Building a Frame-Semantic Model of the Healthcare Domain: Towards the identification of gender-based violence in public health data

Lívia Dutra[1,2], Arthur Lorenzi[1], Lorena Larré[1], Frederico Belcavello[1], Ely Matos[1], Amanda Pestana[1], Kenneth Brown[1], Mariana Gonçalves[1], Victor Herbst[1], Sofia Reinach[3], Renato Teixeira[3], Pedro de Paula[3], Alessandra Pellini[4], Cibele Sequeira[5] , Ester Sabino[4], Fábio Leal[4], Mônica Conde[4], Regina Grespan[5] , Tiago Torrent[1,6]

[1] Universidade Federal de Juiz de Fora (UFJF)

[2] Göteborgs Universitet (GU)

[3] Vital Strategies Brasil

[4] Universidade Municipal de São Caetano do Sul (USCS)

[5] Secretaria da Saúde do Município de São Caetano do Sul

[6] Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)

{livia.dutra,arthur.lorenzi, lorena.tasca,mariana.goncalves, victor.herbst,amanda.pestana, kenneth.cyrill}@estudante.ufjf.br, {fred.belcavello, ely.matos}@ufjf.br, {sreinach, rteixeira, pcbpaula}@vitalstrategies.org {acgpellini,cibelesequeira}@gmail.com, sabinoec@usp.br, fabio.leal@inca.gov.br, monica.tilli@online.uscs.edu.br, regina.maura@saocaetanodosul.sp.gov.br,tiago.torrent@ufjf.br

Abstract. *Public data systems gather different information about Brazilian citizens. Such information is inserted in the system both via the selection of parameterized options and via open text fields. In this paper we describe the effort of modeling semantic frames for the lexicon of the healthcare domain as a means of tagging the open text fields in public health data to make them more easily interpretable by machine learning systems. This effort is one of the steps in a larger project aiming at using data science and machine learning techniques for the identification of territories prone to suffer from gender based-violence. The modeling effort currently covers 1,787 lexical units in the healthcare domain in Brazilian Portuguese, distributed in 29 semantic frames.*

1. Introduction

According to the World Health Organization, one in three women[1] has been a victim of physical or sexual violence by their partner at some point in their lives. In Brazil, the notification of violence cases by healthcare professionals in SINAN[2] is mandatory. However, underreporting is a serious problem in tackling Gender-Based Violence (GBV). Reasons for underreporting described in the literature [GARBIN et al., 2015; KIND et al., 2013] include excessive workload of healthcare teams, lack of knowledge about the importance of the notification process, fear of possible retaliation by the aggressors, and, finally, difficulties identifying that the injuries and other health conditions are related to a violent episode. On the other hand, victims of GBV, when

[1] https://www.who.int/publications/i/item/9789241564793.

[2] SINAN is the Brazilian national information system for the notification of violence and diseases.

seeking healthcare services, may be included in other public health systems, such as e-medical records, SIM[3] and SIH[4].

Some of the Brazilian healthcare information systems feature, on top of parameterized data fields—that is, those where data is inserted via the selection of one option from a closed list—open text fields. So far, information present in those open text fields has been of little to no use in studies tackling GBV and other public health issues. The reasons for this lack of use of open text fields relate to the fact that language form can be ambiguous, polysemous and highly variable. Therefore, as a means to represent the semantics in the text and make the information present in open fields of public health systems more suitable for large scale data analyses, metadata can be associated to the text form. In this paper, we present the first part of an effort of modeling, in terms of Frame Semantics [FILLMORE, 1982] and using the FrameNet Brasil database structure [TORRENT et al., 2022], the lexical domains of Healthcare and Violence so as to represent the semantics of the linguistic forms present in the open text fields of national information systems.

The effort is part of a larger project aimed at using data integration and textual analysis to identify patterns that suggest that women registered in the health systems are victims of violence. Identification of patterns of GBV will be treated at the level of the territories where candidate victims live, and the resulting system will not keep present information on individuals. Hence, the main goal is to better equip policy makers, local authorities and health teams acting on said territories to design and apply public policies for both raising awareness and eventually reducing GBV in the territories. The hypothesis motivating the work presented in this paper is that data present in medical records, when linked to those in other databases and properly analyzed for their semantic content, can contribute to the identification of augmented risk of gender-based violence (GBV) at a given territory.

In this paper, we will report how frames in the Healthcare domain were modeled, from the first contact with the corpora, through their compilation, analysis of the lexicographic affordances of words and subsequent clusterization of them. We also approach the creation of frames from the most cohesive clusters, the establishment of relationships between frames and the association of lexical units to the frames created for lexicographic annotation.

2. The FrameNet Model

FrameNet[5] is a lexicographic resource that originally applied the theory of Frame Semantics [FILLMORE, 1982] to the analysis of the lexicographic affordances of lexical items in English. From the original project, founded in 1997, other framenets have been developed for several languages, including Brazilian Portuguese [TORRENT et al., 2022].

The foundational principle behind any FrameNet analysis is the one according to which "meaning is relativized to scenes" [FILLMORE, 1977]. This is to say that, for every lexical item in any given language, the meaning of such an item is a function of a background scene defined in terms of the participants and props taking part in it. As an

[3] SIM is the Brazilian mortality information system.
[4] SIH is the Brazilian hospital admission system.
[5] https://framenet.icsi.berkeley.edu/

example, consider a lexical item such as *arthritis.n*. To properly understand the meaning of this lexical item, we have to consider a scene—or frame—where an Ailment affects a Patient. Those two participants are necessary for the frame to be instantiated and, therefore, are the core frame elements (FEs) in the `Health_conditions` frame, shown in Figure 1. Other FEs may also be mentioned, such as the Body_part affected by the condition, or a Symptom of the Ailment. In Frame Semantics terms, the lexical unit (LU) *arthritis.n* evokes the `Health_conditions` frame.

Moreover, the background scenes evoked by LUs are connected to each other via a series of typed relations, forming a network of frames, or a FrameNet. For the `Health_conditions` frame, such relations model: (a) via Inheritance, that this frame is a more specific type of the `Gradable_attributes` and of the `State` frames; (b) also via Inheritance, that an `Epidemic` is a specific type of `Health_condition`; (c) via Using, that `Symptoms` and `Body_parts` may be required for understanding the conditions; (d) that the `Health_conditions` frame is referenced by frames modeling the notions of `Recovery` and `Cure`; and (e) that this frame is a part of `Healthcare_scenario`, among other relations.

In FrameNet methodology (RUPPENHOFER et al., 2016), frames are proposed based on a combination of domain knowledge and corpus analysis providing evidence of the valence affordances of LUs. For the project reported in this paper, a domain-specific corpus was used. We describe the corpus next.

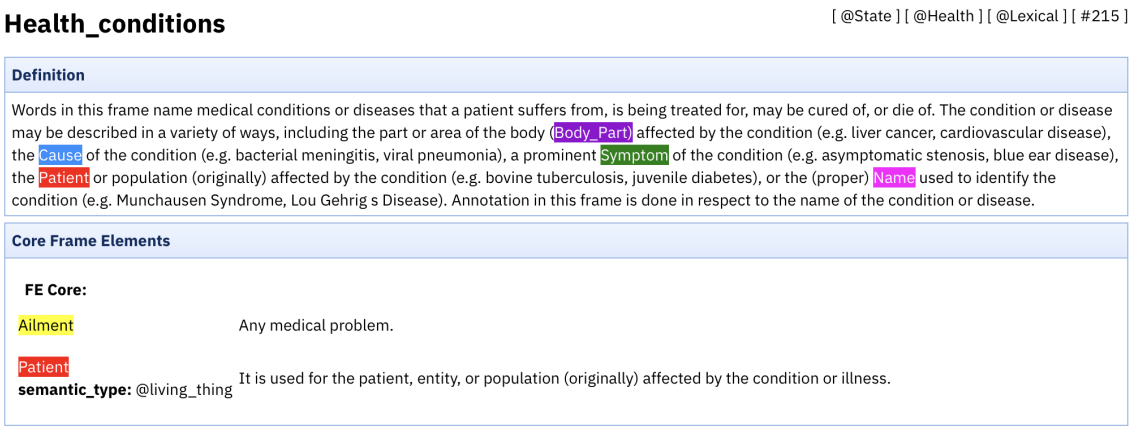

Figure 1. The `Health_conditions` frame

3. Materials

The methodology adopted for modeling the frames relies on the availability of a corpus of texts related to the Healthcare domain. The corpus used in this study was built using records extracted from the hospital admission information system—SIH—from the city of São Caetano do Sul in the state of São Paulo.For this study, the only piece of information that was used to build the corpus was the field in which health workers write down the patients' main complaints. These complaints are, in most cases, a single sentence stating the patient's symptoms and existing medical conditions, as well as for how long they have been occurring. Although considerably less common, some records contain other types of information, such as previous medical interventions, which in turn, can contain personally identifiable information (PII).

Taking into consideration the patients' rights to the privacy of their own data, especially when some could be victims of GBV, the most important steps in building the corpus were (a) extracting only the open text fields from the database, so that the lexicographers could not access the other data in the system, and (b) removing any PII from the texts. For this last process, any proper noun, dates and number were considered potential PIIs and excluded from the final corpus. All of the PII was replaced by tags in the text indicating that in the original text there was a name, date or number in that position. As a safeguarding measure, before any work was done with the corpus, one of the authors of this work[6] manually checked that there were no remaining PII in the final corpus.

The final corpus contains 32,980 sentences and a total of 225,416 tokens, with an average of ~6.8 tokens per sentence.

4. Modeling the Healthcare Domain

The process of computational modeling for a specific domain begins with a general study of that domain, involving the identification of potential participants, events, and other essential elements. Costa (2020) presents a nine-step methodology for modeling a specific domain in FrameNet Brasil, as shown in Figure 2. This methodology was effectively used to model the Healthcare domain, apart from minor adjustments.

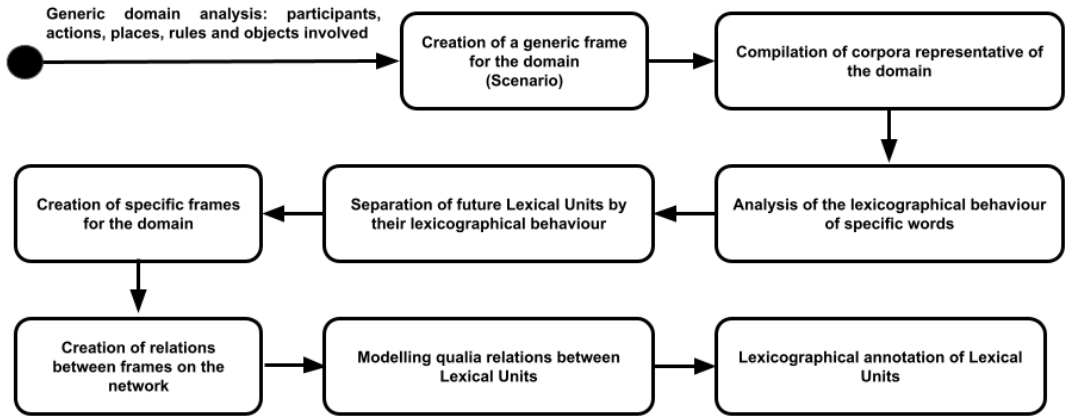

Figure 2. Steps in the process of computational modeling of a specific domain

The existing FrameNet database already included a generic frame called `Medical_interaction_scenario` along with 11 other frames intended for a potential Medical domain project. However, in order to capture a larger scope and cover the Healthcare domain, modifications to the FrameNet Brasil database were necessary. These adjustments were made in order to expand the frames beyond medical interactions and scenarios to include a more diverse range of scenes and participants. As a result, the generic frame was altered to `Healthcare_scenario`, and similar changes were made to the existing frames to align them with the Healthcare domain.

As predicted at the beginning of this work, these frames were not enough to cover the entire domain.Therefore, an analysis was made, grouping semantically related

[6] Said author only had access to the corpus prior to the anonymization process in an encrypted computer environment with appropriate credentials, complying to the methodology registered and approved for the project by the National Research Ethics Committee (process number 64733922.3.0000.5083).

words from the corpus described in section 3, which identified the need to create new frames to meet the demands of the domain. In the following sections, we will discuss the details of these processes, as well as the other steps involved in modeling a domain.

4.1. Clusterization of Lexical Units

The first step of the methodology consists of finding a set of single and multi-word expressions as candidate LUs. These candidate terms are not proposed arbitrarily. Instead, they are extracted from the corpus described in Section 3. Using the 'Keywords & Terms' feature of the Sketch Engine software [KILGARRIFF et al, 2014a], the top scored 1000 single-word and the top 1000 multi-word expressions were selected. Scoring in the Keywords & Tems tool is based on the ratio between the expression frequency on the focus and on a reference corpus. The Portuguese Web 2020 Corpus was used as the reference corpus in this comparison.

The following phase was dedicated to the arrangement of the candidate LUs within an initial set of existing frames, which varied from frames that were not directly related to the Healthcare domain—e.g. the `Kinship` frame**,** which contains lexical units like *pai.n 'dad', mãe.n 'mom'*, *irmão.n 'brother'*—and the existing frames associated to the Healthcare domain as the `Health_conditions` frame, shown in Figure 1.

From this point onwards, the remaining expressions were grouped and sorted according to their semantic similarity. As a result, domain-specific frames—such as the `Symptoms` frame, containing LUs like *náusea.n 'nausea'*, *tosse.n 'cough'*, and *choro excessivo.n 'excessive crying'*—were also created. The last example shows an interesting occurrence and showcases how this sorting process may vary due to the domain where the LU is inserted. Normally, LUs such as *choro.n 'crying'* and *excessivo.a 'excessive'* would both be considered and added into their respective frames individually: the first to the `Communication_Noise` and the second to the `Degree` frame, since the adjective describes the crying as excessive. However, within a specific domain, it is crucial to consider particularities such as considering the intensity marker as a part, and not as an accessory, of the LU which, thus, is added to the `Symptoms` frame.

As another example to illustrate the grouping and sorting process, take the `Health_intervention` frame, which consists of LUs that semantically represent intentional procedures performed to treat patients. Even if they differ a lot in magnitude—such as *curativo.n 'bandage'*, *amputação.n 'amputation'*, *exame.n 'exam'*, *cirurgia.n 'surgery'*—, they are still within the same semantical boundary. As the number of candidate terms grew larger, so did the number of prototypical frames containing these terms, all sorted based upon their semantic function, which were then to be elaborated into frames *per se*.

4.2. Frame Creation Process

The frame creation process can follow the bottom-up, the top-down perspective, or a combination of both. The bottom-up approach consists in studying a corpus to create a frame, i.e., it takes into account linguistic evidence to structure the frame, analyzing patterns and relations. On the other hand, the top-down approach starts from the researcher's intuition, creating the frame structure as a starting point, almost a reverse process if compared to the first method. To model a specific domain both approaches

are combined [TORRENT et al., 2014]: we start with the bottom-up approach and check the intermediate analyses against the systematized knowledge of the domain.

In the bottom-up approach, the process begins with the previously mentioned grouping of lemmas derived from the selected corpus. Once this step is completed, the semantic and syntactic valence of the grouped lemmas is examined within the sentences in which they appear.[7] This allows for the extraction of patterns, which helps to define and structure the frame. In parallel, the frame elements and their coreness are determined. Figure 3 shows the relevant information of a frame and demonstrates the structure of the created frame `Symptoms`.

Figure 3. The `Symptoms` frame

Up to this point, 17 frames have been created to integrate the network that composes the Healthcare domain. Taking into account the existing frames that have been adjusted, the domain now consists of 29 frames with 1,787 lexical units associated

[7] Access to the sentences in the corpora is limited to a selected number of researchers in the team, all of which are bound by non-disclosure policies established in the ethics protocol for the project.

with it. As already stated, the domain is a network—not a list—which means that the frames are interconnected at some level. This is established by means of frame relations.

4.3. Frame Relations

The existing frame-level relations in FrameNet Brasil can be of three types: between Frames (F-F), between Frame Elements (FE-FE) and between Frames and Frame Elements(F-FE) [TORRENT et al., 2022]. These relations not only allow for the organization of frames within a framework, but also improve the understanding of frames by providing additional semantic information.

The types of F-F relations are represented by arrows of different colors and were defined by Ruppenhofer et al. (2016) as: Inheritance (red), Subframe (blue), Precedes (black), Using (green), Causative_of (yellow), Incoative_of (brown), Perspective_of (pink) and See_also(purple). The current state of the network is shown in Figure 4.

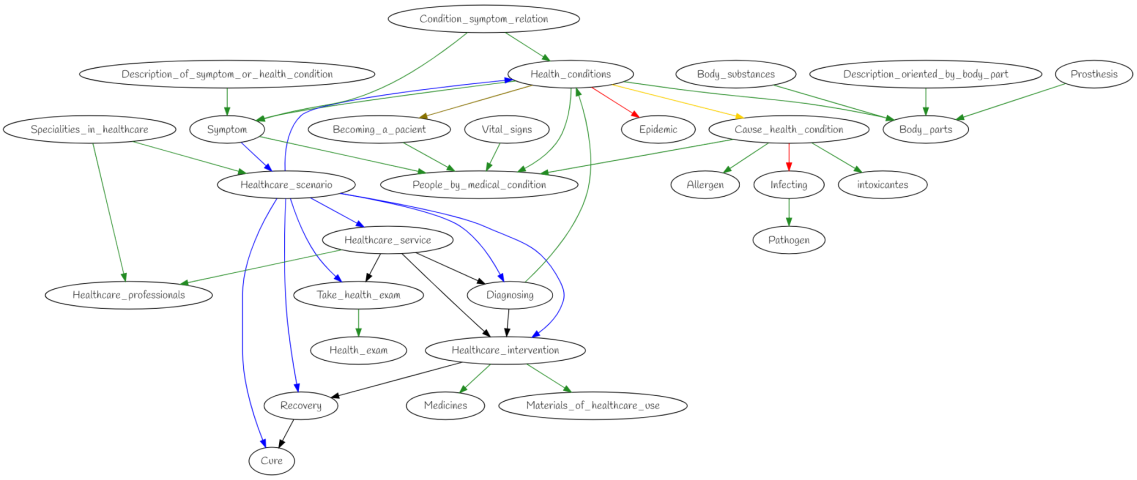

Figure 4. F-F Relations in the Healthcare Domain

Regarding FE-FE relations, it is possible to indicate similarities and correspondences between FEs from different frames. Within the Healthcare domain, for example, the Healthcare_Professional FE of the `Healthcare_intervention` frame is directly connected to the Professional FE of the `Health_care_service` frame. This connection indicates that the professionals engaged in the mentioned interventions are those related to healthcare. Furthermore, the F-FE relations, in a similar manner to the previous one, indicate a FE of a given frame that makes reference to another frame. This is the case, for instance, of the frame `Take_a_health_exam` which has the Health_exam FE that is directly related to the `Health_exam` frame. Such a relation points out that the exams covered by this FE are the LUs that fall under the `Health_exam` frame.

4.4. Qualia Relations

A different type of relation found in the FrameNet Brasil database structure are qualia relations based on the qualia structure proposed by Pustejovsky (1995). Unlike the relations discussed so far, qualia relations express associations between LUs. Qualia relations can be subdivided into four different types: agentive, constitutive, formal and

telic. This small set is, as expected, very general. To further specify these associations, FrameNet Brasil uses frames to mediate the relations between two LUs.

These types of connection are essential to capture aspects of meaning that cannot be adequately represented in the original FrameNet model. For instance, the LU *alergia.n 'allergy* in the `Health_conditions` frame, can be associated to *amendoim.n 'peanut'* in the `Food_and_beverages` frame, as peanuts are known to cause allergic reactions in some individuals. This is an instance of the agentive quale that in FrameNet Brasil is also mediated by the `Cause_health_condition` frame, with *amendoim.n ´peanut´* filling the Cause_of_the_health_condition FE slot and *alergia.n ´allergy´* filling the Health_condition FE. This mediated association models the intuitive notion that when *alergia.n ´allergy´* evokes `Health_conditions`, frames such as `Food_and_beverages` may be relevant in the context.

4.5. Annotation

The *corpus* annotation represents the last step of the frame modeling process, and validates the model. For this project, following the methodology adopted by FrameNet Brasil, the full-text perspective was chosen for better results. In this approach, the researcher uses the entire *corpus* as a source, rather than predetermined words in selected sentences. Thus, all the meaningful LUs are annotated in relation to other constituents in the sentence. It allows for a much wider semantic and syntactic analysis upon the textual genre being annotated which aligns with the goals of the project.

Moreover, the annotated corpus, containing sentences in the Healthcare domain, the frames evoked in these sentences and the distribution of the relevant FEs, is an important resource to train semantically enriched machine learning models.

5. Conclusions

This paper has presented an overview of the effort to model the Healthcare domain according to FrameNet Brasil's methodology. It has shown a bottom-up methodology for semantic modeling, based on a corpus of the healthcare domain. The resulting network of frames, consisting of 29 frames, 17 of which are newly-created, already covers 1,787 LUs. This model of the domain was created as part of a larger project with the long-term goal of identifying traces and tendencies of GBV in certain territories in Brazil, based on public health system records. To achieve this goal, we plan on using, among other techniques, machine learning models. For future work, we propose the use of annotated data, connecting the texts written by health professionals to the frame structure created in this work, to enhance the quality of those models. A semantically enriched machine learning model is more likely to be informative (e.g. by also relating output to frames), but also has the potential of having better performance because it does not solely rely on raw texts. In this context, a better model is also essential to identify patterns that could be related to GBV, which could be in turn used by policy makers to take more informed actions.

Acknowledgements

Research presented in this paper was funded by the Data to Safeguard Human Rights Accelerator Program of the Patrick McGovern Junior Foundation.

References

Costa, Alexandre Diniz. (2020) “A tradução por máquina enriquecida semanticamente com frames e papéis qualia.” (Ph.D. thesis in Linguistics. Universidade Federal de Juiz de Fora, Juiz de Fora.)

Fillmore, Charles J.(1982) The case for case reopened. In: Grammatical relations. Brill, 1977. p. 59-81.

Fillmore, C. J. (1982). Frame semantics. In: Linguistic Society of Korea (ed.), “Linguistics in The Morning Calm”. Seoul: Hanshin, p.111-138.

Garbin, Cléa Adas Saliba et al. (2015) “Desafios do profissional de saúde na notificação da violência: obrigatoriedade, efetivação e encaminhamento.” In: Ciência & Saúde Coletiva, v. 20, p. 1879-1890.

Kilgarriff, Adam et al. (2014) “The Sketch Engine: ten years on.” In: Lexicography, v. 1, n. 1, p. 7-36.

Kilgarriff, Adam et al. (2014) “PtTenTen: A corpus for Portuguese lexicography.” In: Working with Portuguese Corpora, p. 111-30.

Kind, Luciana et al. (2013) “Subnotificação e (in) visibilidade da violência contra mulheres na atenção primária à saúde.” In: Cadernos de Saúde Pública, v. 29, p. 1805-1815.

Pustejovsky, James.(1998) The generative lexicon. MIT press.

Ruppenhofer, Josef et al. (2016) Framenet II: Extended Theory And Practice. https://Framenet2.Icsi.Berkeley.Edu/Docs/R1.7/Book.Pdf).

Torrent, Tiago Timponi et al. (2014) “Multilingual lexicographic annotation for domain-specific electronic dictionaries: The Copa 2014 FrameNet Brasil project.” In: Constructions and Frames, v. 6, n. 1, p. 73-91.

Torrent, Tiago Timponi et al. (2022) “Representing context in framenet: A multidimensional, multimodal approach.” In: Frontiers in Psychology, v. 13. doi: 10.3389/fpsyg.2022.838441