# Ontological Analysis of E-News: A Case for Terrorism Domain

Svetlana Sheremetyeva[1], Anastasia Zinoveva[2]

South Ural State University
[1]sheremetyevaso@susu.ru
[2]zinovevaaiu@bk.ru

**Abstract.** This paper reports on an on-going project aimed at developing a model of multilingual ontological analysis of e-news. The research methodology is data-driven and involves several interwoven stages directed from analysis to representation: extraction and semantic classification of lexical units from three comparable corpora of e-news on terrorism in the English, French, and Russian languages, construction of a core ontology and its application to the ontological analysis of terrorist e-news. The development procedures are described through the terrorist domain case study. Special attention is paid to ontological concept metrics that can facilitate disambiguation in lexical-ontological mappings. The findings are illustrated by applying the developed multilingual model to the ontological analysis of e-news on terrorism in the French language.

**Keywords:** ontological analysis, domain ontology, e-news, terrorism

## 1 Introduction

With the advent of the public Internet, electronic news on terrorism has been a subject of electronic management as an essential part of counter-terrorism. Nowadays, most techniques for classification, search, information extraction, question answering, content analysis, etc. applied to e-news mainly rely on shallow text mining or parsing, without deep linguistic analysis due to the complexity of the latter. However, as it is widely recognized, high-quality solutions for information processing tasks can only be obtained with proper meaning understanding, for which ontological analysis is recognized to be the most promising [4].

Ontological analysis is defined as the study of content as such, or more specifically, as the process of eliciting content knowledge on the entities involved in a certain domain. In practice, ontological analysis consists in mapping lexical units of textual information into an ontology followed by formalizing and interpreting the results of such mapping depending on the particular task in question.

Independently of whether ontological analysis is done manually or involves automation (which is a separate problem), it has serious limitations. The first one is the availability of an appropriate pre-defined and well-established ontology**.** Though



In: P. Sosnin, V. Maklaev, E. Sosnina (eds.): Proceedings of the IS-2019 Conference, Ulyanovsk, Russia, 24-27 September 2019, published at http://ceur-ws.org

quite a number of ontological libraries are currently publicly available, their suitabil-suitability for every particular R&D project involving ontological analysis is, as a rule, problematic. Therefore, in most works on ontological analysis, the first task (that can also be the goal of the research) is to build a domain- and/or application-tuned ontology. The second major limitation lies in the practical realization of ontological analysis as such with the focus on the bi-directional mapping of the textual elements with the ontological concepts. The shortcomings here are well-known and pertain to both the process of ontology building and subsequent application of the ontology to document analysis. It is the difficulty of clearly specifying the boundaries of the analysis as well as the limited consideration of relationships between the ontological concepts. Text elements can be missing in the ontology mapping or one-to-many, many-to-one or many-to-many relationships exist, which leaves ambiguities unresolved. Then, the procedure of the ontological analysis initially done by humans based on objective judgments can influence the results of the analysis [4]. There is no universal recipe for ideal ontological analysis and, as a rule, in every practical project, specific approaches are developed to deal with the problems.

In this paper, we describe our experience in developing a model of multilingual ontological analysis that is data-driven and investigate ontological metrics. We illustrate our findings by applying the developed model to the ontological analysis of e-news on terrorism in the French language. The rest of the paper is organized as follows. Section 2 gives an overview of major trends in ontological analysis of e-news. Section 3 describes our methodology. Section 4 presents our multilingual ontological resource tuned to the terrorism domain. In section 5, the workflow of ontological analysis and ontology metrics findings are described on the example of French-language e-news on terrorism. We conclude with a summary and future work.

## 2 Related work

Ontological analysis involves the comparison of unstructured text with ontologies, followed by the semantic annotation of text elements. The number of works on ontologies and ontological analysis has drastically increased since 2001 when the Semantic Web was popularized [2] with its promise of data interoperability at the semantic level. Quite a number of research projects concentrate on ontology-based techniques for e-news classification systems. For example, the ePaper system reported in [11] uses the ontology as a common language for content-based personalized e-news filtering, while in the NEWS system, ontological knowledge is meant to support content-based classification in three languages: English, Spanish and Italian [3].

Ontological analysis of terrorism domain that has already acquired a large body of studies was boosted by proliferation of e-news on terrorism online from thousands of different sources. The scope of R&D in this field ranges from linguistic and methodological issues to tools and actual knowledge bases that are mainly application-specific and focus on certain limited aspects of the domain. For instance, the PiT (Profiles in Terror) ontology [6] is designed to represent knowledge about the terrorist network, which includes a set of individuals and organizations, as well as

numerous relations between them. Another terrorism ontology, AIT (Adversary–Intent–Target) is designed to predict terror attacks based on data on terrorist organizations, their intentions, and weapons [12]. In [5], ontological analysis based on built-in-house Terrorism Ontology for terrorism event extraction from Thai e-news is described. The development of the core of the RiskTrack domain ontology, which defines the radicalization indicators and incorporates important information about existing terrorist organizations and groups, is presented in [1]. The work [7] is devoted to ontology-related research for the prediction of the terrorist threat on the basis of semantic association acquisition and complex network evolution.

## 3 Methodology

Our research methodology is data-driven and involves several interwoven stages directed from analysis to representation. The road map for this research is shown in Fig. 1. First, the data set of this study — three-language (Russian, French, and English) corpora of the e-news articles from the web — were acquired and divided into two parts for training and testing. Next, the terrorism-domain-relevant lexical items from the training corpora were extracted for semantic classification and decision on the set of ontology concepts encompassing all the three languages. Then, the upper-level ontology and representation formalism were decided on followed by the development of the seed e-news terrorism ontology and lists of lexical items from the training corpora that map into ontology concepts.

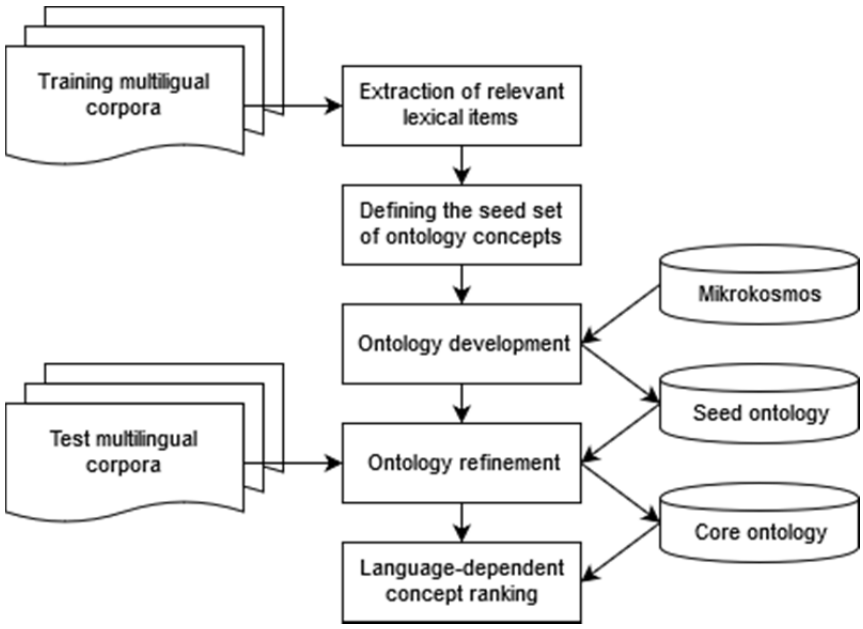


**Fig. 1.** The road map of e-news ontological analysis

The seed ontology is applied to the analysis of the testing corpora and is refined into the core ontology for the terrorist e-news. We then calculate language-dependent ranks of ontology concepts that can be used for semantic tags disambiguation, trend mining and ontology results interpretation, e.g., for the identification of the terrorism perception national specificity. We base our research on the following methodological assumptions:

- An ontology is a reusable language-independent source, hence a good intermediary between multilingual lexicons.
- Domain-specific knowledge is an integral part of general world knowledge. Therefore, a domain ontology should be linked to an upper ontology.
- A mixed ontology knowledge acquisition technique is the most appropriate for our task, as we define key concepts first based on corpus lexical data and then specify and/or generalize them to obtain more detailed or abstract concepts.
- The boundaries of the analysis and the limited consideration of the sets of the ontological concepts and relationships between them are data-driven.

## 4 Building ontology for e-news on terrorism

### 4.1 Data set analysis

Our data set consists of three domain corpora of e-news in the French, English, and Russian languages ca. 500,000 words each acquired from the web. The scope of topics covers e-news about terror attacks all over the world. The corpora were further divided into training and testing corpora.

The acquisition and analysis of the corpora were performed semi-automatically with the use of built-in-house tools, such as a web crawler, an automatic extractor of multiword typed expressions [10], and manual application of component analysis, opposition analysis, and text template analysis, etc., see [9] for the details. We thus obtained seed sets of multicomponent typed phrases (NPs, VPs, AdjPs, etc.) of up to ten-component length. The extracted phrases were grouped into semantic classes and subclasses based on their semantic similarity in the corpora. Note that attributing some of the phrases to a certain semantic class was purely corpus-based, which was the case, for example, with the word *attaque* that does not generally imply a terrorist meaning without a terrorism-domain attribute, e.g., *terroriste*. However, in the corpora, the word *attaque* alone was frequently used to refer to a terror attack specifically. This does not exclude lexical ambiguity even in the domain corpora, which leads to the overlapping lists of different classes. For example, the French named entity *Charlie Hebdo* can mean both OBJECT OF ATTACK (its office was targeted by terrorists in 2015) and SOURCE (it is a weekly newspaper).

This stage of analysis resulted in the set of key object concepts and subconcepts of the domain with their main attributes and relationships. The pool of concepts and relations was further augmented by analyzing the corpora with text templates (or patterns). For instance, in the French-language corpus, the RELATION concept IS-A can

be detected (though not exclusively) by means of the following text templates: A *est* B, B *comme* A, A et *autres* B, wherein B stands for a parent concept, while A is a child concept. The RELATION concept INSTRUMENT can be manifested in the following French-language templates: *attaque* / *attentat avec* / *á* / *au moyen de* A, wherein A stands for a weapon type.

Table 1 shows a fragment of the list of upper-level concepts acquired for the seed terrorism ontology with their definitions illustrated by French lexemes linked to the concepts. To give examples of subclasses, the concept TERROR ATTACK was subdivided into BOMB ATTACK, SUICIDE ATTACK, VEHICLE-RAMMING ATTACK, ARMED ATTACK, CHEMICAL ATTACK, HOSTAGE-TAKING, PSYCHOLOGICAL PRESSURE, and ARSON concepts, while the concept CONSEQUENCES was subclassified into CONSEQUENCES FOR PEOPLE, POSITIVE CONSEQUENCES FOR TERRORISTS, NEGATIVE CONSEQUENCES FOR TERRORISTS, and DAMAGE FOR BUILDINGS.

For the representation of our ontological knowledge, we decided on the formalism of the Mikrokosmos ontology [8] and used it as our upper-level ontology following the Mikrokosmos division of the reality into OBJECTS, EVENTS, and PROPERTIES (RELATIONS and ATTRIBUTES) linking our domain concepts to the appropriate Mikrokosmos parent nodes. We also keep concept labels worded in English. The resulting resource is called the terrorism-domain seed ontology, a fragment of which is shown in Fig. 2.

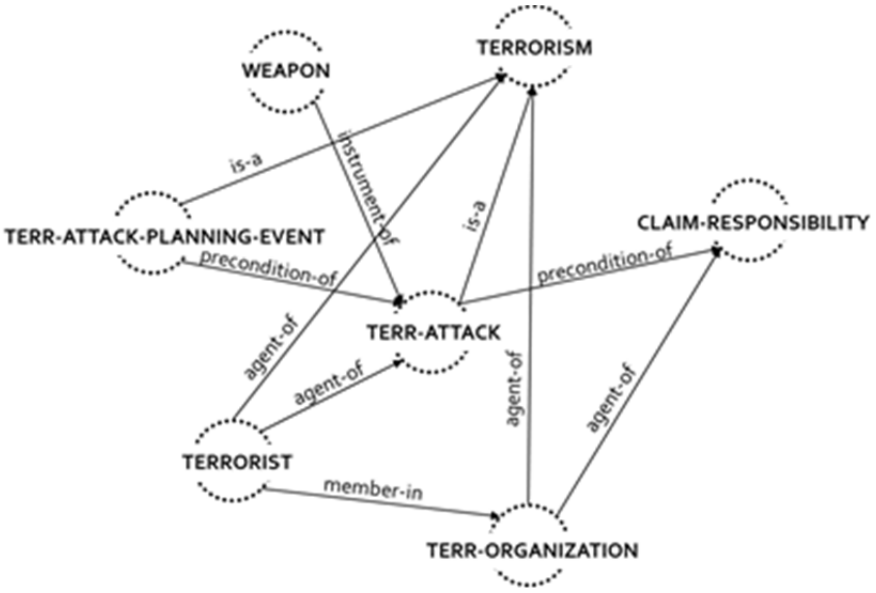


**Fig. 2.** A fragment of the terrorism-domain seed ontology

Based on the lexical-ontological knowledge acquired at this stage, we have developed a platform with flexible settings that allow knowledge administration and different analysis depth to automate tagging texts with ontological concepts.

**Table 1.** A fragment of the list of the seed ontology concepts

| Concept | Definition | Lexical examples |
|---|---|---|
| ADVERSARY'S PLANS. | Intended activities of a terrorist or a terrorist group. | Planification d'attentat terroriste |
| AGENT | The perpetrator of the attack. | Terroriste, combattant, femme kamikaze |
| ASSUMPTION | Assumptions of "good guys" about a probable terrorist group behind the attack or a suspect. | Attribué, présumé, suspecté |
| CONSEQUENCES | All the results of the terrorist attack, such as human victims, destroyed objects, terrorists' destiny, and the condition of those. | Femme, homme, personne, policier turc, terroriste, blessé, mort, otage, tué, neutralisé |
| CHARACTER OF ATTACK | The concept indicates whether the victims of the attack were numerous and one person was the only target. | Meurtier, sanglant, tuerie, carnage, assassiné |
| GOAL OF ATTACK | The goal terrorists are trying to achieve by committing the attack. It can also be used to indicate the reason for the attack as sometimes it is hard to distinguish between them. | Renverser le gouvernement, assassiner des juifs, causer un grand nombre de victimes, venger le drame de la ville d'Alep |
| LOCATION | The country, region, city, district, or geographical entity where the attack took place. | À environ 5 km du bâtiment de la police, Afghanistan, Etat du Minnesota, frontière syrienne |
| MEANS OF ATTACK | The weapons or weapon-like objects (e. g., a truck) used to commit the attack, also functional weapon parts, such as explosives, bullets, etc. | Arme à feu, camion, ceinture, couteau, véhicule |
| OBJECT OF ATTACK | The animate or inanimate object the attack is directed to, which is hurt or damaged in the attack. | Convoi militaire, discothèque, école, église, endroit très fréquenté, femme |
| SOURCE | The sources of the message about the attack, such as newspapers, TV channels, news agencies, or authorities. | Agence de presse Reuters, Al-Jazeera, ambulanciers, autorités israéliennes, CNN Türk, témoins |
| TERRORIST ORGANIZATION | The organization responsible for the attack or any terrorist organization mentioned in the text. | Al Qaïda, Daech, Faucons de la liberté du Kurdistan, talibans |
| TIME | The time and date of the attack. | À la veille du Nouvel An, au cours de la nuit |
| TYPE OF | The type of attack, such as an | Acte terroriste, attentat, |

| | | |
|---|---|---|
| ATTACK | explosion, kidnapping, arson, etc. | attaque au camion belier, explosion |

### 4.2 Ontology refinement

At this stage, we automatically tagged the testing part of our corpora with the seed ontology concept tags and analyzed a list of lexical units left untagged. Terrorism-related lexical items were further mapped to either the existing ontology concepts or to new ones that were added to the ontology following the results of the analysis. Some of the newly added concepts with the examples of French lexical units mapped to them are shown in Table 2.

**Table 2**. A fragment of the list of newly added concepts

| Concept | Definition | Lexical examples |
|---|---|---|
| CLAIM RESPONSIBILITY | To claim responsibility for an attack. | Revendiquer, prendre la responsibilité |
| DECLARATION | To say, to declare, to announce (the concept is normally linked to verbs and adverbial phrases that mean the transfer of information). | Ajouter, citer, commenter, dire, indiquer, rapporter, selon |
| DIRECTION OF ATTACK | To target smth. or smb. | Viser, être la cible, cibler, touché |
| HAVE MEANS OF ATTACK | To have a weapon or a weapon-like object (the concept is normally linked to verbal phrases that mean the process of application of MEANS OF ATTACK). | Armé, chargé |
| NATION | The origin of terrorist and victims; it should not be confused with LOCATION, which only covers the places where particular attacks were committed. | Turc, kurde, russe, franco-tunisienne, de nationalité française |
| OTHER TERRORIST ACTIVITIES | Types of terrorist activities that are not literary terror attacks, e.g., terrorism financing, recruiting, involvement in war conflicts, etc., but appear sporadically in terrorism domain e-news and are therefore considered relevant. | Guerre syrienne, combattre, financer le terrorisme |

The multilingual e-news terrorism-domain core ontology was thus created, which currently contains 107 OBJECT and EVENT concepts, 20 RELATION concepts, and 7 ATTRIBUTE concepts. Created also were the terrorism-domain-related lexicons in French, English, and Russian, mapped to the ontology concepts. Fragments of

multilingual lexical lists mapped to the TERROR ATTACK concept are shown in Table 3 (absolute frequencies in the corpora are provided in brackets).

The new data obtained at this stage were added to the knowledge base of the ontological tagging platform.

**Table 3.** Fragments of multilingual lexical lists mapped to the TERROR ATTACK concept

| English (F) | French (F) | Russian (F) |
|---|---|---|
| attack (1218) | attentat (2447) | теракт (2839) |
| terrorist attack (202) | attentat terroriste (369) | стрельба (210) |
| bombing (136) | attaque terroriste (345) | террористический акт (179) |
| terror attack (129) | attentat-suicide (128) | террористическая атака (93) |
| act of terrorism (59) | fusillade (126) | акт терроризма (30) |
| shooting (39) | acte terroriste (110) | двойной теракт (21) |
| gun attack (23) | attentat suicide (58) | захват заложников (15) |
| terrorist act (16) | attentat à la bombe (37) | взрыв бомбы (11) |
| knife-attack (16) | prise des otages (29) | поджог (6) |
| lone-wolf attack (10) | acte de terrorisme (16) | угон самолета (5) |
| act of terror (8) | attentat à l'explosif (6) | атака смертника (3) |
| terror act (3) | agression terroriste (3) | акт террора (2) |

## 5 Language-dependent concept ranking

In this section, we describe the feasibility study of the language-dependent concept ranking procedure as exemplified by its application to the 20,000-wordform French e-news corpus that was randomly cut out of the initial corpus used to build the core ontology and the French terrorism-related concept-mapped lexicon. The lexicon, so far, contains 1,334 lexical units (single words and multicomponent phrases) and 21 high-level concepts of our core ontology, which we at this stage use in our calculations.

In general, this stage of research was motivated by our hypothesis that the extent, to which multilingual ontology concepts are used to code (tag) lexical meanings in the domain texts, differs, and it was worth investigating this issue as applied to every national corpus. The idea is that these findings might contribute to solving the major information processing problem, — ambiguity, in particular, concept-mapping ambiguity that is relevant, e.g., for information extraction or question answering.

The concept-mapping ambiguity problem in the French terrorism domain can be illustrated by the word *policier (police officer)* that maps to the following concepts:

OBJECT OF ATTACK: *Un **policier** est tué. (A **police officer** was killed.)*

SOURCE: *Selon des **policiers**, l'homme aurait crié « Allah akbar ». (According to **police**, the man shouted "Allahu akbar".)*

COUNTER-TERRORISM: *Après les explosions, les autorités ont déployé des **policiers.** (After the explosions, the authorities deployed **police officers**.)*

AGENT: *L'ambassadeur de Russie est assassiné à Ankara par un* ***policie****r (Russia's ambassador is assassinated in Ankara by a* ***policeman.)***

This means that on tagging, this word will be assigned four concept tags, hence the need for disambiguation. The straightforward solution will be to use the sentence context; however, it requires a lot of knowledge and might not always give correct results. The situation can be improved by means of quantitative parameters. We attempted just this. The feasibility study corpus was automatically concept-tagged with the following tags: P = CONSEQUENCES, L = LOCATION, Z = OBJECT OF ATTACK, T = TYPE OF ATTACK, S = SOURCE, B = TIME, D = DECLARATION, A = AGENT, R = COUNTER-TERRORISM, U = TERRORIST ORGANIZATION, C = MEANS OF ATTACK, CR = CLAIM RESPONSIBILITY, N = NATION, DA = DIRECTION OF ATTACK, I = ASSUMPTION, M = CHARACTER OF ATTACK, O = OTHER, E = OTHER TERRORIST ACTIVITIES, X = GOAL OF ATTACK, HA = HAVE MEANS OF ATTACK.

Then, the concept frequencies (CF), i.e. the numbers of occurrences of every concept tag in this corpus, were calculated. Fig. 3 shows the frequency distribution of the key ontology concepts in the French terrorism corpus.

We then calculated the frequency of concept multitags (the number of words that were assigned more than one concept tag) and discovered that multiple tags amount to 9.67% of the total concept tag frequency, which shows that the concept ambiguity rate as applied to the French corpus is quite high. The frequency distribution of multiple concept tags is shown in Fig. 4 and can directly be used to set priorities when developing disambiguation procedures.

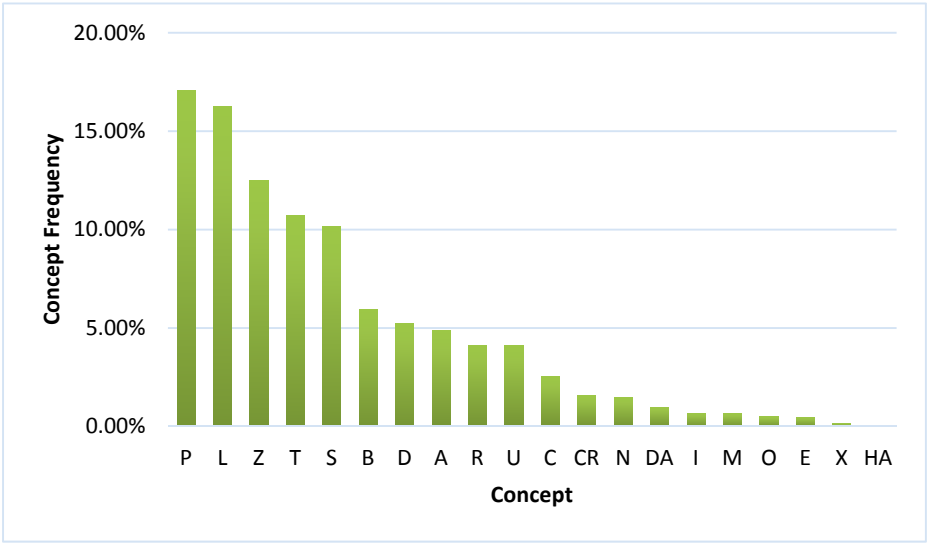


**Fig. 3.** Distribution of concept frequency in the French corpus; 100% is the total number of ontology-mapped lexical units in the corpus in question

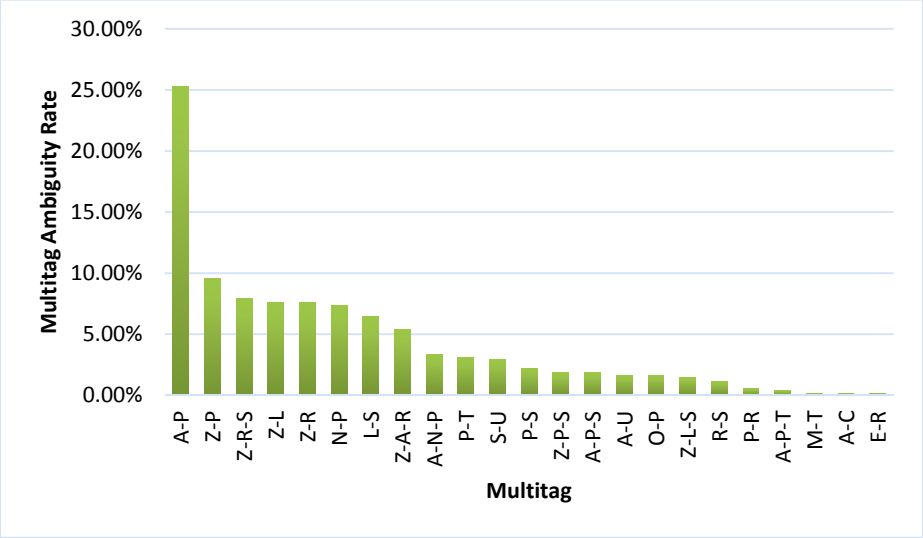


**Fig. 4.** Distribution of the multilingual concept multitags in the French corpus; 100% is the total number of multitags

To have a larger feature space for finer grain ontology concept ranking and disambiguation, two more measures were introduced, — the ratio of concept fillers (RCF) and the concept usage relevancy (CUR). The ratio of concept fillers accounts for the variety of lexical units mapped into a concept and is calculated as follows:

RCF = n/N,

where n is the number of the ontology-linked unilingual (French in our example) lexicon items mapped into a particular concept and N is the total number of items in the ontology-linked unilingual (French in our example) lexicon.

The concept usage relevancy measure in our research is considered to depend on the ratio of concept fillers and the concept frequency when applied to code (tag) a lexeme sense in a unilingual concept corpus. The empirical formula we used at the current stage of research is given below.

CUR = (RCF*10+CF)/T,

where CUR is a concept usage relevancy, RCF is a ratio of concept fillers, CF is a concept frequency and T is the number of wordforms in a unilingual corpus (French in our example).

Ranking ontology concepts according to their CUR measure can be helpful in developing heuristics for concept multitags disambiguation by taking into account the

CUR values calculated for every ontological concept as related to a particular unilin-unilingual corpus. The higher a concept CUR value, the more prioritized its tag can be in the set of the other ones assigned to the same lexical unit. Table 4 shows the values of the suggested measures for the concepts whose tags are included in the first three most frequent multitags shown in Fig. 4. According to the calculated CUR values, the multitags A-P, Z-P, Z-R-S can most probably be disambiguated as P, P and S, correspondingly. We are fully aware that much more research should be done in this direction and in practice a number of different disambiguating parameters might need to be used, but it follows from our findings that the CUR measure could definitely be at least one of them.

**Table 4.** Values of RCF, CF and CUR measures for selected concepts

| Concept | RCF | CF | CUR |
|---|---|---|---|
| A (AGENT) | 0.03 | 0.05 | 1.56 |
| P (CONSEQUENCES) | 1.02 | 0.17 | 6.68 |
| R (COUNTER-TERRORISM) | 0.05 | 0.04 | 2.84 |
| S (SOURCE) | 1.04 | 0.10 | 7.66 |
| Z (OBJECT OF ATTACK) | 1.19 | 0.13 | 6.47 |

## 6 Conclusions

We have presented an ongoing project aimed at the development of an ontological analysis model for multilingual e-news on terrorism. The research covers the acquisition of ontological knowledge and its formal representation based on the data extracted from the multilingual corpus of English, French, and Russian e-news on terrorism. The proposed methodology for ontology development is based on extracting multicomponent lexical units from unilingual corpora, grouping them into semantic classes and using textual templates to enlarge the ontology-related knowledge. Language-dependent knowledge thus obtained is further accumulated into a single ontological resource with language-dependent ontology-mapped lexicons. The methodology can most likely be used on the material of a broader set of languages that would, of course, include the development of corresponding language-dependent textual templates.

We have made an attempt to contribute to solving the lexical-concept mapping ambiguity problem by calculating frequency-related parameters of ontology concepts as applied to the ontological analysis of a unilingual corpus. Two new quantitative measures, a ratio of concept fillers and a concept usage relevancy, were introduced. Our findings show that these measures could definitely be used as at least one of the disambiguating parameters, though we are fully aware that much more research should be done in this direction. We, therefore, see it as our future work.

We will also proceed with enlarging both the depth and the breadth of the ontology and the size of language-dependent ontology-mapped lexicons as well as refining the ontological analysis model.

## References


1. Barhamgi, M., Faci, N., Masmoudi, A. RiskTrack Ontology for On-line Radicalization Domain (2018), available at: https://projet.liris.cnrs.fr/radicali/
2. Berners-Lee, T., Hendler, J., Lassila, O. The Semantic Web. Scientific American 284 (5), pp. 34–43 (2001)
3. Fernández, N., Fuentes, D., Sánchez, L., Fisteus, J.A. The NEWS ontology: Design and applications. Expert Systems with Applications 37 (12), pp. 8694–8704 (2010). DOI: 10.1016/j.eswa.2010.06.055
4. Green, P.S., Rosemann, M., Indulska, M. The Practice of Ontological Analysis (2005), available at: https://pdfs.semanticscholar.org/513c/a04a8132a723cf47d9d9504983a98dd9ec08.pdf
5. Inyaem, U., Haruechaiyasak, Ch., Meesad, Ph., Tran, D. Ontology-Based Terrorism Event Extraction. Proceedings of the 1st International Conference on Information Science and Engineering, pp. 912–915, Nanjing, China (2009). DOI: 10.1109/ICISE.2009.804
6. Mannes, A. J. Golbeck. Building a Terrorism Ontology. ISWC Workshop on Ontology Patterns for the Semantic Web 36 (2005), available at: https://pdfs.semanticscholar.org/9bcb/90e48677e39da7b84939e8c8da2b2a63cde7.pdf
7. Najgebauer, A. Antkiewicz, R., Chmielewski, M., Kasprzyk, R. The Prediction of Terrorist Threat on the basis of Semantic Association acquisition and Complex Network Evolution. Journal of Telecommunications and Information Technology 2, pp. 14–20 (2008)
8. Nirenburg, S., Raskin, V.: Ontological Semantics. MIT Press, Cambridge (2004)
9. Sheremetyeva, S., Zinovyeva, A. On Modelling Domain Ontology Knowledge for Processing Multilingual Texts of Terroristic Content. Communications in Computer and Information Science 859. Springer, Cham, pp. 368–379 (2018). DOI: 10.1007/978-3-030-02846-6_30
10. Sheremetyeva, S.: Automatic Extraction of Linguistic Resources in Multiple Languages. Proceedings of NLPCS 2012, 9th International Workshop on Natural Language Processing and Cognitive Science in conjunction with ICEIS 2012, pp. 44–52. Wroclaw, Poland (2012)
11. Tenenboim, L., Shapira, B., Shoval, P. Ontology-Based Classification of News in an Electronic Newspaper. International Book Series “Information Science and Computing”, pp. 89–97 (2008)
12. Turner, M.D., Turner, J., Weinberg, D. A Simple Ontology for the Analysis of Terrorist Attacks (2011), available at: https://digitalrepository.unm.edu/ece_rpts/41