LOD-connected offensive language ontology and tagset enrichment

Barbara Lewandowska-Tomaszczyk[1], Slavko Žitnik[2], Anna Bączkowska[3], Chaya Liebeskind[4], Jelena Mitrović[5,6] and Giedre Valunaite Oleskeviciene[7]

[1]*Department of Language and Communication, State University of Applied Sciences in Konin, Poland*
[2]*University of Ljubljana, Faculty for computer and information science, Večna pot 113, SI-1000 Ljubljana, Slovenia*
[3]*Department of Glottodidactics and Natural Language Processing, University of Gdansk, Poland*
[4]*Department of Computer Science, Jerusalem College of Technology, Jerusalem, Israel*
[5]*University of Passau, Germany*
[6]*The Institute for Artificial Intelligence Research and Development of Serbia*
[7]*Institute of Humanities, Mykolas Romeris University, Ateities 20, LT-08303, Vilnius, Lietuva*

Abstract

The main focus of the paper is the definitional revision and enrichment of offensive language typology, making reference to publicly available offensive language datasets and testing them on available pre-trained lexical embedding systems. We review over 60 available corpora and compare tagging schemas applied there while making an attempt to explain semantic differences between particular concepts of the category OFFENSIVE in English. A finite set of classes that cover aspects of offensive language representation along with linguistically sound explanations is presented, based on the categories originally proposed by Zampieri et al. [1, 2] in terms of offensive language categorization schemata and tested by means of Sketch Engine tools on a large web-based corpus. The schemata are juxtaposed and discussed with reference to non-contextual word embeddings FastText, Word2Vec, and Glove. The methodology for mapping from existing corpora to a unified ontology as presented in this paper is provided. The proposed schema will enable further comparable research and effective use of corpora of languages other than English. It will also be applied in building an enriched tagset to be trained and used on new data, with the application of recently developed LLOD techniques [3].

Keywords

offensive language, automatic offensive language detection, tagset enrichment, ontologies, ontological modelling, Sketch Engine, tagsets, word embeddings

1. Introduction

The paper addresses one of the key challenges for automatic offensive language detection, which concerns its coverage and ontological categorization. Technological advancement and

SALLD-1: Workshop on Sentiment Analysis & Linguistic Linked Data, September 1, 2021, Zaragoza, Spain
barbara.lewandowska.tomaszczyk@gmail.com (B. Lewandowska-Tomaszczyk); slavko.zitnik@fri.uni-lj.si (S. Žitnik); anna.k.baczkowska@gmail.com (A. Bączkowska); liebchaya@gmail.com (C. Liebeskind); jelena.mitrovic@uni-passau.de (J. Mitrović); gvalunaite@mruni.eu (G. V. Oleskeviciene)
0000-0002-6836-3321 (B. Lewandowska-Tomaszczyk); 0000-0003-3452-1106 (S. Žitnik); 0000-0002-0147-2718 (A. Bączkowska); 0000-0003-0476-3796 (C. Liebeskind); 0000-0003-3220-8749 (J. Mitrović); 0000-0001-5688-2469 (G. V. Oleskeviciene)

increasing online communication has brought in the necessity to automatically detect offensive or abusive language to protect the users and to process the extensive communication data automatically.

Based on available datasets and tagset resources, the present paper aims at two basic objectives. Firstly, we identify and critically review existing offensive language tagsets and, secondly, we create an ontology basis that is proposed as a schema for offensive language identification. We also include additional metadata that enable encoding of an arbitrary dataset into the ontology. Furthermore, we analyse offensive language cases from the datasets and propose an enrichment of the existing ontology, based on a semantic and contextual analysis of particular linguistic categories in the publicly available social media textual resources.

2. Research Methodology

Over 60 available corpus data sets, such as Qian et al. [4], Gomez et al. [5], Wulczyn et al. [6] and others, are scrutinized and the relevant tagging schemas originally applied compared, while making an attempt to explain semantic differences between particular concepts of the category OFFENSIVE in English. We adopt a finite set of classes that cover aspects of offensive language representation, based on the categories originally proposed by Zampieri et al. [1, 2]. Particular offensive words were tested by means of Sketch Engine (SE) and Thesaurus tools on a large web-based corpus (19 billion items), available in SE. The schemata are juxtaposed and discussed with reference to non-contextual word embeddings FastText, Word2Vec, and Glove, originally on a smaller number of offensive terms, in the second phase with four additional offence-related items added, and eventually in the final phase – with an additional inclusion of the item *taboo*. The discussion in the present paper makes reference to the final phase of our study. The study discusses the methodology for mapping from existing corpora to a unified ontology as presented in this paper.

The proposed schema enables further comparative research and an effective use of corpora of languages other than English. It will also be applied in building an enriched tagset to be trained and used on new data, with the application of recently developed LLOD techniques [3].

3. Critical review of available offensive language terminology and categorization

Terminological problem The main problems for automatic identification of offensive language (as well as other related phenomena e.g. hate-speech, abusive language) has been a lack of consensus concerning the definitions of such phenomena. Likewise, the tagsets used for annotating such texts vary. Our research is aimed at overcoming this problem for the offensive language annotation and detection tasks.

Survey of available categories and terms Authors researching language detection for the user protection adopt various terms in their ontological and tagset systems, which are not clearly defined, identifiable or retrievable. Poletto et al. [7] offer a classification in which abusive language is used as an umbrella term for other, similar phenomena – hate speech, offensive

language, etc. Likewise, Founta et al. [8] define abusive language as language used to refer to hurtful language, including hate speech, derogatory language and also profanity. Offensive language, on the other hand, has been defined by Razavi et al. [9] as profanity, strongly impolite, rude or vulgar language expressed with fighting or hurtful words in order to insult a targeted individual or group.

Based on the Sketch Engine data and collocate frequencies, we propose the concept of offensive language as a superordinate category in our system to cover instances of language which upsets or embarrasses people because of its insulting character. In legal contexts [10], offensive language is defined as the term indicating hurtful or derogatory comments by one person or a group to another person or a group. In terms of socio-cultural standards, offensive language identifies a number of hierarchically arranged subcategories such as taunts, directed to ridicule the addressee, references to handicaps, squalid language, which includes allusions to sexual fetishes, slurs which are directed to attack certain culture or ethnicity, homophobia, racism, extremism, crude language which refers either to sexual matters or excrements, disguise which may carry ambiguous meaning, or else direct insults, which contain so-called four-letter words, or provocative language which may cause anger as well the use of taboo-words. Jay and Janschewitz [11] distinguish three main divisions of offensive language including vulgar, referring to rude sexual comments or references, pornographic which comprises obscene comments, and hateful, which denotes offensive remarks related to race, religion, country, etc., also personal attacks and demeaning comments.

We propose to consider abusive language to be viewed as a constituent of the superordinate category of offensive language, characterized in legal terms as harsh, violent, profane, or derogatory language which is directed to violate the dignity of an individual, including profanity and slurs of racial, ethnic, or sexist manner. According to Nobata et al. [12] abusive language includes dominance, derailing, harassment and threat; it may discredit race, religion or gender and may also include stereotypes and spam. Other offensive and abusive language categorization systems identify such subclasses as benevolent, derailing, discredit, dominance, harassment, hate, hostile, insult, obscene, profane, spam, and stereotype [13]. Still other proposals include various categories a number of which require less vague definitions, absent in the relevant publications as e.g., the notion of toxic language and its levels [14], left undefined and attributed to the superordinate category of the same form, which makes them circular and indistinguishable from each other.

The category of hate speech is a similarly fuzzy concept in the previous studies, causing familiar problems in the inter-annotator agreement [15]. We constrain it and define it as a group-targeted or an individual-targeted offence, based on one or more of the identifiable negative stereotypes referring to ethnicity, gender, religion, or ridiculed properties attributed to this group. Harassment overlaps as part of a larger cyberbullying category. Cyberbullying refers to the online intimidating and threatening content and embraces not only harassment, possible hate speech and other offensive modes but also to massive behaviour of individual or group attackers targeted towards discrimination or exclusion of particular individuals from their groups by various kinds of offence, defaming, deceit, etc., both online as well as by means of the devirtualization of the offence from online to offline real world spaces [16].

4. Available automatic offensive language identification and detection tools

The most prominent efforts in detecting offensive language have been the HatEval Task 5 of Semeval-2019: Multilingual detection of hate speech against immigrants and women in Twitter [17] and the OffensEval Task 6 of SemEval-2019: Identifying and Categorizing Offensive Language in Social Media [2], which has seen its second edition in 2020 as Task 12 of SemEval-2020: Multilingual Offensive Language Identification in Social Media [18].

The development of systems for automatic identification of offensive language and similar phenomena has followed a well-known trend in NLP: feature-based linear classifiers [19], neural network architectures applied to corpora in different languages [20, 21, 22, 23, 24], and, fine-tuned pre-trained language models, such as BERT and RoBERTa [25, 26]. Results vary both across datasets and architectures, with linear classifiers qualifying as very competitive, if not better, when compared to neural networks. On the other hand, systems based on pre-trained language models have proved to have the best performance in this area, reaching new state-of-the-art results. For example, HateBERT [27], a pre-trained BERT model for investigating hate speech, offensive and abusive language in social media is currently outperforming the other models [15].

Apart from Machine Learning systems, lexicon-based approaches to tackling hate speech and related phenomena have been facilitated with resources such as HurtLex [28]. HurtLex is a lexicon of offensive, aggressive, and hateful words in over 50 languages. The words are divided into 17 categories, plus a macro-category indicating whether there is a stereotype involved. Special attention has also been given to multiword expressions used in the phenomena we are discussing here, with interesting results of a combined lexical and Machine Learning approach presented in [29].

Still, very few attempts at using LOD for modelling and detection of offensive language were made, with limited success. One such resource is the Colloquial WordNet [30], an extension of the Princeton WordNet [31] that focuses on the use of neologisms and vulgar terminology. Some authors have approached the task of tackling ambiguity in hate speech detection using a combined ontology, sentiment analysis and fuzzy logic-based approach [32]. Battistelli et al. [33] built a hate speech ontology in French (having four main concepts - Action (Action), Target (Cible), Context (Contexte) and Orientation (Orientation)), but no such ontology has been built for English.

5. Extended offensive language tagset

The extended ontological tagset presented in the paper identifies two basic levels, Level I covering sub-levels which refer to lexicon, POS, and syntax, and Level II – for further multimodal uses, which additionally includes subcategories connected to visual elements (in social media datasets) and, considered for further extension, prosodic elements of speech parameters. In the present paper we elaborate on the basic language level of the extended Offensive Language ontological schema presented below.

5.1. Ontology

The category system of offensive language we propose was inspired by the three-level hierarchy of offensive language put forward by Zampieri et al. [1, 2]. Contrary to Zampieri et al. [1, 2], however, in our research, as mentioned above, offensive language is further refined and is divided into two basic levels of analysis (Level I and II) and four sublevels (A, B, C, D) within Level I. Firstly, at Level I, we distinguish offensive from non-offensive language (Level A: offensive vs. non-offensive). The non-offensive cases are beyond the scope of our research. Secondly (Level B: targeted vs. non-targeted), the question of Target is taken up - if there is no identifiable addressee of offence the language is considered an example of self-expression, having, e.g., an exclamatory function (e.g. swear words used to express anger, frustration, pain, etc.). Andersson and Trudgill (2007) dub such instances "abusive swearing". They are often used as stand-alone discourse segments, not integrated with a sentence, typically known as inserts [34, 35]. Such cases of non-targeted offensive language are not included in the present study. Similarly, excluded from this category are contexts in which bad language (in the form of swearing) is targeted at some addressee but for purposes other than expressing offence, e.g., in the contexts of "humorous swearing" [36]. Offensive words may also express solidarity and bonding with the interlocutor [37, 38], alignment with the community of practice [39, 38, 40], friendly affection [41, p. 148], or can be considered an in-group identity strategy in a community [42, 43] [44, p. 297]. Targeted offensive terms are further divided into implicit or explicit cases (Level C: implicit vs. explicit language). While implicitness may be encoded by, for example, irony, in which offence is not straightforward, explicitness entails more direct forms of verbal attack. Along with covert forms of implicit offence, the design of participation framework is considered (e.g., a CMC user criticizes person A by addressing the comment to person B, yet with person A being a part of the discussion). Such cases too are instances of implicit offensive language, parallel to off-line communication contexts in which interactant A is a hearer [45], i.e. the participant who is sanctioned (ratified) and who shares the communicative context. There are numerous mixed contexts for further consideration, e.g., intentional versus surreptitious listeners (unknown and unratified). Such subtle variations of participant status are beyond the scope of our current research. Explicit forms of offensive language involve either individual recipients or groups of individuals. The latter, which relates to cases of hate speech, considered here a recursive behaviour (also linguistic), relying on stereotypes. Individual and group addressees may refer to recipients who are either participants of verbal exchange (by default they are ratified participants), and they are dubbed internal addressees, or to those who are absent, and then they have the status of external addressees. Such types as "You are like all those gays" are also members of the hate speech category, in which an offensive message is communicated to an individual by making reference to a community the individual is a member of. Intentionality is another aspect included, in which an intentional offence is or is not recognized as offensive by the addressee. Intentionality is a complex parameter and has not received attention in NLP-based research on the automatic detection of offensive language. It is beyond the aim of our present research, yet it is an important issue for further considerations of contextual underpinnings of such actions. A separate category of offensive language concerns metaphorical language, both conventional and creative metaphors. While conventional metaphors are identified as fixed (phraseological or sentential) expressions and

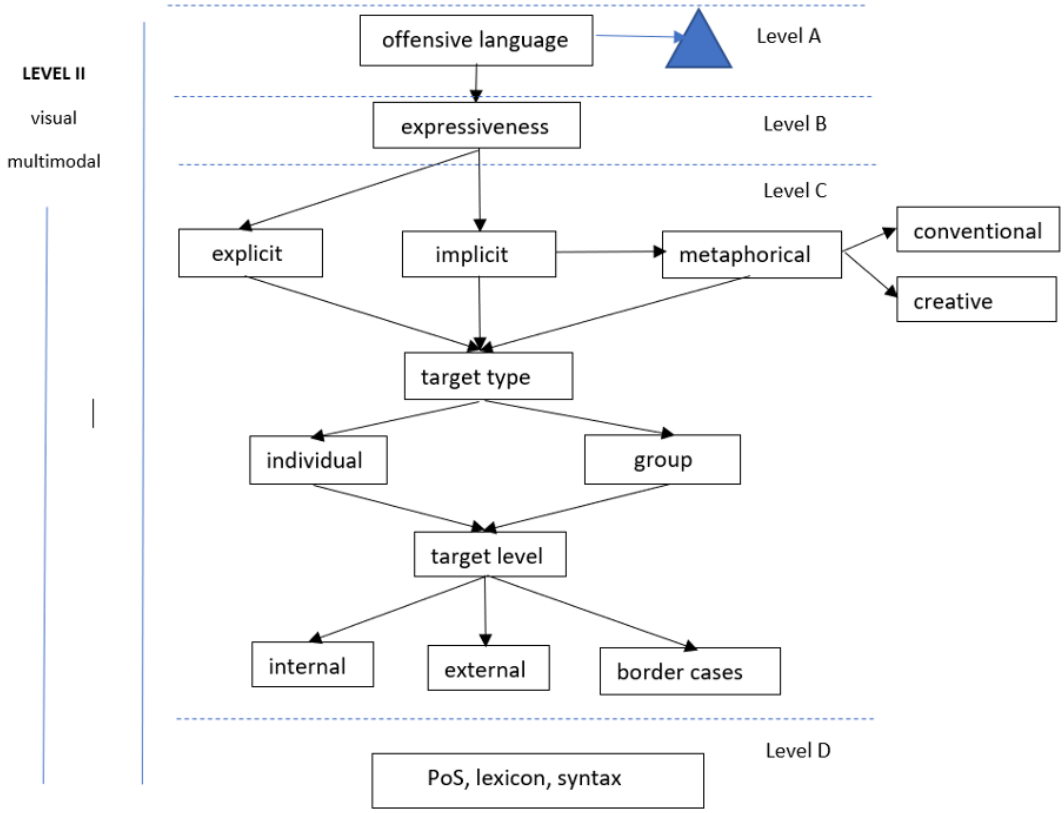

Figure 1: Ontology of offensive language (OL) and methodology of OL detection.

are easy to detect, creative metaphors are notoriously unaccounted for in NLP models. Finally, a fine-grained sublevel (D) requires an analysis of morphosyntactic features, i.e. aspects at the word (parts of speech tagging, lexical analysis) and sentence level. Importantly, sublevel C and sublevel D comprise not only verbal offence (Level I) but also visual and multimodal forms that combine gestures, proxemics, kinesics, gaze, as well as paralinguistic/prosodic features, etc. (Level II) [46, 47, 48]. The extended offensive language ontological system elaborated on in the present research is shown in Figure 1.

To reassume, the concept of offensive language is used here as a generic term, superordinate to abuse, hate speech, insults, taboos, vulgarity, slurs, profane, toxic language, flaming, harassment, etc. The terms are mapped onto the classification described above (Fig. 1.), and thus, for example, abusive language may be targeted at an individual or a group of addressees, it may be expressed implicitly or explicitly as well as literally or metaphorically. The same procedure applies to other types of offensive terms. As the ordinary use of the linguistic forms offensive, abusive, insulting, and hateful as well as taboos, slurs, and harassment are notoriously under defined, we used the corpus management platform the Sketch Engine, which allows both qualitative and quantitative analysis. The enTenTen15 corpus [49] was employed as it is large in size

(over 13 billion segments) and contains online texts. The enTenTen15 consists of web content (crawled by Spiderlink), cleaned and processed heuristically (by useText), and deduplicated (by onion). Three tools have been used to identify the core features of the four concepts under investigation: Thesaurus (Th), Word Sketch (WS) and Word Sketch Difference (WSD). Based on numerical data retrieved by Th and WS, some preliminary core features have been identified, as shown below. Thesaurus classified synonym forms, arranged according to the frequency of occurrence in relation to the query concept, and a similarity score. WS, on the other hand, surveys collocations and grammatical constructions (using logDice). The sample words cited below are based on the similarity score:

Offensive. Th: violent, destructive, inappropriate, ridiculous, disturbing, oppressive, absurd, controversial, etc. WS: [Adj] patently, morally, mildly; [N] odour, remark, [V] find sth., consider, think.

Abusive. Th: violent, oppressive, discriminatory, hateful, cruel, racist, destructive, etc. WS: [Adj] verbally, emotionally, sexually, physically; [N] husband, relationship, behaviour; [V] not tolerate, report, consider, deem.

Insulting. Th: disrespectful, hurtful, derogatory, obscene, vulgar, disgusting, rude, etc. WS: [Adj] grossly, mildly, outrageously, vaguely; [N] Islam, remarks, Erdogan, nickname, insinuation; [V] say sth., find, consider, think.

Hateful. Th: vile, bigoted, hurtful, homophobic, racist, vulgar, anti-Semitic, etc. WS: [Adj] outright, overtly, blatantly, racially, terribly; [N] curse, rhetoric, slur, invective; [V] spewing, called, acted, grow.

The third tool, the WSD, was used to compare the differences between pairs of concepts, e.g. offensive vs. abusive, offensive vs. insulting, and vs. abusive. From this study it has transpired that offensive is the weakest term, abusive is the strongest, and insulting stands midway. The qualitative and quantitative analysis of offensive allowed us to identify the term as short, punctual, i.e. bounded actions (events), opposed to, for example, harassment, which profiles actions of prolonged behaviour, with no clear onset and ending (aspectual, unbounded concept of duration). Slurs imply attacks on individuals and/or groups and denote alleged untruthfulness, while hateful has clear sociological implications.

A dataset of Twitter posts will be extracted from the Brexit corpus (available on the Sketch Engine platform) in order to validate the proposed categorization scheme. The contexts will be retrieved by a number of query words that will illustrate the categories presented in our model.

Based on the above categorization, we encoded the schema into a generic ontology[1]. An example encoding of a dataset is presented in Fig. 3. In the Figure, we encode a sentence "Move away, you idiot!" whose source is Slavic news and is part of Nexus Ligarium toxic dataset. The example is categorized based on the proposed schema into Offensive type, Explicit expressiveness, Individual Stranger target type and Internal level context.

[1] https://ul-fri-zitnik.github.io/offensive-language-ontology/offensive-language-ontology.owl (Accessed: June 3, 2021)

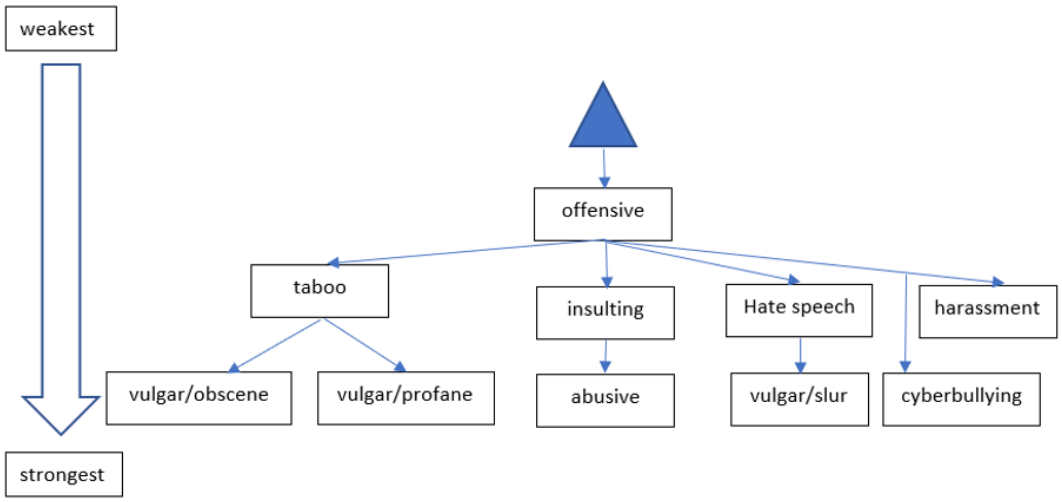

Figure 2: Typology of *offensive language*.

6. Validation of offensive language terms using pre-trained word embeddings

In addition to linguistically sound interpretation of offensive language terminology, we support the categorization using computational linguistics aspects. We use pre-trained word embeddings (i.e. Word2Vec, Glove and fastText) to draw conclusions regarding meaning and interpretation of the selected terms. We draw results from Word2Vec [50] (i.e. pre-trained Google News corpus having 3 million 300-dimensional English word vectors), Glove [51] (i.e. pre-trained Wikipedia and Gigaword corpus having 0.4 million 300-dimensional word vectors) and fastText [52] (i.e. pre-trained CommonCrawl and Wikipedia having 2 million 300-dimensional English word vectors) embeddings.

We selected 16 keywords in their lemma forms for the analysis - offensive, abusive, cyberbullying, vulgar, racist, homophobic, profane, slur, harassment, obscene, threat, discredit, hateful, insult, hostile and taboo. For each keyword 30 neighbouring words were retrieved in each word embedding and filtering was performed where neighbouring words were omitted that contain the keyword or its lemma, or else a stem as a substring. Multiple types of visualization techniques were performed, while empirically the t-SNE with the perplexity value of 15 produced the most conspicuous representations. It is important to note that in the case of t-SNE application (a) cluster sizes seem to play no particular role, (b) neither do inter-cluster distances, (c) random noise may present some non-random significance, and (d) to uncover an adequate topology, multiple empirical visualizations are needed [53].

Our pre-trained word embeddings were trained on three different datasets of a distinct vocabulary size, which shows an impact on the results of term clustering. Thus, we accept each of the pre-training methods to provide its own insights on the tested terms. First, we focus on the interesting findings that are assessed by all the pre-training methods. Then, terms

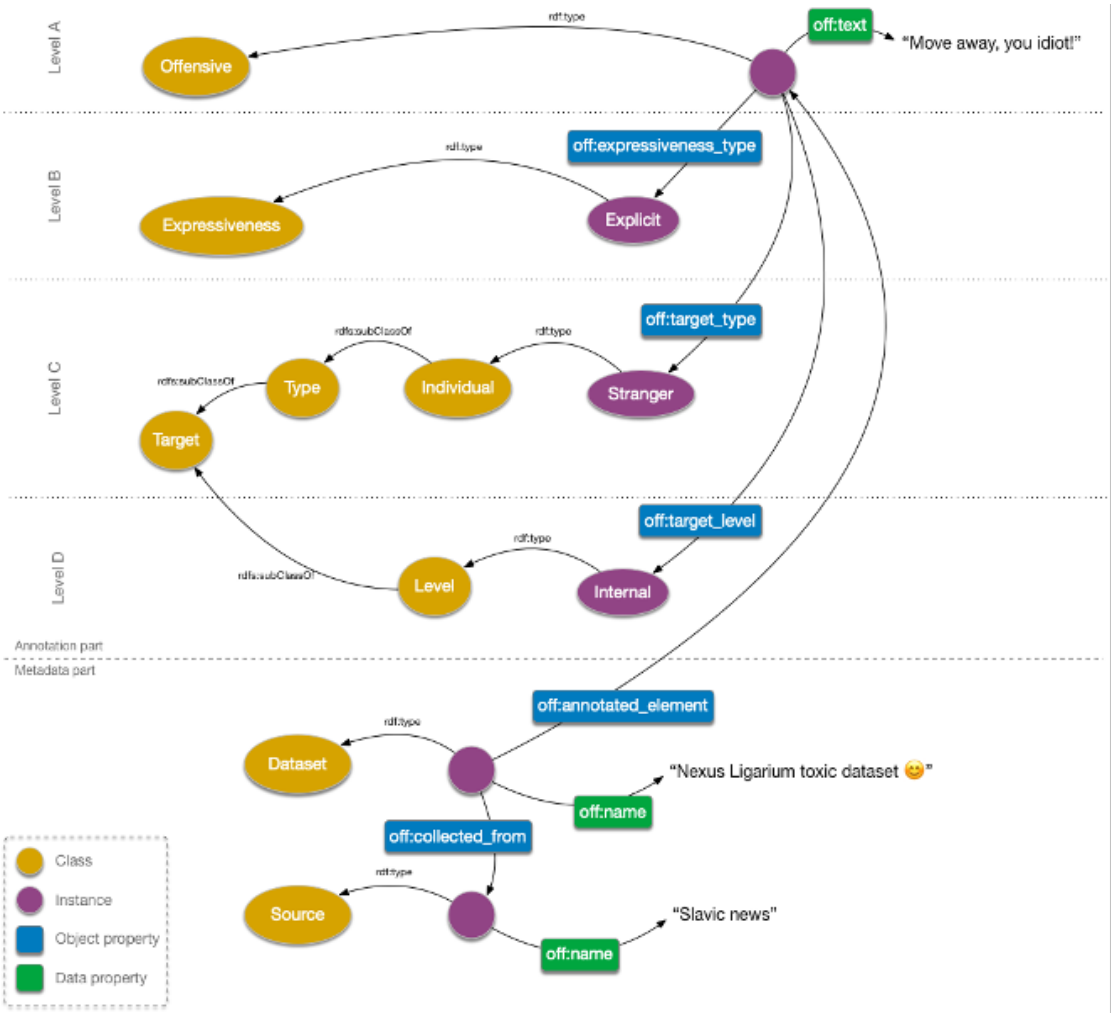

Figure 3: Proposed ontology schema along with an example, encoded using it.

presenting different behavior across the three methods are discussed.

7. Word embeddings versus linguistic analysis categories

FastText is believed to be more accurate than word2vec and Glove as it is based on character n-grams (subword embeddings) rather than words only; thus, the original corpus of words is enriched by the words stemming from subword analysis (along with word-level analysis). The FastText output shows the most discrete clusters of the three methods of word embeddings, which facilitate establishing word classification baselines. For this reason, in this analysis concept categorizations are based primarily on fastText.

Discrete items form a separate category and they comprise offensive, cyberbullying, harass-

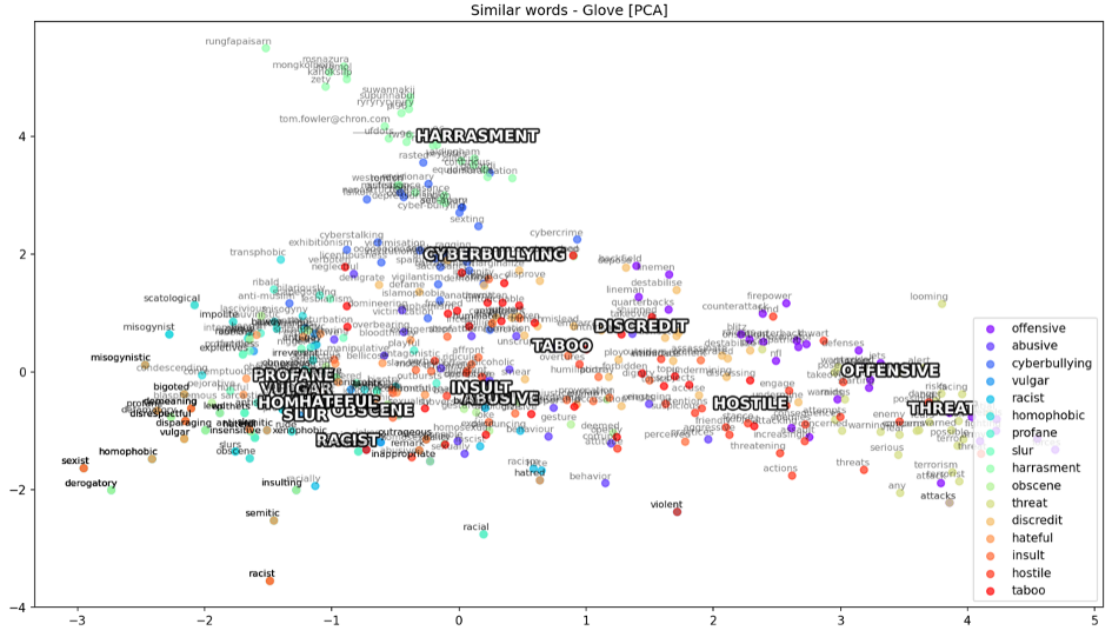

Figure 4: Glove word vectors representation using PCA.

ment, threat, and homophobic. Relatively well-delineated clusters are also formed by: slur, racist, discredit, hateful. Concepts with the least clear-cut borderlines and which spread widely include: profane/obscure/vulgar, hostile/abusive, and insult as well as taboo.

Concepts with show overlap with other elements are insult (intermingles with abusive and slur), abusive (intermingles with hostile, insult, hateful), profane (intermingles with obscene and vulgar).

7.1. Detailed analysis of word embeddings

In the second phase of the embeddings validation, the term taboo was added to the list to explore its influence on the item configuration in the embeddings using the same procedure with the application of three datasets and three different tools.

According to Maaten and Hinton [53] topology identification may require multiple empirical visualizations, so in the next step an overview of a number of representations allowed a clearer hierarchical interpretation applied for methodological consideration.

Word2Vec diagram demonstrates a certain degree of diffusion in the representation of the results; however, there are certain areas revealing the close-knitted relations among the analysed concepts. We can observe a twofold manifestation of the analysed concepts of some forming more discrete clusters and others falling into an overlapping mode.

In the Glove tool, a closely related cluster or cloud of concepts of profane, vulgar, hateful, slur, obscene, homophobic and racist demonstrates a high degree of overlap. The remaining terms show distancing between the represented notions, nevertheless the previously mentioned

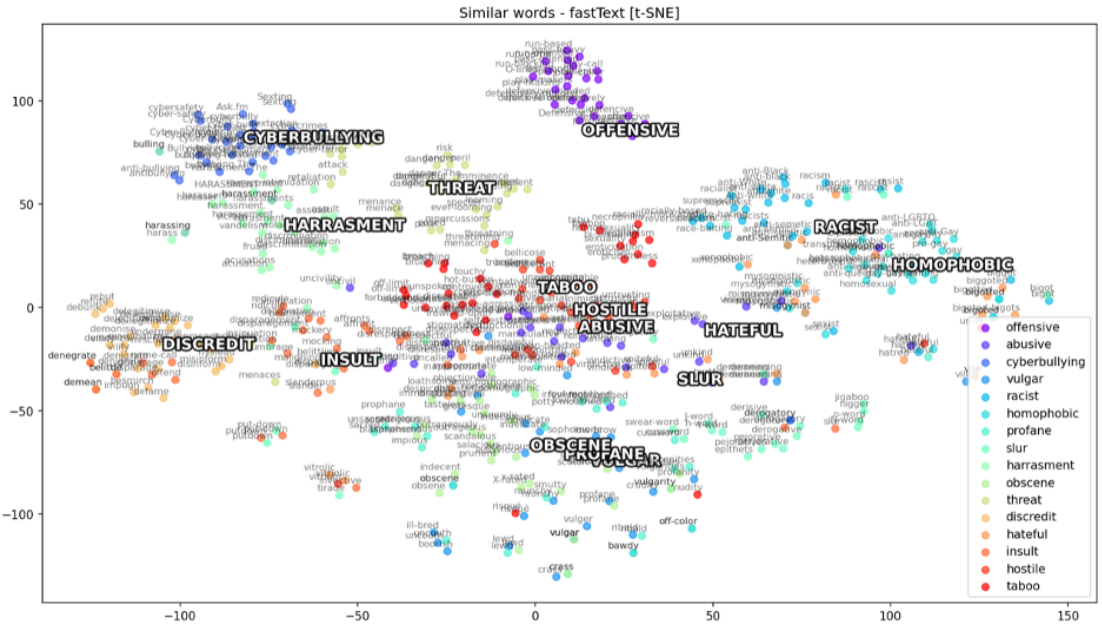

Figure 5: FastText word vectors representation using t-SNE.

cloud of concepts in its own right relates to insult and abusive further expanding towards taboo, cyber-bullying, discredit, hostile and offensive.

The representation of FastText also demonstrates certain overlapping of taboo, hostile, abusive, slur, hateful, obscene, profane, vulgar, and insult. This overlap could be also supported by the SE word sketch representations, which reveal that collocations with taboo really embrace some collocates of a vulgar and obscene character. It also allows observing certain relatedness between more discretely prominent clusters of insult and discredit as well as hateful related to racist and homophobic. It could be identified that taboo is loosely related to threat and harassment. The notion offensive in the representation forms a discrete group that is of a more general, possibly superordinate, character and forms a cloud above the overlapping notions. This allows considering the term offensive as embracing the other terms in the diagram.

8. Relevance to LLOD

The ultimate aim and backbone of our methodology are its complementarity to the use of offensive language ontologies and tagset systems, as well as their integration with the public LLOD resources. The results of our study will be further extended to languages other than English such as German from the Germanic group, Slovenian, Serbian, Croatian, and Polish from the Slavic group, Lithuanian from the Baltic one, as well as Hebrew as a Semitic language. Some of the most recent contributions towards LLOD development is a new methodology for building data value chains, applicable to a wide range of sectors and applications and based

around language resources and language technologies, presented in the paper by Declerck et al. [54] with reference to results of the European H2020 project "Pret-a-LLOD" ('Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors'). The project, proposing an advanced LLOD methodology, uses the Linguistic Linked Open Data cloud and the OntoLex-Lemon representation model for lexical data. The LLOD cloud is used to ensure the application to multiple languages and data modalities (corpora, thesauri, dictionaries), while the OntoLex-Lemon incorporates tools to use richer linguistic grounding for particular lexical concepts and their morphological and syntactic contexts in the conceptual-semantic ontology schemas. To ensure smooth applicability to languages other than English the approach proposed in Gracia et al. [3] - the Ontolex-lemon model and the Apertium initiative using cross-lingual transfer learning, will be tested to detect sentiment/emotionality patterns in our data.

9. Conclusions

The analyzed Glove (PCA) and fastText (t-SNE) embeddings confirm the close conceptual relation of insult and abusive. As abusive tends to mingle with stronger forms of offence, such as hostile (both in Glove/PCA and fastText/t-SNE), and racist (in Glove/PCA), in our typology it is situated at the bottom of the diagram, i.e. below insult, in the region where the strongest forms of offence are marshalled. The words profane, vulgar, obscene, hateful, homophobic and slur intermingle (which is most obvious in Glove/PCA, somewhat less conspicuous but still discernible in fastText/t-SNE) and create a single cloud of concepts, yet the cluster is not well delineated. However, the fact that they tend to occur in similar contexts speaks for their close conceptual affinity, and thus they manifest strong emotions and strong offence. Hence, in our typology vulgar goes together with these words (vulgar/obscene, vulgar/profane, vulgar/slur), and the other terms are also located at the lowest level in the diagram indicating strong offence. Taboo (in Glove/PCA), on the one hand, shares similarities both to weaker forms of offence, encoded by discredit and insult, and on the other, it tends to blend with stronger forms of offence expressed by vulgar, abusive, and cyberbullying. Thus, taboo seems to stand midway between weaker and stronger forms of offence yet with inclinations towards its stronger forms. This observation is corroborated by fastText/t-SNE wherein taboo is circumscribed by abusive and threat on the one side and insult on the other side; it intermingles with hostile and abusive, which seems to place taboo closer to the strong offence on the cline. The term offensive as depicted in the diagrams, particularly in the Glove-based embeddings, selects the nominal sense in terms of a military attack, almost uniquely, while the adjectival offensive as in e.g., an offensive remark, which would be more relevant for abusive language identification, did not surface in the consulted corpora, hence it was not possible here to identify the 'insulting' sense of offence in these embeddings. In the consulted SE data, on the other hand, the primary sense of offence in terms of abuse and insult was the dominant category.

Due to a large number of corpora and corpus annotation systems used for semantic analysis, and more specifically for offensive language identification, our linguistically grounded schema unifies and accounts for all the appropriate concepts applied in it. Since some of the terms proposed in previous studies may have been applied with insufficient precision, they have been dispensed with in our model or conflated with well-defined terms grounded in linguistics. The

proposed ontology will be further investigated for contextual embeddings within the existing annotated datasets (ELMo, BERT).

Acknowledgments

The present study has been conducted within the Use Case WG 4.1.1. Incivility in Media and Social Media, COST Action CA 18209 - European network for Web-centred linguistic data science Nexus Linguarum. The authors acknowledge the contribution of Milica Vuković Stamatović, Ana Ostroški, Branka Ivković, and Lobel Filipić to the identification of datasets and tagsets within the framework of WG 4.1.1; The project on which this report is based was partly funded by the German Federal Ministry of Education and Research (BMBF) under the funding code 01-S20049. The authors are responsible for the content of this publication.

References

[1] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, Predicting the type and target of offensive posts in social media, 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. 1415–1420.

[2] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval), in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 75–86.

[3] J. Gracia, C. Fäth, M. Hartung, M. Ionov, J. Bosque-Gil, S. Veríssimo, C. Chiarcos, M. Orlikowski, Leveraging linguistic linked data for cross-lingual model transfer in the pharmaceutical domain, in: International Semantic Web Conference, Springer, 2020, pp. 499–514.

[4] J. Qian, A. Bethke, Y. Liu, E. Belding, W. Y. Wang, A benchmark dataset for learning to intervene in online hate speech, arXiv preprint arXiv:1909.04251 (2019).

[5] R. Gomez, J. Gibert, L. Gomez, D. Karatzas, Exploring hate speech detection in multimodal publications, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 1470–1478.

[6] E. Wulczyn, N. Thain, L. Dixon, Ex machina: Personal attacks seen at scale, in: Proceedings of the 26th international conference on world wide web, 2017, pp. 1391–1399.

[7] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Language Resources and Evaluation (2020) 1–47.

[8] A. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, N. Kourtellis, Large scale crowdsourcing and characterization of twitter abusive behavior, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 12, 2018.

[9] A. H. Razavi, D. Inkpen, S. Uritsky, S. Matwin, Offensive language detection using multi-level classification, in: Canadian Conference on Artificial Intelligence, Springer, 2010, pp. 16–27.

[10] E. Methven, Offensive language crimes in law, media, and popular culture, in: Oxford Research Encyclopedia of Criminology and Criminal Justice, 2017.
[11] T. Jay, K. Janschewitz, The pragmatics of swearing, Journal of Politeness Research 4 (2008) 267–288.
[12] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, Y. Chang, Abusive language detection in online user content, in: Proceedings of the 25th International Conference on World Wide Web, 2016, p. 145–153.
[13] E. W. Pamungkas, V. Patti, Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon, in: Proceedings of the 57th annual meeting of the association for computational linguistics: Student research workshop, 2019, pp. 363–370.
[14] D. Kunupudi, S. Godbole, P. Kumar, S. Pai, Toxic language detection using robust filters, SMU Data Science Review 3 (2020) 12.
[15] T. Caselli, V. Basile, J. Mitrović, I. Kartoziya, M. Granitzer, I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language, in: Proceedings of The 12th Language Resources and Evaluation Conference, 2020, pp. 6193–6202.
[16] P. Wilson, L.-T. Barbara, The Role of Devirtualization from Online to Offline Spaces in the Honour-Killing of Qandeel Baloch, 2021. International Linguistic and Social Aspects of Hate Speech in Modern Societies Conference, 22-23 March, 2021, Odense University.
[17] V. Basile, C. Bosco, E. Fersini, N. Debora, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti, et al., Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter, in: 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 54–63.
[18] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, Ç. Çöltekin, Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020), in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 1425–1447.
[19] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: Proceedings of the NAACL student research workshop, 2016, pp. 88–93.
[20] R. Kshirsagar, T. Cukuvac, K. McKeown, S. McGregor, Predictive embeddings for hate speech detection on twitter, in: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), 2018, pp. 26–32.
[21] P. Mishra, H. Yannakoudakis, E. Shutova, Neural character-based composition models for abuse detection, in: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), 2018, pp. 1–10.
[22] C. Liebeskind, S. Liebeskind, Identifying abusive comments in hebrew facebook, in: 2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE), 2018, pp. 1–5. doi:10.1109/ICSEE.2018.8646190.
[23] B. Birkeneder, J. Mitrovic, J. Niemeier, L. Teubert, S. Handschuh, upinf-offensive language detection in german tweets, in: Proceedings of the GermEval 2018 Workshop, 2018, pp. 71–78.
[24] J. Mitrović, B. Birkeneder, M. Granitzer, nlpup at semeval-2019 task 6: A deep neural language model for offensive language detection, in: Proceedings of the 13th International

Workshop on Semantic Evaluation, 2019, pp. 722–726.
[25] P. Liu, W. Li, L. Zou, Nuli at semeval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers, in: Proceedings of the 13th international workshop on semantic evaluation, 2019, pp. 87–91.
[26] S. D. Swamy, A. Jamatia, B. Gambäck, Studying generalisability across abusive language detection datasets, in: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), 2019, pp. 940–950.
[27] T. Caselli, V. Basile, J. Mitrović, M. Granitzer, Hatebert: Retraining bert for abusive language detection in english, arXiv preprint arXiv:2010.12472 (2020).
[28] E. Bassignana, V. Basile, V. Patti, Hurtlex: A multilingual lexicon of words to hurt, in: 5th Italian Conference on Computational Linguistics, CLiC-it 2018, volume 2253, CEUR-WS, 2018, pp. 1–6.
[29] R. Stankovic, J. Mitrović, D. Jokic, C. Krstev, Multi-word expressions for abusive speech detection in serbian, in: Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons, 2020, pp. 74–84.
[30] J. P. McCrae, I. Wood, A. Hicks, The colloquial wordnet: Extending princeton wordnet with neologisms, in: J. Gracia, F. Bond, J. P. McCrae, P. Buitelaar, C. Chiarcos, S. Hellmann (Eds.), Language, Data, and Knowledge, Springer, Cham, 2017, pp. 194–202.
[31] C. Fellbaum, Wordnet, An Electronic Lexical Database (Language, Speech and Communication) (1998).
[32] K. Kumaresan, K. Vidanage, Hatesense: Tackling ambiguity in hate speech detection, in: 2019 National Information Technology Conference (NITC), IEEE, 2019, pp. 20–26.
[33] D. Battistelli, C. Bruneau, V. Dragos, Building a formal model for hate detection in french corpora, Procedia Computer Science 176 (2020) 2358–2365.
[34] D. Biber, S. Johansson, G. Leech, S. Conrad, E. Finegan, Longman grammar of spoken and written english (lgswe), Harlow: Pearson Education (1999).
[35] A. Bączkowska, Learner corpus of subtitles and subtitler training, Subtitling Today: Shapes and Their Meanings (2015) 221–247.
[36] L. G. Andersson, P. Trudgill, Cultural approach to interpersonal communication, Blackwell, Oxford, 2007, pp. 195–199.
[37] K. Norman, The ironic body obscene joking among swedish working-class women, Ethnos 59 (1994) 187–211.
[38] N. Daly, J. Holmes, J. Newton, M. Stubbe, Expletives as solidarity signals in ftas on the factory floor, Journal of Pragmatics 36 (2004) 945–964.
[39] P. Trudgill, Sociolinguistics, Harmondsworth: Penguin (1974).
[40] Y. Baruch, S. Jenkins, Swearing at work and permissive leadership culture: When anti-social becomes social and incivility is acceptable, Leadership & Organization Development Journal 28 (2007) 492.
[41] A. Pavlenko, Emotion and emotion-laden words in the bilingual lexicon, Bilingualism 11 (2008) 147–164.
[42] V. De Klerk, The role of expletives in the construction of masculinity, Language and masculinity (1997) 144–158.
[43] C. Gregory, Among the dockhands: Another look at working-class male culture, Men and Masculinities 9 (2006) 252–260.

[44] K. Stapleton, 12. Swearing, De Gruyter Mouton, 2010, pp. 289–306. URL: https://doi.org/10.1515/9783110214338.2.289. doi:doi:10.1515/9783110214338.2.289.
[45] E. Goffman, Forms of talk, University of Pennsylvania Press, 1981.
[46] G. Kress, T. Van Leeuwen, Multimodal discourse, The modes and media of contemporary communication.(Cappelen, London 2001) (2001).
[47] E. Ventola, C. Charles, M. Kaltenbacher, Perspectives on multimodality, volume 6, John Benjamins Publishing, 2004.
[48] K. L. O'Halloran, Multimodal discourse analysis: Systemic-functional perspectives, Bloomsbury Publishing, 2006.
[49] M. Jakubíček, A. Kilgarriff, V. Kovář, P. Rychlỳ, V. Suchomel, The tenten corpus family, in: 7th International Corpus Linguistics Conference CL, 2013, pp. 125–127.
[50] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).
[51] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.
[52] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146.
[53] L. Van der Maaten, G. Hinton, Visualizing data using t-sne., Journal of machine learning research 9 (2008) 2579–—2605.
[54] T. Declerck, J. P. McCrae, M. Hartung, J. Gracia, C. Chiarcos, E. Montiel-Ponsoda, P. Cimiano, A. Revenko, R. Sauri, D. Lee, et al., Recent developments for the linguistic linked open data infrastructure, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 5660–5667.