rules. We used Wikidata⁹ and our enhanced ontology to find the CUI that matches the biomedical entity. We compared our entities with the NER-extracted entities. If the whole entity matches our gold label, it is considered an exact match; if the model only extracts a part of the entity, this is considered a partial match. We also used our NER validation for deciding if and how we needed to strip the entities.

We had two finetuned BEL models, the one finetuned only on original Dutch data and the one finetuned on the combined data of the Dutch data and the translated English data. We tested four decision thresholds: 0.7, 0.8, 0.9, and 1. This gave us a total of eight different variants, which we all ran on the entities that the NER model extracted. We chose our final BEL model based on a balance between matching many entities to the gold standard, while not linking many NER false positives and generic terms, and not missing many entities from the gold standard.

4.3.5 Evaluating final performance

For estimating the performance of the final NER and BEL model, we created a separate evaluation set that consisted of 100 newly sampled and manually annotated sentences. These sentences were chosen in the same way as the validation set, but the sentences from the validation set were excluded from the candidate sentences. For assessing the performance of the NER model, we calculated the precision, recall, and F1-score compared to the gold medical terms. Partial matches were considered false negatives. For assessing the performance of the end-to-end process, we calculated the precision, recall, and F1-score from the BEL model compared to the linkable gold entities.

⁹<https://www.wikidata.org/wiki/>

4.4 Language complexity

For estimating the language complexity in the consultations, we measured and compared the general language complexity of doctor and patient, measured if and how much doctor and patient adjust their language complexity, and calculated the correlation between the pace of the dialogue and the language complexity.

4.4.1 General language complexity

We measured the general language complexity per consultation by the average syllables per word, average words per sentence, Flesch Reading Ease Score (FRE score), percentage of passive sentences, and type-token ratio. These metrics are similar to Roten et al. [25], but we do not keep the Flesch-Kincaid Reading Level Score since it uses the same variables as the FRE score. The FRE score is calculated by

$$206.835 - (1.015 * \frac{\#words}{\#sentences}) - (84.6 * \frac{\#syllables}{\#words})$$

A higher FRE score indicates easier language, while a lower FRE score indicates more complex language. Like Koch-Weser et al. [44], we added the type-token ratio to these measures, which is calculated by

$$\frac{\#unique\ words}{\#words}$$

. We used *Spacy* with the Dutch *nl_core_news_md* pipeline for extracting the number of sentences, syllables, and words. We used *num2words* for converting numbers in the text to written-out numbers.

For determining whether or not a sentence was passive, we used *Spacy* to check if there is a token in the sentence with a dependency called *nsubj:pass*. This means that the sentence has a passive nominal subject, which means that (a part of) the sentence is passive. This means that combined sentences that are partially passive are considered to be passive sentences. We divided the number of passive sentences by the total number of sentences to get the percentage of passive sentences of all sentences. We calculated the type-token ratio by dividing the length of the set of all words by the total number of words. We used *pyplot* from *matplotlib* to create boxplots for each metric for doctor and patient to visualize the differences.

4.4.2 Correlation with patient education level

The text difficulty metrics as described in Section 4.4.1 are generally used for estimating the literacy level for written text, not for spoken text. To determine whether these metrics can actually give an estimation for the education level of the patient, we calculated the correlation between the five text difficulty metrics and the education level of the patients. We used the education levels of the ABEL dataset, since that dataset has the most education level data available and uses a more specific range for education level (1-8) than IBIS (1-3), and the data is more evenly distributed over the education levels. Since the education is in ordinal scale, we used Spearman's rank correlation. We used *scipy.stats* for calculating the Spearman correlation coefficient (ρ) and the P -value in Python. We considered correlations with $P < 0.05$ to be statistically significant.

4.4.3 Language complexity adaptation

Besides the level of the doctor and patient over the whole text, we also analyzed whether or not the doctor and patient adapted their language complexity to each other over the time of the consultation. For this goal, we split each consultation into two parts, based on the total number of turns, which we called P1 and P2, respectively.

We used the FRE score as a metric to compare the change in language complexity, since this metric contains both the average words per sentence and the average syllables per word. We calculated the FRE score for both doctor and patient for P1 and P2. The change in complexity for each speaker was calculated as

$$\Delta_{\text{doctor}} = \text{FRE}_{\text{doc},P2} - \text{FRE}_{\text{doc},P1}, \quad \Delta_{\text{patient}} = \text{FRE}_{\text{pat},P2} - \text{FRE}_{\text{pat},P1},$$

where a positive value means that the language of the speaker became easier, and a negative value indicates it became more complex. We then analyzed the difference in language complexity between the doctor and the patient by calculating

$$\Delta_{\text{distance}} = |\text{FRE}_{\text{doc},P2} - \text{FRE}_{\text{pat},P2}| - |\text{FRE}_{\text{doc},P1} - \text{FRE}_{\text{pat},P1}|$$

A negative value indicates convergence (the language complexity of the doctor and patient became more similar), whereas a positive value indicates divergence (the language complexity became more dissimilar). We also calculated the mean percentage change (relative to P1) and standard deviation within both the converging and diverging consultations.

4.4.4 Dialogue pace

For the dialogue pace, we estimated the pace of the consultations to estimate if there is a correlation between the pace of the consultation and other metrics, and we calculated for both patient and doctor the mean amount of words per turn.

For a subset of IBIS consultations and all IBIS-colon consultations, we have the consultation durations in minutes. We extracted the duration from the document by taking the sentence that starts with “Duur:” and converted the time to seconds. If there are multiple time entries (e.g., 12:10 + 1:10), we added them together. If the duration was written in textual format (e.g., 33 min 5 sec), we extracted the hours, minutes, and seconds separately and converted them to seconds before summing them. In some of the consultations, the “Duur:” line was empty, and instead the duration was at the bottom of the consultation. In those cases, we took the last timestamp that appeared in the consultation as the duration.

During many consultations, a physical exam was performed, or parts were discussed that were deemed not interesting for the transcript, or contained personal information. These parts did take up part of the total consultation time but do not contain written text and are written as (xx:xx-xx:xx occupation). We fixed this by calculating the difference between the two timestamps and subtracting this duration from the total session time.

To estimate the dialogue pace, we used the total number of syllables computed in Section 4.4.2 and divided it by the dialogue length in seconds. The pace of a consultation is relevant for how easily a patient can process information and participate in the conversation, and the absolute values could be used for comparison in future research. However, the value of the pace on its own is difficult to interpret, and we cannot determine the difference in pace between doctor and patient. That is why we did not use the pace as a metric on its own, but used

it for analyzing if there is a relation between the speed of the consultation and the language complexity of the doctor.

We merged the dialogue pace from both IBIS and IBIS colon in one CSV file and do the same thing for the FRE score results. Since we did not have the pace of the doctor, we used the pace of the full consultation. Since the doctor is leading the consultation and talks more during the consultations, we assumed that the speech of the doctor mainly affects the pace of the consultation. We calculated the Spearman correlation coefficient and P -value between the pacing of the dialogue and the FRE score of the doctor. We considered a correlation with $P < 0.05$ to be statistically significant.

We calculated the average turn length from the doctor by dividing the amount of words from the doctor by the number of turns from the doctor and do the same thing for the patient. We created boxplots to show the differences in turn length between doctor and patient for each dataset.

4.5 Medical terminology use

For estimating the medical terminology use in the consultations, we measured and compared the amount of terminology used for doctor and patient per dataset, determined the most used medical terms, categories, medical terminology differences and differences in mentions and calculated the correlation between the language complexity and medical terminology use.

4.5.1 Amount of medical terminology used

We split each document into sentences using pySBD's sentence splitter. We used our finetuned NER model for extracting the entities from all sentences. The NER model outputs BIO labels, which are post-processed to fix invalid sequences. For example, an I label following an O is converted to B, and predictions for subword tokens are combined into whole words. After this, we applied additional filtering:

1. Entities must contain at least one noun (NOUN) or proper noun (PROPN).
2. Determiners and pronouns are stripped from each entity.
3. Entities are normalized by lowercasing, and we remove the punctuation.

We used Spacy with the Dutch *nl_core_news_md* pipeline to identify nouns and proper nouns and to strip them. We applied our BEL model to all extracted entities. These linked entities are considered to be medical terminology. For each consultation, we calculated for both doctor and patient the total amount of medical terms and the amount of unique words, both as a percentage of their total amount of words. Some mentions were very generic, such as "hoofd" (head). We used a list of the 1000 most frequent Dutch words from Wictionary¹⁰ to calculate the percentage of non-generic words of both doctor and patient, where words that are not in the top-1000 list are considered non-generic words. We performed our analysis on all words; we did not exclude the generic words for reproducibility.

¹⁰https://nl.wiktionary.org/wiki/WikiWoordenboek:Lijst_met_1000_basiswoorden

4.5.2 Most used medical terms, categories, medical terminology differences, and differences in mentions

For both doctor and patient, we sorted their used CUIs based on the total number of times that CUI was used in all consultations combined per medical term for both doctor and patient. We used our enhanced UMLS to map all CUIs to their UMLS category, called a TUI, and reported the top 5 categories for both patients and doctors. Furthermore, we determined the top 10 medical terms that are the most used by doctors that are not used by the patients at all, and the top 10 words that the patients used but the doctors did not. Lastly, we determined for all three datasets all CUIs that both doctor and patient referred to, but used at least one different mention to refer to that CUI.

4.5.3 Correlation language complexity and medical terminology use

Finally, we calculated the correlation between language complexity and medical terminology use for both doctor and patient. This is interesting to see if doctors and patients who used more medical terms also used more difficult language in their speech. For this purpose, we combined the BEL data of all datasets into one dataset and did the same with the language complexity data. We calculated the Spearman correlation coefficient and P-value between the FRE score and percentage of BEL mentions compared to the total amount of words. We considered correlations with $P < 0.05$ to be statistically significant.

5 Experiments and results

5.1 Distinguishing doctor and patient turns with a classifier (RQ1)

In Table 2 are the precision, recall, and F1-score from the classifier on the test set for both doctor and patient. In Table 3a are per datasets the top 10 words that were most distinguishing for doctor and patient. The bar plots with the exact coefficient value per word can be found in Appendix B.

Table 2: Classification performance across datasets

Dataset	Doctor				Patient			
	Precision	Recall	F1	Support	Precision	Recall	F1	Support
ABEL	0.78	0.80	0.79	3709	0.77	0.74	0.76	3274
IBIS	0.78	0.80	0.79	2794	0.77	0.74	0.75	2432
IBIS colon	0.79	0.86	0.82	392	0.79	0.70	0.74	297

Table 3: Top 10 features per class

(a) Doctor				(b) Patient			
#	ABEL	IBIS	IBIS colon	#	ABEL	IBIS	IBIS colon
1	heeft	bent	bent	1	ik	ik	ik
2	uw	uw	als	2	mijn	mijn	mijn
3	bent	heeft	heeft	3	me	ik had	euh
4	even	geven	uw	4	jij	dat ik	weer
5	kunt	chemotherapie	chemotherapie	5	hij	me	eh
6	gebruikt	even	want	6	heb	dat was	zei
7	krijgt	mevrouw	de	7	moet ik	jij	ben
8	mensen	we	we	8	ik had	chemo	bent dokter
9	vragen	kunt	mensen	9	wat is dat	hij	met lymfeklieren
10	wilt	zal	heel	10	ofzo	ben	ie

5.2 Dutch Biomedical Entity Linking (RQ2)

5.2.1 Validation

In our 100 random sentences, we labeled 109 medical terms that contained a noun, thereby matching rule 1 and 2. Of those, 34 terms were deemed too generic to be considered meaningful (rule 3), for example “people”. 32 of the remaining entities did not appear in our enhanced UMLS (rule 4), such as “Isotil” and “her2neu”. This is shown in Figure 3.

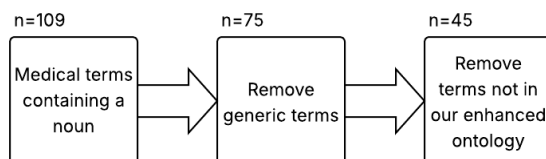


Figure 3: Flow chart of the extraction of gold labels for the **validation** set.

This means that we want our NER model to extract 109 entities and the BEL model to link 45 of them.

The NER model extracted in total 160 entities, of which 98 completely matched our labeled medical terms and 2 partially (e.g. “MRI” instead of “MRI scan”). This means that 60 of the entities that the NER model extracted were not deemed to be a medical entity. Our entities most matched our gold standard if we only kept nouns, proper nouns and adjectives in the entities. Of our 45 gold labels, NER extracted 38 entities correctly and 2 partially. These results can be seen in Table 4.

Table 4: Performance of the NER model on the validation set

Description	Count
Extracted (total)	160
Exact matches	98
Partial matches	2
False positives	60
False negatives	10
Gold entities found by NER	38 / 45

In Table 5 we show the results of the different BEL model variants. We chose C0.9 as our final model.

Table 5: Results from the different BEL models compared to the gold standard where D is the Dutch only data, C is the combination of the Dutch and translated English data and the digit is the decision threshold. The bold underlined cell is the best result per row.

Category / Data	D0.7	C0.7	D0.8	C0.8	D0.9	C0.9	D1	C1
Total BEL-linked entities	85	84	67	68	47	48	24	21
Matching to gold standard	<u>36</u>	<u>36</u>	<u>36</u>	35	31	32	21	18
Wrong CUI	2	3	1	2	<u>0</u>	2	<u>0</u>	<u>0</u>
Missed from gold standard	<u>2</u>	<u>2</u>	3	<u>2</u>	9	6	19	22
Generic/ Not in UMLS	34	36	24	25	13	13	<u>3</u>	<u>3</u>
NER false positives	13	10	6	6	3	1	<u>0</u>	<u>0</u>

5.2.2 Evaluation

In our 100 random sentences, we labeled 99 medical terms matching rule 1 and 2. Of those, 35 terms were deemed too generic to be considered meaningful. 19 of the remaining entities did not appear in our enhanced UMLS. This is shown in Figure 4.

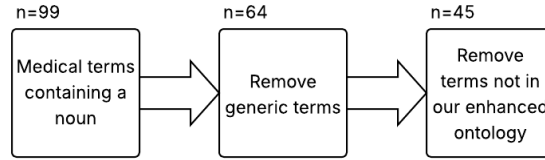


Figure 4: Flow chart of the extraction of gold labels for the **evaluation** set.

The performance of the NER model and the BEL model on the evaluation set is shown in Table 6. The precision, recall and F1 are in Table 7.

Table 6: NER and BEL model evaluation results

(a) Performance of the NER model on the evaluation set

Description	Count
Extracted (total)	160
Exact matches	92
Partial matches	0
False positives	68
False negatives	7
Gold entities found by NER	44 / 45

(b) BEL model results (combined data, threshold = 0.9)

Category	C0.9
Total BEL-linked entities	56
Matching to gold standard	38
Wrong CUI	1
Missed from gold standard	5
Generic / Not in UMLS	14
NER false positives	3

Table 7: Performance of the NER model compared to the 99 medical terms and the end-to-end performance compared to the 45 linkable gold entities.

	Precision	Recall	F1
NER	0.58	0.93	0.71
End-to-end	0.68	0.84	0.75

5.3 Consult complexity (RQ3)

5.3.1 General language complexity

Figures 5 to 7 show the boxplots for the text difficulty metrics for the different datasets.

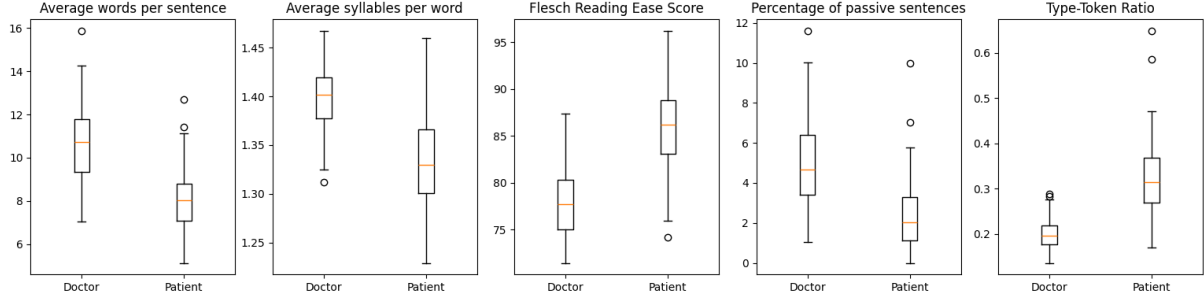


Figure 5: Text difficulty metrics of the ABEL dataset.

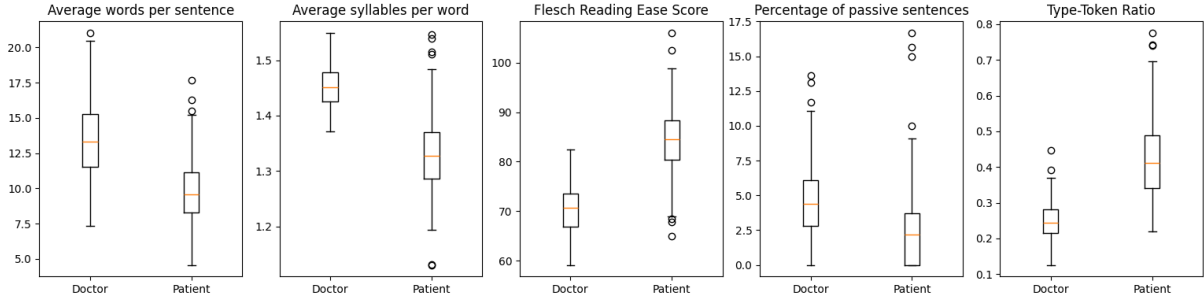


Figure 6: Text difficulty metrics of the IBIS dataset.

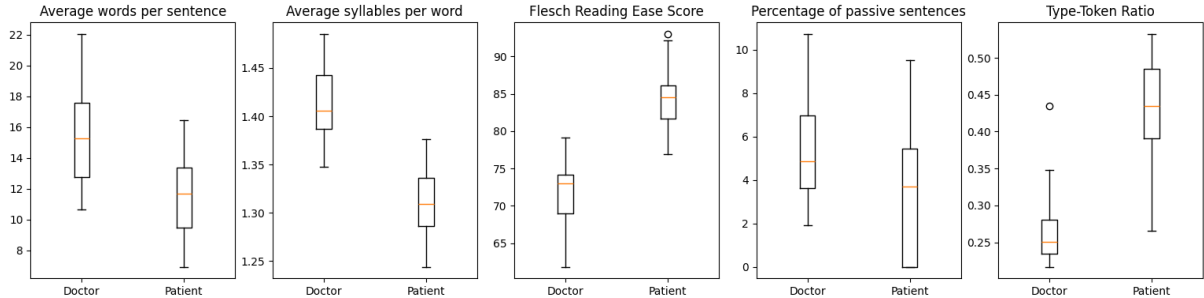


Figure 7: Text difficulty metrics of the IBIS colon dataset.

5.3.2 Correlation text difficulty metrics with patient education level

Table 8 shows the correlation between our text difficulty metrics and the education level. Average syllables per word has a strong positive correlation with education level ($\rho = 0.649$, $P < 0.001$), which indicates higher education levels are associated with longer, more complex words. The FRE score shows a strong negative correlation ($\rho = -0.640$, $P < 0.001$), which indicates that a higher education level is associated with a lower FRE score, so a higher text complexity. The other three metrics have no significant correlation with the education level.

Table 8: Spearman correlation between text difficulty metrics and the education level for 73 of the patients in the ABEL dataset.

Metric	ρ	<i>P</i>-value
Average words per sentence	0.123	.299
Average syllables per word	0.649	<.001
Flesch Reading Ease Score	-0.640	<.001
Percentage of passive sentences	0.226	.055
Type-Token Ratio	0.000	.999

5.3.3 Language complexity adaptation

Table 9 shows the change in the language complexity of the doctor and patient between the first and second half of the consultation. In 49.04% of the consultations the doctor and patient converged in language complexity and in 50.96% their language complexity diverged.

Table 9: Distribution of language complexity change between consultation halves.

Speaker / Direction	% of consultations	Avg % change	SD
Doctor Harder	36.44	4.84	3.92
Doctor Easier	63.56	7.68	6.52
Patient Harder	41.10	5.74	4.99
Patient Easier	58.90	6.67	6.88

In converged consultations, the doctor language became on average 7.0% easier, while the patient language became on average 3.0% harder. In diverged consultations, the doctor language became on average 0.6% harder, while the patient language became on average 5.9% easier. Over all consultations, doctor language became on average 3.1% easier and patient language 1.6% easier. Overall, no clear pattern can be observed in the change of language complexity. Convergence and divergence both occur approximately half of the consultations, and both speakers change in both directions. The changes are small and with a relatively big standard deviation, suggesting that there is no strong or consistent adaptation from either doctor or patient.

5.3.4 Dialogue pace

The Spearman correlation between the FRE score of the doctor and the dialogue pace (syllables per second) is $\rho = -0.166$ with a *P*-value of .021. This means that there is a significant weak negative association between the FRE score and the dialogue pace. This suggests that a higher pace of the dialogue is correlated with a higher complexity of the doctor speech. In Figure 8 the number of words per turn for both doctor and patients are per dataset shown in a boxplot.

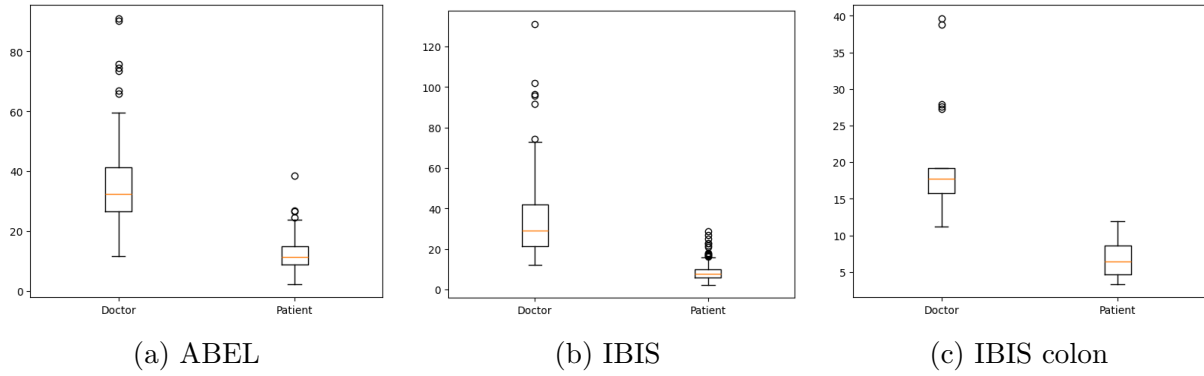


Figure 8: Amount of words per turn for doctor and patient.

5.4 Medical terminology use (RQ4)

Figure 9 to 11 show boxplots of the percentage of BEL mentions compared to the total amount of words per speaker, the percentage of unique BEL mentions and the percentage of unique BEL mentions which did not appear in the top 1000 most frequent Dutch words. We considered all mentions that our BEL model linked to the UMLS as medical terminology.

5.4.1 Amount of medical terminology used

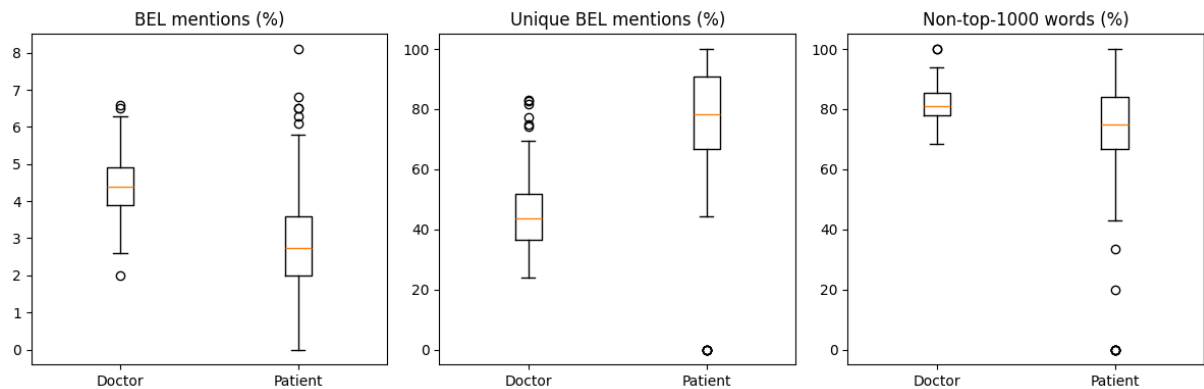


Figure 9: For IBIS, the percentage of BEL mentions compared to the total amount of words, the percentage of unique BEL mentions, and the percentage of BEL mentions which do not occur in the top-1000 most frequent Dutch words.

5.4.2 Most used medical terminology

Table 10 to 15 show the 10 most frequent used medical terms for each dataset for both doctor and patient.

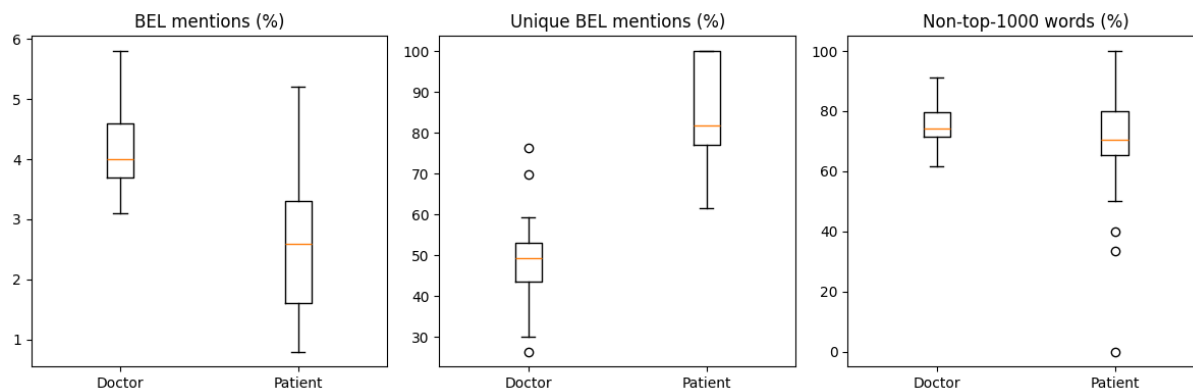


Figure 10: For IBIS colon the percentage of BEL mentions compared to the total amount of words, the percentage of unique BEL mentions, and the percentage of BEL mentions which do not occur in the top-1000 most frequent Dutch words.

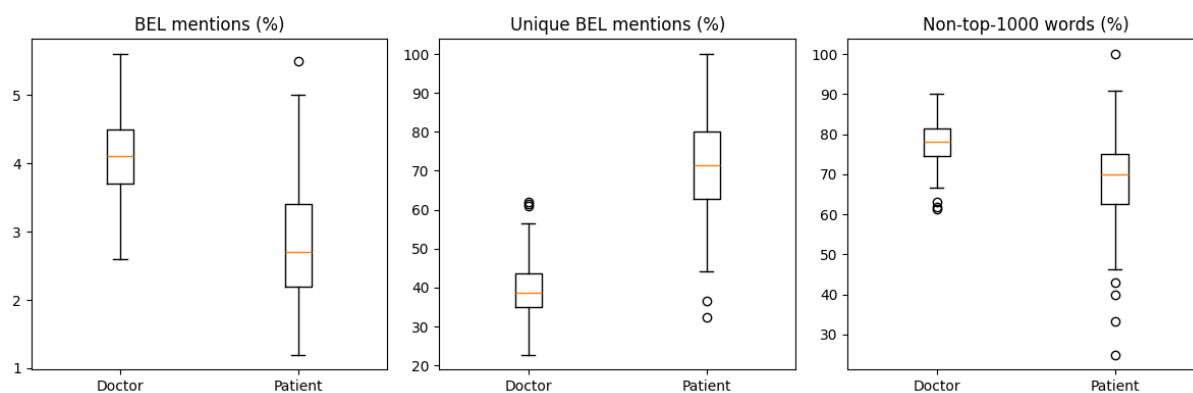


Figure 11: For ABEL, the percentage of BEL mentions compared to the total amount of words, the percentage of unique BEL mentions, and the percentage of BEL mentions which do not occur in the top-1000 most frequent Dutch words.

Table 10: ABEL - Top 10 most used medical terms of the doctor.

CUI	Meaning	Mentions	#
C1522449	bestraling voor oncologie en/of radiotherapie	bestraling, bestralingstraject, radiotherapie	3600
C0543467	chirurgische verrichting	chirurgie, operatie	1874
C0086418	homo sapiens	mens, mensen, mensheid	1632
C0027651	neoplasma	tumor	1065
C0015733	feces	ontlasting	912
C0441633	scan	scan	713
C0040405	computertomografie	CT, CT scan, CT-scan	672
C0009429	combinatietherapie	behandeling	571
C0079941	ORF's	gebied	513
C0003461	structuur van anus	anus	447

Table 11: ABEL - Top 10 most used medical terms of the patient.

CUI	Meaning	Mentions	#
C1522449	bestraling voor oncologie en/of radiotherapie	bestraling, radiotherapie	311
C0543467	chirurgische verrichting	chirurgie, operatie	240
C0016117	financieel beheer, ziekenhuis	ziekenhuis	228
C0015733	feces	ontlasting	211
C0448970	uitwendig stoma	stoma	203
C0005767	bloed	bloed	182
C0237935	geloof	geloof, hoop, vertrouwen	120
C0204774	weken	weken	111
C0030193	pijn	pijn	110
C0027651	neoplasma	tumor	103

Table 12: IBIS - Top 10 most used medical terms of the doctor.

CUI	Meaning	Mentions	#
C3665472	chemotherapie	chemo, chemo therapie, chemokuur, chemotherapie, chemotherapieën	2399
C0279025	hormoontherapie	hormonaal therapie, hormonale behandeling, hormonale behandelingen, hormonale therapie, hormoontherapie, hormoontherapie/, hormoontherapieën	976
C0086418	homo sapiens	mens, mensen	831
C0043210	vrouw	vrouw, vrouwen	738
C0009429	combinatietherapie	behandeling, combinatietherapie	665
C0027651	neoplasma	tumor	573
C0013227	drug of geneesmiddel	geneesmiddel, medicament, medicamenten, medicatie, medicijn, medicijnen	535
C0204774	weken	weken	532
C1522449	bestraling voor oncologie en/of radiotherapie	bestraling, bestralingstraject, radiotherapie	472
C0006826	maligne neoplasma	kanker, kankers, maligniteit	392

Table 13: IBIS - Top 10 most used medical terms of the patient.

CUI	Meaning	Mentions	#
C3665472	chemotherapie	chemo, chemokuur, chemotherapie, chemotherapieën	338
C0204774	weken	weken	126
C1522449	bestraling voor oncologie en/of radiotherapie	bestraling, radiotherapie	105
C0006141	borst	borst, mamma	102
C0279025	hormoontherapie	endocriene therapie, hormonale therapie, hormoontherapie	91
C0543467	chirurgische verrichting	chirurgie, operatie	89
C0013227	drug of geneesmiddel	medicatie, medicijn, medicijnen	83
C0006826	maligne neoplasma	kanker, kankers	76
C0086418	homo sapiens	mens, mensen	74
C0026591	moeder	moeder	63

Table 14: IBIS colon - Top 10 most used medical terms of the doctor.

CUI	Meaning	Mentions	#
C3665472	chemotherapie	chemo, chemokuur, chemotherapie	211
C0086418	homo sapiens	mens, mensen	135
C0204774	weken	weken	72
C0009429	combinatietherapie	behandeling	70
C0543467	chirurgische verrichting	operatie	69
C0006826	maligne neoplasma	kanker	54
C0856249	lymfeklieren, vergroot (excl. lymfadenitis)	lymfeklieren	47
C0013227	drug of geneesmiddel	medicatie, medicijn, medicijnen	44
C0221423	toestand van ziekte	kwaal, ziekte	44
C0346627	maligne neoplasma van darm	darmkanker	34

Table 15: IBIS colon - Top 10 most used medical terms of the patient.

CUI	Meaning	Mentions	#
C0543467	chirurgische verrichting	operatie	13
C3665472	chemotherapie	chemo, chemokuur, chemotherapie	12
C0856249	lymfeklieren, vergroot (excl. lymfadenitis)	lymfeklieren	11
C0030193	pijn	pijn	11
C0204774	weken	weken	10
C0574032	infuus	infuus	10
C0027651	neoplasma	tumor	9
C0043210	vrouw	vrouw	7
C0013227	drug of geneesmiddel	medicijnen	6
C0016117	financieel beheer, ziekenhuis	ziekenhuis	6

5.4.3 Most used categories

Table 16 to 18 show the top five used UMLS categories for each dataset for both doctor and patient.

Table 16: ABEL - Most used medical term categories.

a) Doctor		b) Patient	
Category	%	Category	%
Therapeutic or Preventive Procedure	27.3	Therapeutic or Preventive Procedure	16.0
Body Part, Organ, or Organ Component	9.3	Body Substance	8.6
Diagnostic Procedure	6.9	Body Part, Organ, or Organ Component	8.5
Human	6.4	Neoplastic Process	5.4
Neoplastic Process	6.4	Sign or Symptom	5.4

Table 17: IBIS - Most used medical term categories.

a) Doctor		b) Patient	
Category	%	Category	%
Therapeutic or Preventive Procedure	32.1	Therapeutic or Preventive Procedure	27.6
Neoplastic Process	9.3	Body Part, Organ, or Organ Component	8.2
Disease or Syndrome	5.6	Neoplastic Process	7.3
Sign or Symptom	5.4	Pharmacologic Substance	4.9
Body Part, Organ, or Organ Component	5.4	Sign or Symptom	4.4

Table 18: IBIS colon - Most used medical term categories.

a) Doctor		b) Patient	
Category	%	Category	%
Therapeutic or Preventive Procedure	31.6	Therapeutic or Preventive Procedure	18.8
Human	8.7	Body Part, Organ, or Organ Component	9.1
Disease or Syndrome	7.4	Neoplastic Process	7.1
Neoplastic Process	7.3	Disease or Syndrome	6.5
Sign or Symptom	6.8	Sign or Symptom	6.5

5.4.4 Medical terminology use differences

In Table 19 to 21 we show the top 10 CUIs used by patient but not doctor and reversed. In Table 22 is per dataset the percentage of CUIs and mentions that the doctor used that were not used by the patient and reversed.

Table 19: ABEL - Most used CUIs only used by only doctors (a) or only patients (b).

(a) CUIs only used by doctors.

CUI	Mentions
C0001527	vetweefsel (76)
C0031809	lichamelijk onderzoek (67)
C5769145	bestralingsplan (59), bestralingsplannen (1)
C0005682	blaas (44), blaas weg (1)
C0230278	kleine bekken (30), kleine bekkengebied (2), kleine bekken gebied (1)
C0024094	regio (29)
C0043240	wondgenezing (27)
C0034014	schaambeek (27)
C0026724	slijmvlies (23), slijmvlieslaag (1)
C0030851	penis (24)

(b) CUIs only used by patients.

CUI	Mentions
C0022408	artrose (10)
C0037366	rook (9)
C0230403	pink (6)
C0005684	blaaskanker (6)
C0005615	bevalling (3), geboorte (3)
C0005558	biopsie (6)
C0087178	jeugd (5)
C0007189	cardiologie (5)
C0037993	milt (4)
C0003483	aorta (4)

Table 20: IBIS - Most used CUIs only used by only doctors (a) or only patients (b).

(a) CUIs only used by doctors.

CUI	Mentions
C0005821	bloedplaatjes (59)
C0002871	bloedarmoede (41), bloedarmoede erg (1)
C0018270	groei (34)
C0161647	geneesmiddelen (32)
C0729594	okselklieren (25), okselklier (7)
C0005841	bloedtransfusie (20), transfusie (4)
C0007798	tumoren (23)
C0005823	bloeddruk (21)
C0005773	bloedcellen (17)
C0206692	lobulair carcinoom (14), lobulaire carcinoom (3)

(b) CUIs only used by patients.

CUI	Mentions
C0024202	lymfe (8)
C0186081	heupoperatie (3), heupoperaties (2)
C3495832	reumatische klachten (3)
C0034770	herinnering (3)
C0036572	aanval (3)
C0010709	cyste (3)
C0027740	zenuw (2)
C0027086	myoom (2)
C0029062	operatiekamer (2)
C0019134	heparine (2)

Table 21: IBIS colon - Most used CUIs only used by only doctors (a) or only patients (b).

(a) CUIs only used by doctors.

CUI	Mentions
C0087111	therapie (21), therapieën (2)
C0030705	patiënten (12), patiënt (10)
C0012634	ziekte (18), ziektes (3)
C0011991	diarree (11)
C0005953	beenmerg (8)
C0005821	bloedplaatjes (7)
C0011646	huid (5)
C0001249	infecties (5)
C0020538	hoge bloeddruk (5)
C0032285	longontsteking (5)

(b) Patient

CUI	Mentions
C0448970	stoma (3)
C2711273	icd (3)
C0014245	endoscopie (2)
C0079941	gebied (2)
C0162340	verstand (2)
C0026591	moeder (2)
C0026549	morfine (2)
C0019552	heup (2)
C0009471	thuiszorg (2)
C0007531	blinde darm (2)

Table 22: The percentage of CUIs and mentions of the doctor that were not used by the patient and reversed.

	CUIs		Mentions	
	Unique doctor	Unique patient	Unique doctor	Unique patient
ABEL	41.1	32.7	44.9	34.9
IBIS	51.9	31.0	56.1	32.8
IBIS colon	66.5	42.2	66.8	41.0

5.4.5 Mention differences

Table 23 to 25 show all CUIs that both doctor and patient use, but use different mentions to refer it to.

Table 23: ABEL - Different mentions that doctors (D) and patients (P) use for the same CUI.

CUI	Meaning	Doctor only mentions	Patient only mentions	Shared mentions
C0009653	condoom	schede (29)	condoom (1)	-
C0426649	darmkramp	darmkrampen (13)	darmkramp (1)	-
C0038999	uitstulping	zwelling (9)	bult (3)	-
C0021831	aandoening van darm	darmziekten (2), darmziektes (2)	darmziekte (1)	-
C0241705	mictieproblemen	plasproblemen (4)	plasprobleem (1)	-
C0007097	carcinoom	carcinoom (3)	kwaadaardige gezwel (1)	kwaadaardige tumor (D:20, P:5)
C0032098	plant	plant (3)	planten (1)	-
C0001948	alcoholconsumptie	alcohol gebruik (2)	alcoholgebruik (1)	-
C0006098	brachytherapie	inwendige bestraling (2)	brachytherapie (3)	-
C0001807	agressief gedrag	agressiviteit (1)	agressie (1)	-
C0003811	hartritmestoornis	ritmestoornis (1)	hartritmestoornis (1)	-
C0013227	drug of geneesmiddel	geneesmiddel (1)	medicijntjes (1)	medicijnen (D:228, P:54), medicatie (D:21, P:6), medicijn (D:16, P:6)
C0013987	emotie	emotie (1)	emoties (1)	-
C0014724	boer	boeren (1)	boer (1)	-
C0042449	structuur van vena	aderen (1)	ader (1)	-
C0186081	operatie van heupgewricht	heupoperaties (1)	heupoperatie (2)	-

Table 24: IBIS - Different mentions that doctors (D) and patients (P) use for the same CUI.

CUI	Meaning	Doctor only mentions	Patient only mentions	Shared mentions
C0023516	leucocytus	witte bloedcellen (63)	leukocyten (1)	-
C0279025	hormoontherapie	hormonale behandeling(40), hormoontherapieën(3), hormonaal therapie(2), hormonale behandelingen(2), hormoontherapie/(1)	endocriene therapie (1)	hormoontherapie (D:673,P:79), hormonale therapie (D:255,P:11)
C0677944	schildwachtklier	sentinel node (5)	poortwachtersklier (2)	schildwachtklier (D:46,P:9)
C0014180	structuur van endometrium	baarmoederslijmvlies (4)	endometrium (1)	-
C0005778	bloedstolling	stolling (3)	bloedstolling (2)	-
C0009443	acute rinitis	kou (2)	verkoudheid (2)	-
C0009938	contusie	blauwe plekken (2)	blauwe plek (3)	-
C0030971	perceptie	perceptie (2)	waarneming (1)	-
C0011980	structuur van diafragma	pessarium (1)	middenrif (1)	-
C0013404	dyspneu	kortademigheid (1)	benauwdheid (1)	-
C0156543	zwangerschap met abortus als resultaat	abortus (1)	miskraam (1)	-
C1140680	maligne neoplasma van ovarium	eierstok kanker (1)	eierstokkanker familie (1)	eierstokkanker (D:10, P:2)

Table 25: IBIS colon - Different mentions that doctors (D) and patients (P) use for the same CUI.

CUI	Meaning	Doctor only mentions	Patient only mentions	Shared mentions
C0042449	structuur van vena	aderen (1)	ader (1)	-

5.4.6 Correlation language complexity and medical terminology

Table 26 shows the Spearman correlation between the FRE score of the doctor and patient and their medical terminology use. There is a significant negative correlation for both doctor and

patient. This means that for doctor and patient, a higher language complexity is associated with the use of more medical terminology.

Table 26: Spearman correlation between the FRE score and the percentage of medical terminology use compared to the total amount of words.

	ρ	<i>P-value</i>
Doctor	-.320	<0.001
Patient	-.385	<0.001

6 Discussion

6.1 NLP perspective on evaluation of models

Quality of the transcripts The quality of the transcripts of the consultations affected our results. The texts contained spelling errors, missing punctuation, and were not formatted consistently. This influenced the scores for the different text difficulty metrics and resulted in some of the turns being assigned to the wrong speaker. For example, missing punctuation caused sentence lengths to be reported longer or shorter than they actually are. PySBD is a rule-based sentence splitter, so incorrect punctuation leads to incorrect detection of sentence boundaries. Furthermore, spelling errors influenced syllable count and the total number of words. In addition, these transcripts are based on spoken language, which means sentence boundaries are not always clearly defined. As a result of this, there is variability, since the metrics that are based on sentence length (FRES and average words per sentence) are dependent on how the transcriber added punctuation. In addition, spelling errors resulted in some medical entities being missed by the NER model or incorrectly linked by the BEL model, e.g., the writing of “zieken huis” instead of “ziekenhuis”. Besides entities being missed, it also caused some mentions to be considered different between doctor and patient while they are not actually different, for example the CUI C0001948 being mentioned to as “alcohol gebruik” by the doctor and “alcoholgebruik” by the patient.

We also experienced difficulties because of the inconsistent way that the data was structured. This gave a challenge in the preprocessing for distinguishing the doctor and patient turns, since the prefixes that were used for the companion of the patient and the way that doctor and patient turns were denoted differed both between and within the datasets. This meant that speaker-indicating prefixes had to be manually filtered out. There was also an inconsistency in the way that the duration of the consultation was denoted, both in the placement of the duration in the consultation and in the format of the duration itself. These inconsistencies can be explained by the goal for which these datasets were created. Transcripts are made for a specific goal, and these transcripts were not made for automatic text analysis, but were used for manual analysis. These factors mean that for the original research, typos and inconsistencies in the data were not a problem, while they do cause issues when we used code to extract information from the interviews. For future research, it would be desirable to manually correct typos and inconsistencies in the data for a better quality of the results.

Classifier The classifier achieved a reasonably strong result in distinguishing between doctor and patient turns for all datasets. This indicates that there is a difference in their language. However, when inspecting the top features per class, we can see that this difference seems to be mainly caused by the difference in perspective (e.g. “ik” vs “uw”) rather than by differences in use of medical jargon. This difference can be explained by the topic of the consultations: the topic of the consultations is the disease and treatment of the patient, so it makes sense that the patient is mainly talking about themselves, therefore using words such as “ik” and “mijn” more often, while the doctor is talking about the patient, therefore using “uw”.

NER model The NER model achieved high recall but relatively low precision, indicating that it successfully extracted most relevant medical entities, but also gave a large number of false positives. This was likely caused by the dataset that was used for training the model, which was fully machine-translated. This introduced errors in the dataset. A better and bigger

Dutch NER corpus would improve the performance of the NER model. A part of the errors of the NER model comes from mistakes of the Dutch *spacy* pipeline, such as labeling “Oke” as a noun or “ingroei” as a verb. Besides manually correcting such mistakes there is not one clear solution to fix this.

BEL model Our Dutch BEL model was able to correctly label almost all our entities in the evaluation set. The overall end-to-end performance of our pipeline shows that it is reasonably effective for identifying and linking medical concepts in Dutch clinical consultations. Adding the translated English data to the positive pairs for finetuning gave a slightly better result on the evaluation set. However, since this was a minor improvement on 100 sentences, it does not mean that it improved the performance of the BEL model over the whole dataset. Since translating of the English mentions takes extra time and causes translation errors and the training takes more time because of the bigger dataset, we would advice to only use the Dutch positive pairs in future research.

Our BEL model also had some systematic issues. First, some words were consequently linked while they should not be linked, such as the plural noun “weken” (*weeks*), which is not a medical term but was linked to the verb “weken” (*to soak*), which occurs in the enhanced ontology. Other words were consequently linked to the incorrect entity, such as “lymfeklieren” to C0856249, “lymfeklieren, vergroot (excl. lymfadenitis)” instead of C0024204, “lymfeklier”. If there is a need for more thorough research in the datasets, such mistakes could be manually filtered out or fixed. Secondly, because our model is not context-aware, it gave false positives on words that are medical terms on their own, but not in a context. For example, in the figurative phrase “er is iets aan de hand” (*something is up*), our model linked “hand” to C0018563, hand, the body part, while in this context “hand” is used figuratively. Thirdly, our model often does not link abbreviations, unless their abbreviation is also present in the enhanced ontology. We can also see in our results that many of the seemingly different mentions between doctor and patients are only different because one of the speakers only used the singular form, while the other used the plural form, for example “blauwe plekken” en “blauwe plek” (*bruise*) in IBIS for the CUI C0009938. This seems to be less informative; it might be useful in future research to use the lemma of the mentions instead to filter out these differences.

Enhanced ontology Our enhanced ontology also added noise to our pipeline. We mainly experienced the following issues:

1. A portion of the medical terminology in our texts did not occur in our enhanced UMLS.
2. A small part of concepts in the UMLS have multiple CUIs.
3. Some of the CUIs cover multiple concepts, for example, C0156543 is used for both “abortus” and “miskraam”, while they are different things. In a similar way, C0009653 is used for both “condom” and “schede”.

6.2 Differences in doctor and patient language

Difference in language complexity We observed a consistent difference between the language complexity of the doctor and the patient. In all datasets, the averages of most of the text difficulty metrics were visibly higher for the doctor than for the patient. This indicates that doctors consistently use more complex language than patients, which is consistent with prior

work, which concluded medical professionals generally use more formal language [45] [27]. This difference can partially be explained by the difference in roles in the consultation. The doctor is a medical professional and has to explain terms such as side effects and treatment options, and has to ask questions to the patient, while the patient is mainly asking and answering questions. Especially when the doctor is asking closed questions, this leads to a difference in word length and sentence length. This does not mean that the patients do not use complex language at all, but on average, their language complexity is lower than that of the doctor in all datasets. The doctor using more difficult language than the patient can lead to a worse understanding of the patient, which can hinder shared decision-making.

Relation to education level We found a significant correlation between the FRE score and the education level of the patient, and the average syllables per word and the education level., the average syllables per word, This suggests that, while these metrics are usually used for written text, they also can give a good indication for the education level of the patient for spoken text. This can be used in further research for transcripts where the education level of the patient is not available, but we still want an estimation of the education level.

Adaptation of language Contrary to our expectations, doctor and patient did not consistently converge their language complexity during the consultations. Half of the consultations showed convergence, while the other half showed divergence, and the magnitude of these changes was relatively small. There also did not seem to be a consistent pattern of complexity becoming easier or harder for either doctor or patient. The raw numbers seem to suggest some patterns, for example the doctor language on average becoming slightly easier. However, these numbers are all very small and have a relatively large standard deviation, and we do not want to overinterpret the results. Therefore, we do not draw conclusions from these results.

Relation to dialogue pace We found a significant weak negative association between the FRE score and the dialogue pace, suggesting that higher-paced dialogues also had more difficult language from the doctor. A higher pace means that it is more difficult for the patient to understand the information. When the doctor also uses more complex language, the information becomes even harder to understand. This might indicate that some doctors are more aware of trying to make their words understandable, therefore talking slower and easier, while other doctors talk both fast and more complex. It is also visible that doctors use many more words per turn than the patients. Research indicates that a doctor should not give too much information in a turn, but rather break down the information into smaller chunks, instead of giving information all at once [46]. Too much content in one turn can make the patient overloaded and unable to process the information, and this could hinder shared decision-making.

Difference in medical terminology use We also found differences between doctors and patients in their use of medical terminology. Across all datasets, doctors consistently used a higher proportion of medical terms than patients. This is consistent with the literature, which concluded doctors used more medical jargon than patients in English consultations [28, 29]. However, it is important to note that we considered all mentions that our BEL model linked to as medical terms. We can see in the results that our BEL model also links a substantial amount of words that are medical terms, but are still likely to be understood by the patient, such as “operatie” or “tumor”. On the contrary, very niche medical jargon terms will not be

linked by the BEL model if they do not exist in our enhanced UMLS. The distributions of the medical terminology should be considered an estimation. There does not seem to be a consistent pattern for either doctor or patient to use more non-generic terms.

Some words were used a lot by both patient and doctor, such as “chemotherapie”, “bestraling”, and “operatie”, but we also see words that were frequently used by the doctor but not that much by the patient, and vice versa. We also see this in the most used categories for doctors and patients. For all datasets, the most medical terms for both doctor and patient fell in the category “Therapeutic or Preventive Procedure”. This makes sense given the nature of the consultations, which are mainly about discussing treatment options. However, there are also some categories which were often used by doctors or patients but much less by the other. For example, in the ABEL dataset, we see that the category “Body Substance”, which contains terms such as blood and feces, is the second most used category for medical jargon for the patient, while it is not in the top 5 most used categories of the doctor. We also see many medical terms that were solely used by the doctor and not by the patient at all, such as “vetweefsel” in ABEL, “bloedplaatjes” in IBIS, and “therapie” in IBIS colon. A big part of the medical terms that the doctor used were not used by the patient, ranging from 41.1% to 66.5% across the datasets. In reverse, 32.7% to 42.2% of the medical terms of the patient were not used by the doctor at all. This suggests a mismatch in medical terminology use: a big part of medical words that the doctor uses are not used the patient and the patient uses terms that the doctor does not reuse. The doctor not reusing the words of the patient could partially be caused by the patient talking about things that are unrelated to the current diagnosis or treatment. For example, for ABEL, we see that the patient used the word “artrose” (*osteoarthritis*) while the doctor did not. Osteoarthritis is not related to the topic of the conversation, so this could be a side story of the patient that the doctor did not necessarily needed to repeat. The patient talking about medical topics that are not related to the treatment plan could also explain why the percentage of unique medical terminology was higher for the patient than for the doctor: the patient will most likely talk about more topics besides the treatment than the doctor.

For a small part of the words, the doctor and patient used a different word for the same term. For example, in ABEL, the doctors used the word “zwellen”, while the patients used the word “bult”, and in IBIS, the doctors used “sentinel node” while patients used “poortwachtersklier”. However, a big part of the mention differences were artifacts of the analysis pipeline, caused by the difference in use between the singular and plural form of the word. In future research, it could be possible to transform all words to their singular version or only use the lemma. While we would expect that the doctor would use the “more technical” medical terms and patients the simpler terms, we can see that this is not necessarily true for all words. For example, the patients used “brachytherapie” while the doctors used “inwendige bestraling”. This could be explained by the patients already having knowledge and experience with their disease: they have already spoken with different doctors before and received medical information, which could explain why they know and use the medical terms. The higher use of medical terminology by both doctors and patients aligns with their higher language complexity, which could be explained by higher educated people both using more medical terms and more complicated language.

Difficulty of doctor language Finally, we can analyze how difficult doctors have made it for patients to understand the information. First, the consistently higher language complexity of

the doctors compared to the patients can make it more difficult for patients to fully process the information that the doctor provides. Doctors should try not to make their words, sentences, and turns too long in order to match the language of the patient. Secondly, the higher use of medical terminology by the doctor, in combination with the occurrence of words that the doctor uses that are not used by the patient, may cause patients not to adopt or fully understand the terminology introduced by the doctor. Since the average language complexity of the doctor is substantially more difficult than that of the patient and the doctor uses more terminology, many of which are not used by the patient, the doctor makes it difficult for the patient to understand the information that is discussed.

The differences in language complexity and medical terminology use between doctors and patients do not necessarily mean that the patients had a worse understanding of the information provided by the doctor because of these differences. Patients may have understood the information without explicitly reusing the same terminology, and they could still understand the doctor even if the doctor uses more complex language than they do. In future research the patients could be given a questionnaire about how well they understood the information that the doctor provided, to see if the differences that we found are directly reflected in the patient understanding.

7 Conclusion

In this thesis, we analyzed the differences between the language complexity and medical terminology use of doctors and cancer patients in transcripts of Dutch medical consultations. We used text difficulty metrics and Dutch biomedical entity linking to quantitatively assess these differences. With these results, we are able to answer the research questions.

RQ1. *How well can a classifier distinguish doctor and patient turns and which words mainly contribute to this difference?*

For all three datasets, the classifier had a decent performance in distinguishing doctor and patient turns, which indicates there is a clear difference in their language. The words mostly distinguishing the doctor language were in all datasets, mainly verbs such as “heeft” and “bent”, and pronouns such as “uw”. The words that mostly distinguished the patient language were in all datasets mainly personal pronouns such as “ik” and possessive pronouns such as “mijn”. This difference can be explained by the doctor leading the conversation and the patient being the object of the consultation. This means that the main signal for the classifier is differences in the conversation roles rather than the use of medical terminology.

RQ2. *What is the quality of the MedRoBERTa.nl model trained on weakly supervised Wikipedia data when applied to transcripts of Dutch medical consultations?*

The combination of NER and our Dutch BEL model was reasonably effective for identifying and linking medical concepts in our Dutch medical consultations. The majority of medical terms were correctly identified and labeled, meaning our pipeline is useful for extracting medical terminology from consultations. A major flaw in our pipeline is that our model is not context-aware, causing non-medical terms to be linked to medical terms. The model was also often unable to link abbreviations. Manual corrections in both the transcripts and the pipeline would be needed in order to reach better performance.

RQ3. *What are the differences in language complexity between doctors and patients, estimated with existing text difficulty metrics?*

Doctors used more complex language than the patient in all datasets, for example, they used longer words and longer sentences. The FRE score and average syllables per word were positively correlated with education level, which suggests that these metrics, while originally intended for written text, can serve as an approximation of the education level of the patient in consultation transcriptions. There did not seem to be a clear pattern for either doctor or patient in the change of language complexity. Higher-paced dialogues also had more complex language complexity of the doctor, which potentially made the information for the patient even more difficult to follow. The doctors used, on average, more words per turn than the patients, suggesting that they might give too much information per turn, making it harder to understand and remember.

RQ4. *What are the similarities and differences in medical terminology use between doctors and patients, and how is the terminology use related to language complexity?* Doctors consistently used more medical terminology than patients. Words that are commonly used in cancer consultations, such as “chemotherapie”, “bestraling” and “operatie” were used frequently by both doctor and patient. In all three datasets the doctor used a lot of medical terms that the patient did not use and reversed, suggesting a mismatch in medical terminology use. Doctor and patient also sometimes used different words for the same concept, but this was not very common. It is important to note that the dataset is quite small, so the results are about relatively small numbers. There is a significant positive correlation between the use of medical terminology and language complexity for both doctors and patients, which means that a higher

language complexity often also means that they used more medical terminology.

Based on our results, we can answer the main research question: **To what extent do doctors and patients differ in their language use in Dutch cancer consultations and how well can we measure these differences using Dutch Biomedical Entity linking and text difficulty metrics?**

We found that doctors, on average, used more complex language and more medical terminology than patients, and these differences can be automatically assessed using text difficulty metrics such as the FRE score for language complexity and a pipeline of Named Entity Recognition followed by BEL for the medical jargon use. The higher language complexity of the doctor and higher use of medical terminology means that the doctors made it relatively difficult for the patient to understand the information, which can be not beneficiary for reaching Shared Decision Making. These results do not necessarily mean that the patient actually had trouble understanding the information because of these differences.

This study had several limitations. The quality of the transcripts influenced both the results of the text metrics, as well as the results of the NER and BEL model. In addition, the BEL model we used is not context-aware, and our enhanced UMLS misses medical terms, causing noise in the medical terminology results. In future work, the transcripts could be manually corrected to get a cleaner result. A context-aware model and further enhancement of the ontology, potentially also adding abbreviations, would further improve the quality of the BEL results. In future work, it would also be interesting to let patients fill in a questionnaire or do an interview after the consultation to score their understanding of the provided information. These results could be compared to the language complexity and medical jargon use, which can be analyzed following the method we described in this thesis, to research if these language differences actually affected the patient understanding.

Appendices

A Excluded UMLS categories

TUI	Semantic Type Name
T001	Organism
T008	Animal
T010	Vertebrate
T011	Amphibian
T012	Bird
T013	Fish
T014	Reptile
T015	Mammal
T056	Daily or Recreational Activity
T057	Occupational Activity
T064	Governmental or Regulatory Activity
T065	Educational Activity
T066	Machine Activity
T067	Phenomenon or Process
T072	Physical Object
T073	Manufactured Object
T078	Idea or Concept
T079	Temporal Concept
T080	Qualitative Concept
T081	Quantitative Concept
T082	Spatial Concept
T083	Geographic Area
T089	Regulation or Law
T090	Occupation or Discipline
T092	Organization
T093	Health Care Related Organization
T094	Professional Society
T095	Self-help or Relief Organization
T097	Professional or Occupational Group
T168	Food
T170	Intellectual Product
T171	Language

B Top features classifier

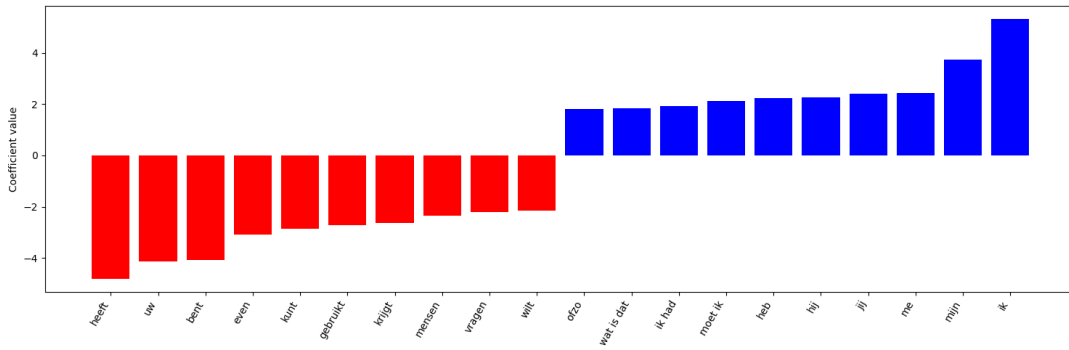


Figure 12: ABEL - Top 10 features for the doctor (red) and patient (blue).

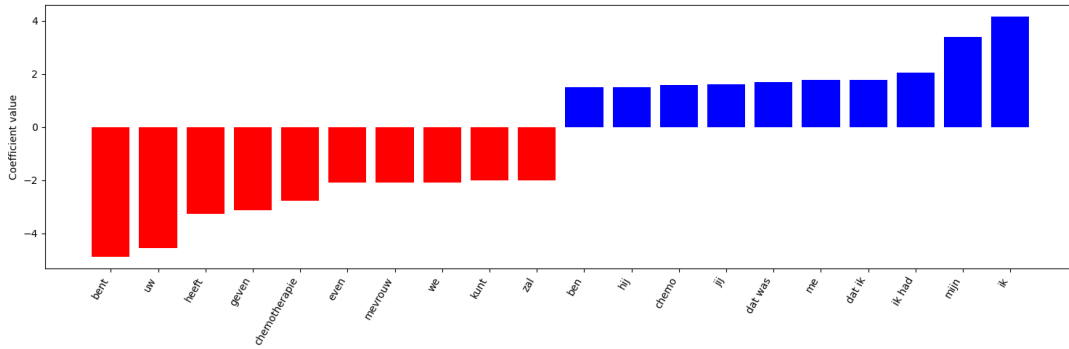


Figure 13: IBIS - Top 10 features for the doctor (red) and patient (blue).

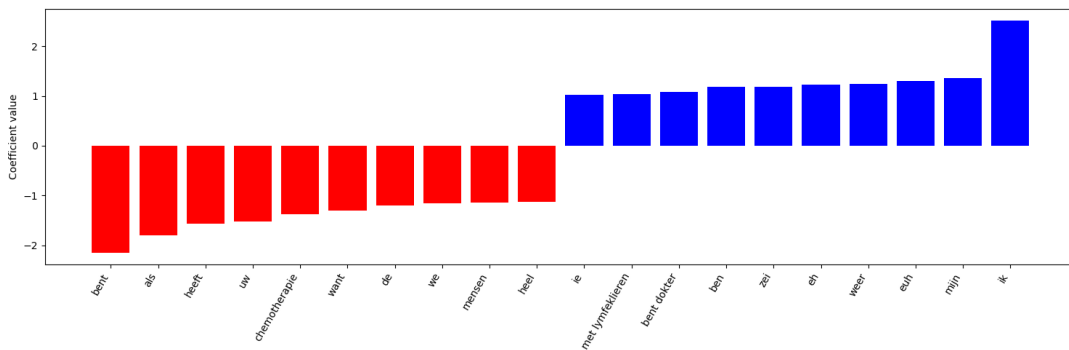


Figure 14: IBIS colon - Top 10 features for the doctor (red) and patient (blue).

References

- [1] Cheryl R Dennison Himmelfarb et al. “Shared decision-making and cardiovascular health: a scientific statement from the American Heart Association”. In: *Circulation* 148.11 (2023), pp. 912–931.
- [2] Alex Waddell et al. “Barriers and facilitators to shared decision-making in hospitals from policy to practice: a systematic review”. In: *Implementation Science* 16.1 (2021), p. 74.
- [3] Heath Pillen et al. “Educating for supported decision making and shared decision making: a scoping review of educational design and outcomes for education and training interventions”. In: *Disability and rehabilitation* 47.2 (2025), pp. 275–286.
- [4] Renata W Yen et al. “Measuring the use of spoken plain language by clinicians in healthcare encounters: a scoping review”. In: *Patient Education and Counseling* (2026), p. 109531.
- [5] Katherine A Allen et al. “Jargon be gone—patient preference in doctor communication”. In: *Journal of patient experience* 10 (2023), p. 23743735231158942.
- [6] Sandra D Vamos et al. “The role of health literacy in patient-physician communication: A scoping review”. In: *Health Behavior and Policy Review* 11.4 (2024), pp. 1624–1654.
- [7] Fons Hartendorp et al. “Biomedical Entity Linking for Dutch: Fine-tuning a Self-alignment BERT Model on an Automatically Generated Wikipedia Corpus”. In: *LREC-COLING 2024* (2024), p. 253.
- [8] Joanne E Jordan et al. “The health literacy management scale (HeLMS): a measure of an individual’s capacity to seek, understand and use health information within the healthcare setting”. In: *Patient education and counseling* 91.2 (2013), pp. 228–235.
- [9] Kristine Sørensen et al. “Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU)”. In: *The European journal of public health* 25.6 (2015), pp. 1053–1058.
- [10] Richard S Safeer and Jann Keenan. “Health literacy: the gap between physicians and patients”. In: *American family physician* 72.3 (2005), pp. 463–468.
- [11] Yuji Kanejima et al. “Impact of health literacy in patients with cardiovascular diseases: a systematic review and meta-analysis”. In: *Patient Education and Counseling* 105.7 (2022), pp. 1793–1800.
- [12] Gildasio S De Oliveira Jr et al. “The impact of health literacy in the care of surgical patients: a qualitative systematic review”. In: *BMC surgery* 15.1 (2015), p. 86.
- [13] Dean Schillinger et al. “Closing the loop: physician communication with diabetic patients who have low health literacy”. In: *Archives of internal medicine* 163.1 (2003), pp. 83–90.
- [14] Youssoufa M Ousseine et al. “Multiple health literacy dimensions are associated with physicians’ efforts to achieve shared decision-making”. In: *Patient education and counseling* 102.11 (2019), pp. 1949–1956.

- [15] Sarah S Nouri and Rima E Rudd. “Health literacy in the “oral exchange”: An important element of patient–provider communication”. In: *Patient education and counseling* 98.5 (2015), pp. 565–571.
- [16] Fatima Al Sayah et al. “Health literacy and nurses’ communication with type 2 diabetes patients in primary care settings”. In: *Nursing Research* 63.6 (2014), pp. 408–417.
- [17] Victoria Charpentier et al. “Say what? Quantifying and classifying jargon use during inpatient rounds”. In: *Hospital Pediatrics* 11.4 (2021), pp. 406–410.
- [18] Ann Neville Miller et al. “Use of seven types of medical jargon by male and female primary care providers at a university health center”. In: *Patient education and counseling* 105.5 (2022), pp. 1261–1267.
- [19] Sarah S Nouri and Rima E Rudd. “Health literacy in the “oral exchange”: An important element of patient–provider communication”. In: *Patient education and counseling* 98.5 (2015), pp. 565–571.
- [20] Victoria Charpentier et al. “Say what? Quantifying and classifying jargon use during inpatient rounds”. In: *Hospital Pediatrics* 11.4 (2021), pp. 406–410.
- [21] George Tewfik et al. “Simulated anesthesia consent discussions demonstrate high level of comprehension and education requirements for patients: A pilot study”. In: *PEC innovation* 2 (2023), p. 100153.
- [22] Elaine Wittenberg-Lyles et al. “Exploring oral literacy in communication with hospice caregivers”. In: *Journal of pain and symptom management* 46.5 (2013), pp. 731–736.
- [23] Daniele Santi et al. “Qualitative and quantitative analysis of doctor–patient interactions during andrological consultations”. In: *Andrology* 10.7 (2022), pp. 1240–1249.
- [24] David Forner et al. “Turn analysis and patient-centredness in paediatric otolaryngology surgical consultations”. In: *Clinical Otolaryngology* 45.5 (2020), pp. 725–731.
- [25] Debra L Roter et al. “Assessing oral literacy demand in genetic counseling dialogue: preliminary test of a conceptual framework”. In: *Social science & medicine* 65.7 (2007), pp. 1442–1457.
- [26] Debra L Roter. “Oral literacy demand of health care communication: challenges and solutions”. In: *Nursing outlook* 59.2 (2011), pp. 79–84.
- [27] James P Meza et al. “Language complexity differs between doctors and patients during contraceptive counseling: A mixed-method study”. In: *Communication & Medicine* 14.1 (2017), pp. 3–14.
- [28] Susan Koch-Weser, Rima E Rudd, and William DeJong. “Quantifying word use to study health literacy in doctor–patient communication”. In: *Journal of health communication* 15.6 (2010), pp. 590–602.
- [29] Brandon Q Tran and Kate Sweeny. “Correlates of physicians’ and patients’ language use during surgical consultations”. In: *Health Communication* (2020).
- [30] Zubair Afzal et al. “ContextD: an algorithm to identify contextual properties of medical terms in a Dutch clinical corpus”. In: *BMC bioinformatics* 15.1 (2014), p. 373.

- [31] Stella Verkijk and Piek Vossen. “MedRoBERTa. nl: a language model for Dutch electronic health records”. In: *Computational Linguistics in the Netherlands*. Vol. 11. Computational Linguistics in the Netherlands. 2021, pp. 141–159.
- [32] Hielke Muizelaar et al. “Extracting patient lifestyle characteristics from Dutch clinical text with BERT models”. In: *BMC medical informatics and decision making* 24.1 (2024), p. 151.
- [33] Isa Spiero et al. “Comparison of local large language models for extraction of signs and symptoms data from electronic health records”. In: *medRxiv* (2025), pp. 2025–12.
- [34] Jenia Kim et al. “Modeling Dutch medical texts for detecting functional categories and levels of COVID-19 patients”. In: *Proceedings of the thirteenth language resources and evaluation conference*. 2022, pp. 4577–4585.
- [35] Evan French and Bridget T McInnes. “An overview of biomedical entity linking throughout the years”. In: *Journal of biomedical informatics* 137 (2023), p. 104252.
- [36] David Kartchner et al. “A comprehensive evaluation of biomedical entity linking models”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023, pp. 14462–14478.
- [37] Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. “DNorm: disease name normalization with pairwise learning to rank”. In: *Bioinformatics* 29.22 (2013), pp. 2909–2917.
- [38] Matthew E Peters et al. “Semi-supervised sequence tagging with bidirectional language models”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2017, pp. 1756–1765.
- [39] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019, pp. 4171–4186.
- [40] Jinhyuk Lee et al. “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”. In: *Bioinformatics* 36.4 (2020), pp. 1234–1240.
- [41] Mujeen Sung et al. “Biomedical entity representations with synonym marginalization”. In: *Proceedings of the 58th annual meeting of the association for computational linguistics*. 2020, pp. 3641–3650.
- [42] Fangyu Liu et al. “Self-alignment pretraining for biomedical entity representations”. In: *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2021, pp. 4228–4238.
- [43] Sheng Zhang et al. “Knowledge-rich self-supervision for biomedical entity linking”. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. 2022, pp. 868–880.
- [44] Susan Koch-Weser, Rima E Rudd, and William DeJong. “Quantifying word use to study health literacy in doctor–patient communication”. In: *Journal of health communication* 15.6 (2010), pp. 590–602.

- [45] P Whittaker. “Do physicians and patients speak different languages: quantitative evidence from linguistic analysis”. In: *European Heart Journal* 45.Supplement_1 (2024), ehae666–3566.
- [46] Valerie A Ubbes and Benedict Njoku. “A curriculum, instruction, and assessment (CIA) framework for health literacy education (HLE) in medical and health professions schools”. In: *World Journal of Social Science Research* 9.1 (2022), pp. 15–55.