Defining Autoimmune Diseases in Expert and Non-Expert Texts

Ana Ostroški Anić[1,*,†], Martina Pavić[2,†]

[1,2] *Institute for the Croatian Language, Ulica Republike Austrije 16, HR-10000 Zagreb, Croatia*

Abstract

The paper analyses the definitions of autoimmune diseases as they are defined in texts of different levels of specialization. Sentences are annotated applying the FrameNet's methodology. The results are discussed in order to verify if FrameNet's annotation procedure can be adequately used to define medical concepts.

Keywords

definitions, medical terminology, non-expert texts, Frame Semantics

1. Introduction

Medicine is one of specialized domains that is of particular interest to different communities of speakers, most of whom cannot be considered experts or even semi-experts, but whose interest in the domain lies simply in the fact that a certain level of medical knowledge is useful in everyday life. It is therefore common to have some medical terms defined in a general dictionary, as many medical terms enter the general vocabulary of a language, whether that be due to their prominent position in non-linguistic context, for various educational purposes or triggered by exceptional events, such was the recent Covid-19 pandemic.

As a prominent characteristic of medical language, terminological variation has been extensively studied [1], [2], [3], focusing mostly on the differences in expertise levels for different speakers, i.e. medical experts and laypeople. The more precise, concise and systematically structured the discourse is, the greater the term density with less term variation. As the degree of specialization decreases, specialized discourse becomes more similar to general discourse in terms of conceptual variation, redundancy, ambiguity, and extensive use of synonyms and paraphrases to explain the concept [4]. The degree of text specialization causes variation in defining concepts, often referred to as contextual variation [5], conceptual variation [6] or vagueness in general language [7]. Following Cabre's seminal theory [8], many argue that the context determines the exact meaning of the term in that context, e.g., San Martín [5], who claims that "the term invokes the same concept, but the activated knowledge differs."

Being of interest for a large population of users, medical concepts related to diseases, conditions, treatments, procedures, etc., are defined and described differently in different contexts and registers, depending on the intended users. The meaning of particular medical concept or its delimiting characteristics remain the same regardless of the context in which the concept is placed, but different characteristics are placed as more prominent depending on the focus of the communicative setting, e.g., the cause of an illness, its symptoms or methods of treating the illness are not always described in the same manner. In other words, if we view a certain disease as a complex conceptual structure consisting of smaller elements, analogue to a semantic frame with its frame elements (FEs), then different elements are in focus depending on the context and the user addressed. This also means that the situation can be framed differently using different terms or

[1,*] Corresponding author.
[†] These authors contributed equally.
aostrosk@ihjj.hr (A. Ostroški Anić); mpavic@ihjj.hr (M. Pavić)
0000-0001-9999-0750 (A. Ostroški Anić); 0000-0001-6061-9495 (M. Pavić)

term variants, which is why there are commonly more terms for one and the same disease. Traditional terminological or analytical definitions, which consist of a superordinate concept and the defined concept's delimiting characteristics, are therefore often replaced with types of definitions that exploit other knowledge patterns, e.g. functional or synonymic [9], and that also underline non-hierarchical relations like those of frame-to-frame relations in FrameNet.

Frame Semantics has been effectively applied in many specialized domains, with biomedical domains comprising a significant portion. These applications serve both to describe specialized knowledge using an established methodological apparatus, and to connect the terminology of a specialized domain to the general vocabulary lexicon [10], [11]. Some of the most recent applications of the FrameNet methodology in the medical and biomedical domains include [12], which evaluates the efficiency of automatic annotation in medication leaflets, and [13], which applies FrameNet analysis to biomedical English to test its applicability.

So far, only one frame-based resource has been developed for Croatian [14], in which the FrameNet methodology was applied with certain adjustments to better accommodate the specificities of the aviation domain. Croatian medical terminology has not yet been described in the framework of Frame Semantics nor has there been an application of the FrameNet methodology to medical texts in Croatian. This paper therefore presents a first such attempt, using the definitions of autoimmune diseases from Croatian texts with varying levels of specialization. The field of autoimmune diseases is chosen as it is a medical field of interest to the general public. Structural and conceptual differences of definitions extracted from two corpora are compared in order to verify if the FrameNet's methodology and annotation procedure can be adequately used to describe medical concepts to non-experts.

The paper is organized as follows: Section 2 describes the corpora used for definition extraction, and the process of definition extraction and validation. In section 3, the FrameNet model is outlined, as well as the semantic frames used in annotation. In Section 4, we discuss the annotation process and compare different definitions for the same diseases, illustrating the issues with examples from corpora. Finally, Section 5 concludes the paper with reflections on the aptness of the annotation process.

2. Methodology

Two specialized corpora in the Croatian language have been used in the analysis: a scientific corpus of medical research papers, consisting of 5,318,395 words, and a corpus of texts taken from medical portals for the general public, consisting of 5,022,639 words. The scientific corpus includes reputable contemporary medical journals taken from *Hrčak*, the Croatian portal of scientific and professional journals. These journals cover various medical fields and include well-known publications such as *Acta Medica Croatica*, *Liječnički vjesnik*, and *Cardiologia Croatica*. The popular corpus consists of texts from widely used online medical portals, such as *ordinacija.hr*, the most extensive and comprehensive database of private practice doctors, and *Cybermed*, Croatia's first health portal, designed for public health education and professional development of medical practitioners. Both corpora had been previously compiled using Sketch Engine tools [15], which were also applied in further analysis and definition extraction.

The first step in the data analysis was to create a list of the 50 most frequent terms for diseases by manually analysing concordances of the Croatian term *bolest* 'disease' in the corpus of texts from medical portals. These 50 terms were then used to query the corpus of scientific medical texts. Due to a broad scope of medical terminology, we focused this analysis on autoimmune diseases for two reasons. First, the high occurrence of terms related to autoimmune diseases among the most frequent terms in both corpora suggested their relevance in both specialized medical texts and texts aimed at the general public. This also confirmed that the rising number of people suffering from autoimmune diseases is accompanied by a growing need for accessible medical information [16]. Second, as a heterogeneous group of medical conditions, each with a complex nature, autoimmune diseases are often defined or explained (as will be seen later) by emphasizing their multifaceted

causes, symptoms, and treatments. This provides a strong foundation for exploring how concept definitions are adapted across different registers and levels of expertise.

Five definitions per each of the 50 most frequent terms were selected for annotation. Additionally, for each term, several extra sentences were taken as explanations that would illustrate the context in which the term is placed and described. In cases where concordances exceeded 1,000 occurrences, a random sample of 300 was used for analysis. The same procedure was then applied to the scientific medical corpus. In total, the dataset comprises around 400 examples, but for this paper, only definitions of autoimmune diseases are analysed – specifically, *celiac disease*, *Hashimoto's thyroiditis*, *rheumatoid arthritis*, *multiple sclerosis*, and *psoriasis* – as these five autoimmune diseases are the most prevalent in the popular medical corpus.

Definitions were annotated following the FrameNet's methodology [17], and using FrameNet's semantic frames related to medicine, with `Medical_conditions` as the starting frame.[2] This approach enabled the identification of frame elements that capture the core characteristics of medical concepts. Verbal patterns and their lexical markers were identified based on the typology outlined in Sierra et al. [9]. These included markers commonly used in definitions to signal conceptual relations and attributes, which helped in distinguishing definitions from explanations, common in popular medical texts. Finally, definitions extracted from the scientific corpus were compared to those extracted from the corpus of texts written for non-experts to assess the level of simplification applied in less specialized texts. This was done by analyzing the terms used, if any, in place of those denoting autoimmune diseases in the scientific corpus and determining whether they were more transparent or closer to general vocabulary than those in the scientific corpus.

As the aim of the analysis was to test whether FrameNet's frames could be successfully applied in the annotation of medical texts in Croatian, frames were applied as they were defined in the Berkeley FrameNet, without any modifications in their original structure in English. Examples given in Section 4 serve to illustrate the challenges we met during the annotation process, e.g., deciding on the suitability of certain FEs in given frames. The results of the annotation served as the basis for proposing elaborations or modifications in the frames, and for creating guidelines for future, more extensive annotation of medical texts.

3. FrameNet annotation

FrameNet is a computational lexical resource built on the theoretical premises of Frame Semantics [18], [19], which exploits the concepts of semantic frame, frame elements, lexical units and frame-to-frame relations [17] in a semantic and syntactic description of English. Each frame consists of core and non-core elements, which have the role of participants, props or other elements in defining the situation or a state represented by the frame. In the `Medical_conditions` frame, which is the central frame in our description of diseases, AILMENT and PATIENT are the core, defining elements, that are needed for the conceptualization of the frame, while BODY_PART, CAUSE, DEGREE, DURATION, NAME, PLACE and SYMPTOM can be instantiated in a sentence, but not necessarily. If we are to make a correlation between a frame-semantic description of a specialized category like medical condition, and a traditional terminological description of it, we could regard the elements of a frame as delimiting and non-delimiting characteristics of a defined concept.

3.1. FrameNet frames used in annotation

As previously stated, the `Medical_conditions` frame was the primary frame used in the annotation and analysis of medical definitions, given that it is used to define medical conditions or diseases from which a patient suffers or for which is being treated. However, although the frame contains elements to denote the affected individuals, as well as the cause and degree of the

[2] Following established conventions, frame names are written in `Courier New`, while frame elements are set in SMALL CAPS.

condition, it represents rather a general conceptualization of a medical condition, and we soon noticed its' limitations in terms of a more specific medical representation.

The FEs of the `Medical_conditions` frame reflect the fundamental semantic structure underlying most medical discourse. Instead of this general view, many definitions in both corpora included elements that are specific to autoimmune diseases, and that could not have been adequately represented by the `Medical_conditions` frame. For instance, the *progression of the disease* or *prevalence in population* were characteristics mentioned in certain definitions, which lacked corresponding elements in the `Medical_conditions` frame. Popular texts frequently use paraphrases, analogies, and simplified language that emphasize communicative clarity over medical precision found in scientific texts. These texts sometimes highlight features like preventive measures or lifestyle advice, which are outside the scope of traditional medical context, but that are nevertheless prominent in descriptions of the diseases in texts written for medical portals.

After a set of test annotation conducted on a sample of 30 examples, we opted for using additional frames from FrameNet that are related to the medical domain: `Condition_symptom_relation`, `Cure`, `Medical_instruments`, `Medical_interaction_scenario`, `Medical_intervention`, and `Medical_professionals`. By adding these to the `Medical_conditions` frame, a more complete and accurate analysis of both scientific and non-expert-oriented medical texts was ensured, considering the complexity and variability of medical terminology across different registers.

4. Results of annotation

Since FrameNet is not a specialized resource, its medical frames have their limitations when annotating medical texts. Additionally, choosing a different frame to annotate the same example can sometimes result in small but meaningful differences in data analysis. E.g., if the definition of *celiac disease* is annotated with regards to the `Medical_conditions` frame, as in example (1a), patients are identified as "genetically predisposed individuals":[3]

(1a) [HR] CELIJAKIJA se pojavljuje kod [genetski sklonih pojedinaca PATIENT] čija prehrana sadrži [gluten [CAUSE], ali i kao posljedica [infekcija i stresa CAUSE].
[EN] CELIAC DISEASE occurs in [genetically predisposed individuals PATIENT] whose diet includes [gluten CAUSE], but also as a result of [infections and stress CAUSE].

Whereas, if we apply the `Condition_symptom_relation` frame, we are able to use the FE INFLUENCE, which is used for identifying genetic, biological, and environmental influences that affect medical conditions:

(1b) [HR] CELIJAKIJA se pojavljuje kod [genetski sklonih INFLUENCE] [pojedinaca PATIENT] čija prehrana sadrži [gluten CAUSE], ali i kao posljedica [infekcija i stresa CAUSE].
[EN] CELIAC DISEASE occurs in [genetically predisposed INFLUENCE] [individuals PATIENT] whose diet includes [gluten CAUSE], but also as a result of [infections and stress CAUSE].

The difference between two annotations may be subtle as in (1a) and (1b), but it is more often than not that significant information would be lost if the frame `Condition_symptom_relation` was not applied, as in the following example in which all the possible influences of psoriasis are listed:

(2) [HR] [Simptomi SYMPTOM] [psorijaze MEDICAL_CONDITION] SE POJAVLJUJU [periodično TIME] i [osobito su izraženi EXTENT] pod utjecajem određenih faktora, kao što su: [hladnije vrijeme INFLUENCE], [infekcije INFLUENCE], [ozljede kože INFLUENCE], [neki lijekovi INFLUENCE], [stres INFLUENCE], [pušenje INFLUENCE] i [alkohol INFLUENCE].

[3] For ease of reference, all definitions are given both in the Croatian original and in English translations.

[EN] The [symptoms SYMPTOM] of [psoriasis MEDICAL_CONDITION] OCCUR [periodically TIME] and are [particularly pronounced EXTENT] under the influence of certain factors, such as [colder weather INFLUENCE], [infections INFLUENCE], [skin injuries INFLUENCE], [certain medications INFLUENCE], [stress INFLUENCE], [smoking INFLUENCE], and [alcohol INFLUENCE].

As opposed to a sentence in which *psoriasis* would be the lexical unit that is the target word of the annotation, in (2) it is *occur* that is the target lexical unit, which immediately evokes the frame `Condition_symptom_relation`, used to define the symptoms of the disease, the period over which they occur, as well as their possible origins. Although symptoms are annotated in this frame, there was no FE in the original frame to identify the manner of symptoms' occurrence, which was needed for the symptoms of diabetes, as expressed by the adverb *naglo* 'suddenly' in (3):

(3) [HR] Kod [dijabetesa tipa 1 MEDICAL_CONDITION] [simptomi SYMPTOM] se obično POJAVLJUJU [*naglo* MANNER], [unutar nekoliko dana ili tjedana TIME].

[EN] 'In [type 1 diabetes MEDICAL_CONDITION], [symptoms SYMPTOM] usually APPEAR [*suddenly* MANNER], [within a few days or weeks TIME].'

In neither of the frames used, there is no frame element to identify an indicator or value, e.g., in the following sentence, where the blood sugar level is a relevant piece of information:

(4) [HR] Ako patite od dijabetesa tipa 1, kontrola *razine šećera u krvi* se može nešto razlikovati u odnosu na osobe koje pate od dijabetesa tipa 2.

[EN] 'If you have type 1 diabetes, *blood sugar level* management may differ somewhat compared to individuals with type 2 diabetes.'

For some definitions, a deeper understanding of the characteristics of the disease is needed in order to choose the right frame or its element. In example (5), the Croatian adjective *upalni* 'inflammatory' in the definition of multiple sclerosis can be interpreted in at least two ways:

(5) [HR] *Multipla skleroza* (MS) je [kronična DURATION] [upalna demijelinizacijska bolest AILMENT] [središnjeg živčanog sustava BODY_SYSTEM] (SŽS).

[EN] '*Multiple sclerosis* (MS) is a [chronic DURATION] [inflammatory demyelinating disease AILMENT] of the [central nervous system BODY_SYSTEM] (CNS).'

From the definition, it is not clear whether multiple sclerosis causes inflammation or it results from inflammation.

The choice of frame can influence the identification of concepts superordinate to the defined disease, which presents a challenge for consistent annotation. The *type_of* hierarchical relation is one of the key relations in any terminological conceptual system, which in FrameNet corresponds to the *inherits_from* frame-to-frame relation. In the `Medical_conditions` frame, AILMENT is used to identify the type of a medical problem the defined condition or disease belongs to, whereas in `Condition_symptom_relation`, this superordinate concept is defined by the element MEDICAL_CONDITION. It would appear that the FEs are placed in a relation one to the other, but since the FE MEDICAL_CONDITION is defined as "a holistic description of the medical state of the PATIENT", it is obvious that there is no intended relation between AILMENT and MEDICAL_CONDITION, but that these are rather used simultaneously for the same semantic role in different frames. Let's illustrate this complexity on the definitions of the celiac disease:

6. [HR] CELIJAKIJA je [autoimuna bolest AILMENT], a ne alergija ili intolerancija na određenu vrstu hrane.
   [ENG] CELIAC DISEASE is an [autoimmune disease AILMENT], not an allergy or intolerance to a specific type of food.
7. [HR] CELIJAKIJA je [česta FREQUENCY] [kronična EXTENT] [autoimuna bolest MEDICAL_CONDITION] koja se javlja u [1% FREQUENCY] [zapadne populacije GROUP].

[ENG] CELIAC DISEASE is a [common FREQUENCY], [chronic EXTENT] [autoimmune disease MEDICAL_CONDITION] that occurs in [1% FREQUENCY] of the [Western population GROUP].

8. [HR] CELIJAKIJA je zapravo [autoimuna bolest AILMENT] u kojoj imunološki sustav napada [stanice crijeva BODY_PART] nakon što u njih uđe [gluten CAUSE].
   [ENG] CELIAC DISEASE is essentially an [autoimmune disease AILMENT] in which the immune system attacks [the cells of the intestine BODY_PART] after [gluten CAUSE] enters them.
9. [HR] [Autoimuna bolest AILMENT] koju karakterizira [nepodnošenje glutena SYMPTOM] već je desetljećima u liječničkim krugovima i među [oboljelima PATIENT] poznata pod nazivom CELIJAKIJA ili [glutenska enteropatija NAME].
   [ENG] [An autoimmune disease AILMENT] characterized by [gluten intolerance SYMPTOM] has been known for decades in medical circles and among [patients PATIENT] as CELIAC DISEASE or [gluten enteropathy NAME].
10. [HR] CELIJAKIJA ili [glutenska enteropatija NAME] je [autoimuna bolest AILMENT] [probavnog sustava BODY_PART] koja podrazumijeva [trajno i doživotno DURATION] [nepodnošenje glutena SYMPTOM] s [različitim stupnjevima DEGREE] [oštećenja sluznice tankog crijeva MEDICAL_CONDITION] i [širokim spektrom DEGREE] [kliničkih simptoma SYMPTOM].
    [ENG] CELIAC DISEASE, or [gluten enteropathy NAME], is [an autoimmune disease AILMENT] of the [digestive system BODY_PART] characterized by a [permanent and lifelong DURATION] [intolerance to gluten SYMPTOM], with [varying degrees DEGREE] of [damage to the small intestine lining MEDICAL_CONDITION] and a [wide range DEGREE] of [clinical symptoms SYMPTOM].
11. [HR] Na temelju iznesenog razvidno je da je CELIJAKIJA [složena bolest MEDICAL_CONDITION] determinirana [pojedinačnim i međusobnim utjecajem velikog broja gena INFLUENCE] i da se može manifestirati u [svakoj dobi AGE] i s [vrlo varijabilnim, širokim rasponom simptoma SYMPTOM].
    [ENG] Based on the above, it is evident that CELIAC DISEASE is a [complex disease MEDICAL_CONDITION] determined by the [individual and interrelated influence of a large number of genes INFLUENCE]. It can manifest at [any age AGE] and with a [highly variable, broad range of symptoms SYMPTOM].

Examples (6), (8), (9), and (10) are annotated according to the `Medical_conditions` frame, whereas examples (7) and (11) by using the `Condition_symptom_relation` frame. The definition in (7) was annotated using `Condition_symptom_relation` because the elements FREQUENCY, GROUP, and EXTENT are not included in `Medical_conditions`. Since we applied this frame, another dilemma arose regarding whether to use the element AILMENT or MEDICAL_CONDITION. Although verbs are not the target lexical units in these frames, in some sentences, they help establish relations between frame elements. For instance, in (9), the Croatian verb *karakterizirati* 'to *characterize*' links the superordinate concept AILMENT to gluten intolerance as its SYMPTOM. In any case, it is difficult for a non-expert to say whether *gluten intolerance* is the symptom or the actual ailment.

As stated above, sentences that were not considered the best fit for defining a disease but still contained relevant information were also extracted and annotated as contexts for concept information. The following sentence is one such example, where the `Cure` frame is used in annotation due to its elements MEDICATION and TREATMENT:

(12) [HR] Jedini je [lijek MEDICATION] za [oboljele PATIENT] od CELIJAKIJE [bezglutenska dijeta TREATMENT] koje se moraju pridržavati [cijeli život DURATION].
[EN] The only [cure MEDICATION] for [individuals PATIENT] with CELIAC DISEASE is [a gluten-free diet TREATMENT], which they must follow [for their entire life DURATION].

The `Medical_conditions` frame contains an element NAME that is used to identify the name of the medical condition, e.g. *Crohn's disease*. In some definitions, however, when the term of the disease is the target lexical unit, it was not clear if the element NAME should be used for both the synonym and the term of the disease, and in that case, which should be the main term:

(13) [HR] HASHIMOTOV, ili preciznije nazvan [kronični autoimuni tireoiditis NAME] je [autoimuna bolest AILMENT] [štitnjače BODY_PART], a u današnje je vrijeme glavni uzrok [poremećaja funkcije štitnjače RESULT].
[EN] HASHIMOTO'S, or more precisely called [chronic autoimmune thyroiditis NAME], is [an autoimmune condition AILMENT] of [the thyroid BODY_PART], and today it is the main cause of [thyroid dysfunction RESULT].

These doubts appear because FrameNet annotation is not designed to serve as a method for terminology extraction, although in (13), it yielded a term variant of the target term *Hashimoto's*.

4.1. Adapting the frames' structure

Another medical frame from FrameNet is `Medical_intervention`, which was less used in the annotation, but since it contains the element RESULT, it was the reference frame for each sentence containing the result of a certain medical intervention. It so happens that in certain examples the elements of a frame are not sufficiently precise or quite apt to be used. For example, whenever a sentence carries an expression of *potential* realization of certain semantic roles, e.g. risk factors for the development of rheumatoid arthritis, one is not certain whether that element could be annotated as CAUSE. Similary, in the sentence *Multipla skleroza može uzrokovati slabost mišića ili grčeve zbog kojih je teško hodati.* 'Multiple sclerosis can cause muscle weakness or spasms that make walking difficult,' *muscle weakness* and *spasms* were not annotated as the result of multiple sclerosis, but rather as its CONSEQUENCE because multiple sclerosis directly causes these effects.

When referring to the element EXPLANATION, found in the *Condition_symptom_relation* frame, there are definitions where it is explicitly stated. In other instances, the context had to be closely examined to make sure the right element was used. Example (14) contains a clear explanation for the occurrence of a SYMPTOM or MEDICAL_CONDITION:

(14) [HR] MULTIPLA SKLEROZA je [sporo napredujuća MANNER] [bolest AILMENT] [središnjeg živčanog sustava BODY_SYSTEM] pri kojoj [imunitet uništava ovojnicu koja prekriva živce EXPLANATION].
[EN] MULTIPLE SCLEROSIS is a [slow-progressing MANNER] [disease AILMENT] of [the central nervous system BODY_SYSTEM] in which [the immune system destroys the sheath that covers the nerves EXPLANATION].

In example (15) the situation is more complex:

(15) [HR] MULTIPLA SKLEROZA je jedna od najčešćih [neuroimunoloških bolesti AILMENT] [središnjeg živčanog sustava BODY_SYSTEM] današnjice – [kronična DURATION] [upalna demijelinizacijska bolest AILMENT] [središnjeg živčanog sustava BODY_SYSTEM] ([mozga i kralježnične moždine BODY_PART]), obilježena [propadanjem mijelinske ovojnice živčanih vlakana autoimunom reakcijom EXPLANATION].
[EN] MULTIPLE SCLEROSIS is one of the most common [neuroimmunological diseases AILMENT] of [the central nervous system BODY_SYSTEM] today – a [chronic DURATION] [inflammatory demyelinating disease AILMENT] of [the central nervous system BODY_SYSTEM] ([brain and spinal cord BODY_PART]), characterized by [the degeneration of the myelin sheath of nerve fibres due to an autoimmune reaction EXPLANATION].

It is clear that the degeneration occurs during the disease, but the question is whether this should be identified as an EXPLANATION of how the condition progresses or by another element. Given that the sentence describes the degeneration of the myelin sheath as a characteristic of multiple sclerosis caused by an autoimmune reaction, this could be seen as an EXPLANATION. However, it could also be viewed as a CAUSE because the autoimmune reaction is the cause of the degeneration or even as a CONSEQUENCE. If we decide to focus on the result of the autoimmune response, CONSEQUENCE or CAUSE might be better suited in the above example. What is also clear from example (15) is that elements from more than one frame are used, and the element of BODY_SYSTEM is used along the element BODY_PART. This is justified by the very content of the

example, as well as the decision to enrich the FrameNet frames. The central nervous system is indeed a system, unlike the brain and spinal cord, which are body parts and given in brackets as an elaboration of the main information. For the annotation of (14) and (15) as well as similar examples, the `Medical_conditions` frame was applied but enriched FEs EXPLANATION, BODY_SYSTEM and MANNER.

Sentence (16) is another example where the frames `Cure`, `Medical_conditions` and `Medical_professionals` are combined to be able to annotate all the elements of the sentence.

(16) [HR] Lijekovi koji mijenjaju tok REUMATOIDNOG ARTRITISA, [antireumatici $_{\text{MEDICATION}}$], trebali bi se primjenjivati [rano $_{\text{TIME}}$] i [agresivno]$_{\text{MANNER}}$], čim se primijete [prvi znaci bolesti $_{\text{SYMPTOM}}$], tvrde [stručnjaci $_{\text{PROFESSIONAL}}$].
[EN] Drugs that modify the course of RHEUMATOID ARTHRITIS, [antirheumatic drugs $_{\text{MEDICATION}}$], should be administered [early $_{\text{TIME}}$] and [aggressively $_{\text{MANNER}}$], as soon as [the first signs of the disease $_{\text{SYMPTOM}}$] are noticed, [experts $_{\text{PROFESSIONAL}}$] claim.

5. Differences in definitions from the scientific corpus and the popular corpus

The definitions in the scientific corpus and the medical portals (or popular) corpus exhibit notable differences in their structure, vocabulary used, and the level of detail. Table 1 (in the Appendix) gives examples of terminological definitions from both corpora, where the superordinate concept is underlined, and verbal lexical markers are written in Italics. Other autoimmune diseases apart from the ones analysed in the previous sections are also given as illustrative examples.

It can be observed that the definitions of diseases in the scientific and popular corpora are based on different superordinate concepts. The scientific corpus is characterized by internationalisms, with terms such as *kardiovaskularni* 'cardiovascular', *infektivni* 'infectious', *koronarni* 'coronary', and *maligni* 'malignant' being more commonly used by medical professionals, while in the corpus of popular texts, the superordinate concepts are closer to general language to ensure they are understandable to the broader audience for whom the texts are intended. To illustrate this, in the scientific corpus, Gaucher's disease is classified under the superordinate concept *autosomal recessive disease* 'autosomno recesivna bolest', whereas in the popular corpus, it is described as a *rare hereditary disease* 'rijetka nasljedna bolest', which is not the exact equivalent as it highlights a different aspect of the condition. Leptospiroza is classified under *zoonoses* 'zoonoza' in the scientific corpus; in the popular corpus the superordinate concept is *infectous disease* 'zarazne bolesti'. The definition of *psoriasis* 'psorijaza' is a *chronic relapsing inflammatory disease* 'kronično recidivirajuća upalna bolest', in contrast to a more understandable *skin disorder* 'kožni poremećaj'.

Based on the provided examples, several key differences between definitions in the scientific and popular corpora can be observed. These differences relate to:

- Specificity and detail: in the popular corpus, terms that are part of general vocabulary are often used, resulting in simplified definitions, as seen in the following example: *Reumatoidni artritis je autoimuna bolest koja uzrokuje nastanak upale.* 'Rheumatoid arthritis is an autoimmune disease that causes inflammation'. In contrast, the scientific corpus contains more specific terms and specialized vocabulary. For example, leptospirosis is described in greater detail in the scientific corpus, specifying its causative agents as *pathogenic spiral bacteria of the genus Leptospira spp.*, while the popular corpus provides a less specific definition, mentioning terms like *summer flu* or *harvest fever*, which are more familiar to the general audience;
- Contextualization and target audience: scientific definitions often contain specific information relevant to professionals in the medical field, such as the pathophysiological processes, enzymes, or bacterial strains, providing a deeper understanding of the disease.

On the other hand, popular definitions tend to avoid highly specialized language to be more accessible to the broader public. This includes using more recognizable terms, such as *summer flu* for leptospirosis;

- Content and expansion of description: scientific definitions include details about causes, pathophysiology, disease progression, and specific characteristics. The scientific definition of psoriasis, for example, mentions the disease's specific mechanisms (i.e., keratinocyte differentiation, apoptosis) and the precise skin regions affected, offering a very detailed insight into the nature of the disease. The popular corpus, in contrast, tends to simplify the description of symptoms and focus on basic information, such as accelerated cell growth on the skin, without delving into the disease's underlying mechanisms;
- Terms that are missing or replaced: medical terms like *autosomal recessive disease*, *demyelinating disease*, or *autoimmune reaction* are often replaced with more transparent term variants in popular definitions, or with terms denoting broader categories like *rare inherited disease*, *skin disorder*, or *inflammation*.

6. Conclusions

Based on this research, it is possible to establish common patterns in the way diseases are defined in texts aimed at lay audiences. Unlike scientific corpora, lay-oriented texts prioritize simplicity, accessibility, and relatability to ensure the content is comprehensible to non-specialists. The following characteristics emerge as defining features of these texts:

- Lay texts tend to replace or avoid specialized medical terminology, opting for everyday language and general descriptions;
- Definitions in lay texts emphasize symptoms that are immediately observable or relatable to everyday experiences. These texts often link the disease to its practical implications on daily life, helping readers understand its relevance;
- Compared to scientific definitions, definitions in lay-oriented texts often omit pathophysiological details or genetic explanations, presenting only the most essential aspects of the disease;
- Non-Scientific texts frequently use analogies, alternative terms, or simplified explanations to clarify medical concepts;
- Diseases are often broadly categorized, such as describing psoriasis as a skin disorder;
- Texts for non-experts tend to highlight actionable insights, such as treatment options or lifestyle changes, to empower readers.

By identifying these patterns, it is possible to develop a structured approach to writing accessible medical definitions. This is particularly beneficial for patient education and public health communication, ensuring that information about diseases is not only accurate but also understandable to broader audiences. This analysis confirms the assumption that a methodology developed for lexicographic purposes, and used to define general language vocabulary can be applied in a specialized context with certain modifications, such as adding FEs to the existing frame structure (e.g., `Medical_conditions`), or using FEs from several related frames to annotate texts. The results of this annotation task will be used to modify medical semantic frames in the Croatian version of FrameNet, while the definitional patterns identified will aid in extracting definitions of other medical concepts for the purpose of creating a dataset of expert and non-expert definitions of medical concepts.

The dataset and typology of definitional patterns will support text simplification experiments and other NLP tasks aimed at developing efficient methods for creating terminological resources for non-experts and the general public. Understanding how different medical conditions are presented and explained in layperson-oriented materials is crucial for improving public awareness and ensuring clear, accurate communication of medical information.

Acknowledgements

This work was created as part of the project *Semantic Frames in the Croatian Language (SEF)* funded by the European Union – NextGenerationEU.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 in order to translate examples into English. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] L. Bowker, S. Hawkins, Variation in the Organization of Medical Terms: Exploring Some Motivations for Term Choice, Terminology 12/1 (2006) 79–110. https://doi.org/10.1075/term.12.1.05bow

[2] M. Lončar, A Ostroški Anić, Eponymous medical terms as a source of terminological variation, in: G. Budin, V. Lušicky (Ed.), Languages for Special Purposes in a Multilingual, Transcultural World, Proceedings of the 19th European Symposium on Languages for Special Purposes, University of Vienna, Vienna, 2014, pp. 36–44.

[3] M. Tercedor Sánchez, C. I. López-Rodríguez, Access to Health in an Intercultural Setting: The Role of Corpora and Images in Grasping Term Variation, Linguistica Antverpiensia 11 (2012) 247–268. https://doi.org/10.52034/lanstts.v11i.306.

[4] T. Cabré Castellví, 1998, in J. Freixa, Causes of Denominative Variation. A Typology Proposal, Terminology 12/1 (2006) 51–77.

[5] A. San Martín, A Flexible Approach to Terminological Definitions: Representing Thematic Variation, International Journal of Lexicography 35/1 (2022) 53–74. doi:10.1093/ijl/ecab013.

[6] J. Freixa, S. Fernández-Silva, Terminological Variation and the Unsaturability of Concepts, in: P. Drouin, A. Francoeur, J. Humbley, A. Picton (Ed.), Multiple Perspectives on Terminological Variation, John Benjamins, Amsterdam, 2017, pp. 155–180. doi:10.1075/tlrp.18.

[7] D. Geeraerts, Vagueness's puzzles, polysemy's vagaries, Cognitive Linguistics 4 (1993). 223–272. doi: 10.1515/cogl.1993.4.3.223.

[8] T. Cabré Castellví, Terminology. Theory, methods and applications, John Benjamins Publishing Company, Amsterdam, 1999.

[9] G. Sierra, R. Alarcón, C. A. Aguilar, C. Bach, Definitional verbal patterns for semantic relation extraction, Terminology 14 (2008) 74–98. https://doi.org/10.1075/term.14.1.05sie.

[10] M. C. L' Homme, C. Subirats, B. Robichaud, A proposal for combining "general" and specialized frames, in: Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex – V), Osaka, Japan, 2016, pp. 156–165. https://aclanthology.org/W165321.

[11] M. C. L 'Homme, B. Robichaud, C. Subirats, Building multilingual specialized resources based on FrameNet: application to the field of the environment, in: T. Torrent, C. F. Baker, O. Czulo, K. Ohara, M. R. L. Petruck (Eds.), International FrameNet Workshop 2020. Towards a Global, Multilingual FrameNet, Marseille, 2020, pp. 85–92.

[12] M. Gamonal, A. Pagano, T. Torrent, E. Matos, A. Lorenzi, Automated semantic frame annotation. An Exploratory Study in the Health Domain, in: K. Despot, A. Ostroški Anić, I. Brač (Eds.), Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress, Institute for the Croatian Language, Zagreb, 2024, pp. 67–80.

[13] E. Castaño, I. Verdaguer Clavera, Frame semantics in the lexical database SciE-Lex, Terminology 30/2 (2024) 190 – 215. https://doi.org/10.1075/term.22035.cas.

[14] A. Ostroški Anić, I. Brač, AirFrame. Mapping the field of aviation through semantic frames, in A. Klosa-Kückelhaus, S. Engelberg, C. Möhrs, P. Storjohann (Eds.), Dictionaries and Society. Proceedings of the XX EURALEX International Congress, Manheim, 2022, pp. 334–345.

[15] A. Kilgarriff et al, The Sketch Engine: ten years on, Lexicography 1/1 (2014) 7–36. http://doi.org/10.1007/s4060701400099.
[16] M. Velasquez_Manoff, An Epidemic of Absence: A New Way of Understanding Allergies and Autoimmune Diseases, Scribner, New York, 2013.
[17] J. Ruppenhofer, M. Ellsworth, M. R. L. Petruck, C. R. Johnson, C. F. Baker, J. Scheffczyk, FrameNet II: Extended Theory and Practice, Revised November 1, 2016.
[18] C. J. Fillmore, Frame semantics, in: Linguistic society of Korea (Ed.), Linguistics in the morning calm, Hanshin Publishing Co, 1982, pp. 111–137.
[19] C. J. Fillmore, Frames and the semantics of understanding, Quaderni di Semantica 6/2 (1985) 222–254.

Appendix

Table 1. Comparison of definitions from two corpora

concept	scientific corpus	medical portals corpus
Gaucherova bolest 'Gaucher's disease'	[HR] Gaucherova bolest autosomno je recesivna bolest *koju karakteriziraju* snižene vrijednosti enzima glukocerebrozidaze u lizosomima.	[HR] Gaucherova bolest je rijetka nasljedna bolest *koja* zbog nedostatka enzima *uzrokuje* nakupljanje tvari u stanicama.
	[EN] Gaucher's disease is an autosomal recessive disease *characterized by* reduced levels of the enzyme glucocerebrosidase in lysosomes.	[EN] Gaucher's disease is a rare inherited disease that, due to enzyme deficiency, *leads to* the accumulation of substances in cells.
leptospiroza 'leptospirosis'	[HR] Leptospiroza je jedna od globalno najraširenijih zoonoza *uzrokovana* patogenim spiralnim bakterijama iz roda Leptospira spp.	[HR] Leptospiroza (ljetna gripa, žetvena/vodena/muljna groznica) *spada u* zarazne bolesti životinja i čovjeka.
	[EN] Leptospirosis is one of the globally widespread zoonoses *caused by* pathogenic spiral bacteria of the genus *Leptospira spp*.	[EN] Leptospirosis (summer flu, harvest fever, water/mud fever) *belongs to* infectious diseases affecting both animals and humans.
psorijaza 'psoriasis'	[HR] Psorijaza je kroničnorecidivirajuća upalna bolest koja je *obilježena* poremećajem diferencijacije i proliferacije keratinocita te sniženom apoptozom keratinocita unutar epidermisa.	[HR] Psorijaza je kožni poremećaj *koji uzrokuje* ubrzani razvoj stanica na površini kože.
	[EN] Psoriasis is a chronic relapsing inflammatory disease *characterized by* a disturbance in the differentiation and proliferation of keratinocytes and reduced apoptosis of keratinocytes within the epidermis.	[EN] Psoriasis is a skin disorder *that causes* accelerated development of cells on the skin's surface.
reumatoidni artritis 'rheumatoid arthritis'	[HR] Kako je reumatoidni artritis kronična upalna bolest koja *često rezultira* progresivnom disfunkcijom zglobova, tijekom bolesti bolesnici mogu razviti kompresiju perifernog živca što se opisuje kao neurološko pogoršanje u sklopu reumatoidnog artritisa.	[HR] Reumatoidni artritis je *teška, progresivna* i kronična bolest cijeloga organizma, *najizraženija je* na zglobovima, ali promjene mogu biti prisutne i na koži i potkožnom tkivu, mišićima, plućima, srcu, krvnim žilama.
	[EN] As rheumatoid arthritis is a chronic inflammatory disease that *often results in* progressive joint dysfunction, patients may develop peripheral nerve compression during the course of the disease, which is described as neurological deterioration within the context of rheumatoid arthritis.	[EN] Rheumatoid arthritis is a *severe*, *progressive*, and chronic disease that affects the entire body, *most pronounced in* the joints, but changes can also be present in the skin and subcutaneous tissue, muscles, lungs, heart, and blood vessels.