Hybrid extraction of multi-word terms: an application on vibration-based condition monitoring technique

Konstantinos Chatzitheodorou[1], **Vassilios Kappatos**[2]
[1]Department of Foreign Languages, Translation and Interpreting, Ionian University, Corfu, GR49100, Greece
[2]Hellenic Institute of Transport, Centre for Research and Technology Hellas, 6th Km Charilaou Thermi, 60361, Thermi, Thessaloniki, Greece
[1]Corresponding author
E-mail: [1]*kchatzitheodorou@ionio.gr*, [2]*vkappatos@certh.gr*

Received 27 December 2020; accepted 10 January 2021
DOI https://doi.org/10.21595/mme.2021.21850

Abstract. In this paper, we present an advanced domain-specific multi-word terminology extraction method. Our hybrid approach for automatic term identification benefits from both statistical and linguistic approaches. Our main goal is to reduce as much as possible the human effort in term selection tasks as well as to provide a wide-range and representative terminology of a domain. We emphasize in identification of verb or noun phrases multi-word terms, in neologisms and technical jargons. Our architecture applies the term frequency-inverse document frequency (TF-IDF) algorithm to a domain-specific textual corpus in order to measure a unit's importance in it. We also use techniques to filter out nested terms of a candidate term taking into consideration its frequency by itself in the corpus. In addition, the exported terms are filtered out based on a stop-word list and linguistic criteria. To further reduce the size of the candidate terms and achieve accurate and precise terminologies, our method automatically validates them against a general-purpose corpus. Our study based on a small corpus of vibration-based condition monitoring domain shows that most extracted terms have nice correspondence to the domain of condition monitoring concepts and notions.

Keywords: hybrid extraction, terminology, multi-word terms, condition monitoring, vibration, specialized languages, term-base, technical language.

1. Introduction

Every scientific area makes use of a special vocabulary to convey specialized concepts by means of technical language. This special vocabulary contains a wealth of information that was gathered over the years in a particular domain. Terminologies, which are the lexical components of specialized languages, have value in the way they condense mass of information into single-word, multi-word or compound word units. They are a crucial component of both technical and scientific writing, ensuring more effective communication. Terminology identification and documentation is a time consuming task, requiring manual validation by humans. It is also subjective and depends on the experience and criteria of the experts who validate it.

Condition monitoring (CM) utilizes specific measurement of equipment parameters, such as the vibrations in a machine, its temperature, or condition of its oil, assessing significant changes (if any) that may be indicative of an impending failure. CM offers a multitude of benefits, including 24×7 remote monitoring, early warnings of potential (serious) failures, reduced unplanned downtime, reduction in maintenance costs, and support for ongoing reliability and risk reduction with fewer inspections. The vibration-based CM is a system which uses non-destructive non-destructive sensing and analysis of the characteristics of the system in the time, frequency or modal domains for the purpose of detecting a change that can indicate damage or degradation.

Different technical terms, jargon, acronyms, and concepts are used in vibration analysis domain. There is a lot of semantic confusion in the area of vibration-based CM when general or

vague terms like *frequency*, *frequency response*, *frequency response function*, etc. are used. Specifically, vocabularies referring to a specific condition are used interchangeably with terms denoting general dysfunction and vice versa. These confusions may reflect research that has changed dramatically in the last decades. We aim to eliminate this confusion by providing a stable and unambiguous terminology work of vibration-based CM.

While terminology recognition, i.e., development of a list with domain specific vocabulary from a textual corpus, has become widespread presently, the notion itself of term is nevertheless not clear, both from a pure linguistic and a computational perspective [1]. Juan C. Sager [2] defines a term as a depository of knowledge and a unit with specific reference in that it "refers to discrete conceptual entities, properties, activities or relations, which constitute the knowledge space of a particular subject field". Similarly, Christian Jacquemin [3] and Maria Teresa Pazienza [4] define a term as "a surface representation of a specific domain concept".

A considerable amount of work has been done in order to define terms for a specific domain. They use statistical methods to extract terminologies from textual corpora, which is a collection of documents in the domain. Many of the results are completely useful. Gerard Salton [5] proposed TF-IDF. It is a statistical measure used to evaluate how relevant a term is to a document within a collection of documents. This is done by considering: (i) how many times a term appears in a document; and (ii) how frequent a term appears across a set of documents in the reverse order. Later, Kenneth Church and Patrick Hanks [6] suggested the mutual information (MI) which is an important concept in information theory. Particularly, it measures how much information, on average, is communicated in one random variable about another. Both variables are sampled at the same time. Ted Dunning [7] proposed a probability measure, which relies on frequency profiling and is based on likelihood ratios. The method produces reasonable results and works for both large and small text samples. Relatively recently, researchers have captured n-grams (i.e., occurrences of one or more words) using corpus comparison based on normalized frequency. For instance, Adam Kilgarrif [8] experiments with several measures, among others $\chi 2$-test, Mann-Whitney rank, t-test, MI, and TF-IDF and concludes that $\chi 2$-test performs best. However, in general, most methods produce endless lists of candidate terms.

In this paper, we present a method for the automatic extraction of multi-word terms from machine-readable textual corpora. We apply the method to a corpus of the vibration-based CM technique domain, which is a keytool in the predictive maintenance of any equipment and machines used in a wide range of industries including transport, oil and gas, processing and manufacturing. The key idea of our approach is to reduce the noisy terms as well as the out-of-domain terms by using two different corpora: a) a domain specific corpus and b) a general purpose corpus. Our experimental results show that our method can extract multi-words terms, i.e., a set of two or more words, with nice correspondence to domain-specific concepts and notions. Thus, the human effort is reduced and it is subjective independent. The rest of this paper is structured as follows: we analyze our terminology extraction method in Section 2. Section 3 presents the results of our experiment as well as the error analysis. Finally, Section 4 concludes our work.

2. Our terminology extraction approach

Automatic term extraction is an open problem of natural language processing. The task is to identify and extract terminological units from textual corpora. In this paper, we extend the methodology proposed by Patrick Drouin [9], where the extraction of terms is done in domain-specific corpora with the help of general corpora. Using TF-IDF and stop-word lists, we restrict terms that are sub-terms of other terms in the list of candidate terms in order to reduce the number of candidate terms.

Our algorithm emphasizes in identification of verb or noun phrases multi-word terms. For verb phrases, the method excludes auxiliary and light verbs and takes their base forms as candidates. For noun phrases, the method extracts structures of noun-noun (NN), adjective-nouns (AN) and

its combinations, including conjunctions (C) or determiners (D). It also recognizes unknown words (e.g., words of the domain-specific corpus which are not found in the general corpus) because most technical jargon is unlikely to be included in a general dictionary. To do so, we split the task into three steps: (i) corpus creation, (ii) term identification and, (iii) automatic validation.

2.1. Corpus creation

To be able to export multi-word terms, i.e., a group of words that are commonly used together, we created a corpus from scientific publications and websites in the domain of vibration-based CM. We call it the Analysis Corpus (AC). The collection consists of full texts of 209 Journal papers, including references, on CM using vibration technology. Additionally, we crawled data from approximately 30 commercial websites that offer various types of customers vibration-based CM for several applications. To verify our results, we then created a new corpus, called reference corpus (RC). This is a general-purpose corpus and it contains data of the Europarl corpus [10], the United Nations Parallel Corpus [11] and the ParaCrawl [12]. It has a 10-times longer length than the AC. Both corpora were pre-processed with Natural Language Toolkit for Python [13] in order to separate the punctuation and the words. Both corpora are tokenized at the sentence level. Furthermore, the initial letters of each sentence are true-cased in accordance with the most frequent case. Besides this, the RC was factored, meaning that each word was tagged with its part-of-speech (POS) and its lemma.

The reason we use two different types of corpora (in terms of content), is that it will help us to compare the behavior of multi-word units in different contents and extract the ones that are specific to the AC. Table 1 summarizes the size of each corpus. The size of the RC ensures that it covers a wide range of topics and that its content is heterogenous. The AC is, in contrast, domain-specific and topic-oriented.

Table 1. Statistics about the corpora

	AC	RC
Sentences	122,628	1,222,000
Tokens	975,878	11,243,504
Word forms	87,785	163,266

As can be seen from Table 1, the corpora are quite small. The size of the AC was determined by the original intention of our research. We decided to use short, concise texts that are representative of those mined manually by researchers in the domain. During the term extraction process by TF-IDF, each line of our corpus is taken into account as a document.

2.2. Term identification

The next step after creating our corpora was to extract the candidate terms. To achieve this, we apply the TF-IDF algorithm to the AC. The TF-IDF measures a word's importance in a text. It is the product of two statistics, term frequency (TF) and inverse document frequency (IDF). The TF for a term is defined as follows:

$$TF_{ij} = \frac{n_{ij}}{|d_j|}, \tag{1}$$

where n_{ij} to be the frequency of term i in document j, while $|d_j|$ corresponds to the total number of terms in the text. The IDF for a term is defined as follows:

$$IDF_j = \log\frac{n}{n_j}, \tag{2}$$

where n corresponds to the total number of texts in the collection and n_j corresponds to the number of texts containing the term. Finally, the TF-IDF value of a term is calculated as follows:

$$(TF - IDF)_{ij} = TF_{ij} * IDF_j. \tag{3}$$

We set the maximum number of words for a candidate term at seven (7-grams) i.e., only strings with a frequency of occurrence of seven or more words are extracted, including symbols such as hyphens and dashes. To do so, we calculate the TF-IDF algorithm as follows:

$$(TF - IDF)_{ij\ n-gram} = TF_{ij\ n-gram} * IDF_{j\ n-gram}. \tag{4}$$

The idea of TF-IDF algorithm is that if a term appears a lot in a few documents, it is a specialized term of the domain. While if a term appears a lot in all documents, it will be less relevant. For instance, the term candidate, case study will have an extremely low TF-IDF. While the term frequency response will have high frequency in few documents and hence will be ranked on top.

After we extracted the terms, we removed the noise using a stop-word list. In stop-words we mean the most frequently used words in a language. Candidates that start or end with a word in the list of stop-words were rejected. We used the default stop-word list provided by the NLTK suite since there is no universal list of stop-words. It contains 179 commonly used words such as *while*, *why*, *further*, *must*, *doing*, etc. Additionally, we filtered out candidates starting and ending with numbers.

2.3. Automatic validation

Our technique relies on statistics observed in the domain-specific corpora, thus results may contain noise, i.e., general terms or vocabulary. We validated the terms against the RC in order to get an even more precise and accurate selection. We removed the subordinate terms or nested terms during this step.

Consider the sequence of words *forced response analysis*. This is a term in vibration-based CM. A methodology based on frequency of occurrence would extract it, given its high frequency in the corpus. Its substrings, *forced response* and *response analysis*, would also be extracted since they would have frequencies at least as high as *forced response analysis*. However, *response analysis* is not a term in vibration-based CM.

An easy solution to this problem is to filter out candidate substrings of a candidate term considering its frequency on its own in the corpus (i.e., within different surrounding words). Then, in order to decide if a candidate will be included in the list, we should subtract the frequency of its candidate terms as a substring from its total frequency [14]:

$$termhood(a) = f(a) \sum_{b \in Ta} f(b), \tag{5}$$

where a is the candidate term, $f(a)$ is its total frequency in the corpora, T_a is the group of candidate terms that contain a, b is such a candidate term and, lastly, $f(b)$ is the frequency of the candidate term b that contains a.

Moreover, we used linguistic knowledge. Namely, we use morphological analysis (lemma) to match the inflected forms of the terms and POS tagging in this step. One of the features of our architecture is that it exports only candidates consisting of patterns of [N N], [A N], [A A], and [V] or [V N] or combinations as described in John Justeson and Slava Katz [15]. It also looks for the pattern [N "C" N] and [A "C" N].

Finally, the algorithm exports all unknown words that are identified in the AC but do not appear in the RC in a separate list. These words are normally neologisms or they might contain

typos. To furthermore distinguish the words with typos from neologisms, we use string matching algorithms to get the most approximate pattern. To do so, we find the candidate P term which has the smallest edit distance to the unknown pattern T.

3. Results

The list of the candidate terms consists of 20,000 English entries obtaining a total of 9,000 term candidates (i.e., 45 % of the total) after filtering and the automatic validation. After filtering out the unigrams which are out of scope of this research, 4,000 term candidates remained (i.e., 20 % of the total). Table 2 shows the number of n-grams for each step of the automatic term extraction.

Table 2. Number of n-grams extracted from each step of automatic term extraction

	Term identification step	Automatic validation step	Multi-word candidate terms
Unigram	7,106	1,672	–
bigram	7,476	4,207	2,340
3-gram	3,854	2,239	1,178
4-gram	1,167	647	337
5-gram	258	157	99
6-gram	108	59	36
7-gram	31	19	10
Total	20,000	9,000	4,000

Some sample fragments of the extracted terms with a high, medium and low frequency of occurrence in the corpus are shown in Table 3. There are no units, acronyms, or abbreviations such as FRF, FTF, FFT, etc. in the following list.

Table 3. The 20 terms with the highest frequency of occurrence

No	Frequency	Term	No	Frequency	Term
1	165	Acoustic emission	11	398	frequency domain
2	112	Anti-aliasing filter	12	158	frequency domain analysis
3	323	Natural frequency	13	112	dynamic stiffness
4	83	Band pass filter	14	150	frequency spectrum
5	73	Bearing frequencies	15	353	fundamental frequency
6	93	Bearing misalignment	16	196	imaginary part
7	48	Coherence function	17	101	impulse response
8	72	Frequency response	18	325	low pass filter
9	66	Discrete fourier transform	19	435	mechanical impedance
10	45	Automatic machine CM using vibration signals	20	52	conventional vibration spectrum analysis

Table 4 shows some examples of candidate terms, grouped by their structure. The lexical items on this list met the probability threshold used in the identification process. The list needs to be further cleaned by hand. For instance, the candidate term electrotechnical standardization with a high frequency would most likely be removed from the list if it is not a representative term of the domain of vibration analysis. This is considered normal given that we have a corpus of scientific papers with standard phraseology used by authors. Consequently, all candidate terms and in particular those of high frequency, such as electrostatic effect and absolute maximum, have to be further validated by experts in the vibration-based CM domain to be representative of their domain.

Validation is done by taking into account: (i) if the term is representative of the domain, (ii) if the term is neutral and free of judgements, (iii) if the term has been deprecated, and, (iv) if the term is not ambiguous. Metadata are attached to each term; namely, a list of contexts, the POS, abbreviations, the frequency, etc. (see Fig. 1). The experts could either approve or reject the candidates based on their experience in the area. In addition, they assign definitions to the terms

and add more metadata, if necessary.

Table 4. Term candidates examples

Structure	Term candidate
NN	frequency domain
AN	acoustic emission
ANN	short-term disturbance noise
NNN	frequency domain analysis
AAN	white gaussian noise
ANNN	conventional vibration spectrum analysis
ANNNVNN	automatic machine CM using vibration signals

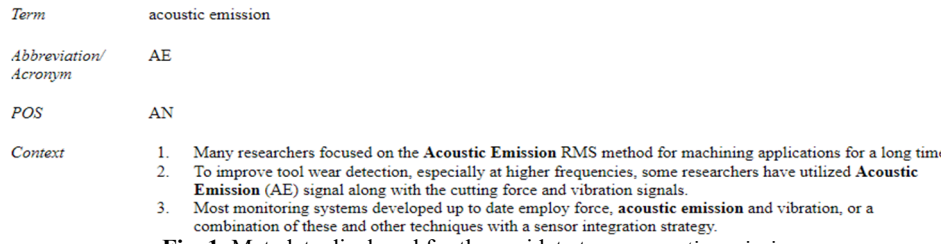

Fig. 1. Metadata displayed for the cacidate term *acoustic emission*

For this reason, 2 hours of training was offered for the experts to introduce them to the terminography. It should be noted that we acknowledge and recognize the valuable terminology work made by the International Organization for Standardization with the help of the International Electrotechnical Commission (IEC) [16] as well as by the Vibration Institute (https://www.vi-institute.org/vibration-terminology-project). We use definitions from both contributions in our work and expand the metadata for the terms.

Based on their relationship to concepts, terms are classified according to the principles of terminography; (i) synonyms are grouped together in a single entry and (ii) homonyms and polysemes (i.e., words with the same pronunciation and/or spelling) are presented separately (different entries) since they represent different concepts. Furthermore, the term formation shall be concise and as neutral as possible, avoiding connotations, especially negative ones [17].

To prepare for and familiarize themselves with the documentation process, experts started with the validation of 20 ambiguous but typical terms in the domain of vibration-based CM [18]. These terms have broad meanings, and are employed in a variety of disciplines including music, physics, acoustics, electronic power transmission, radio technology, and other domains. For example, among others, the term *node* (http://www.electropedia.org/iev/iev.nsf /6d6bdd8667c378f7c12581fa003d80e7?OpenForm&Seq=2), according to IEC's electropedia is used across the domains of acoustics and electroacoustics, circuit theory, mathematics, telecommunication networks, teletraffic and operation, electric traction. In the same way, the term *filter* (http://www.electropedia.org/iev/iev.nsf/SearchView?SearchView&Query=field +SearchFields+contains+filter+and+field+Language=en&SearchOrder=4&SearchMax=0) is used across the domains of radiology and radiological physics, oscillations, signals and related devices, electrical and magnetic devices, lighting, etc. It is important to note that a term from one domain within the same language can be borrowed and attributed to another concept within another domain. Thus, a term may have a conflicting meaning with its general language meaning, or it may confuse the multiplicity of technical meanings.

After a term is verified, it is converted into machine-readable XML (According to [19]) format. The term entry *harmonic* is shown in Fig. 2. More specifically, in the subject domain of vibration CM, the English term *harmonic* is formed as a single-word term, via terminologization of the ordinary term *harmonic*, meaning "repeating signals, such as sinusoidal waves" in order to render the concept "harmonic vibration, the frequency of which is an integral multiple of the fundamental

frequency". It is a single, masculine noun, and its deprecated term is overtone because "the term "overtone" has frequently been used in place of harmonic, the nth harmonic being called the $(n-1)$th overtone". An example of how to use this term in the domain of vibration is "Some of these harmonics have a dominant value in the vibration spectrum due to interaction of machine flu harmonics and the mechanical structure of the machine".

```
<termEntry id="1">
  <descrip type="subjectField">vibration</descrip>
  <langSec xml:lang="en">
    <descripGrp>
      <descrip type="definition">harmonic vibration, the frequency of which is an integral multiple of the fundamental frequency</descrip>
      <descrip type="context">Some of these harmonics have a dominant value in the vibration spectrum due to interaction of machine flu
      harmonics and the mechanical structure of the machine.</descrip>
    </descripGrp>
    <ntig>
      <termGrp>
        <term>harmonic</term>
        <termNote type="partOfSpeech">noun</termNote>
        <termNote type="grammaticalNumber">singlural</termNote>
        <termNote type="grammaticalGender">masculine</termNote>
        <termNote type="administrativeStatus">preferred</termNote>
        <termNote type="usageNote">The term "overtone" has frequently been used in place of harmonic, the nth harmonic being called the
        (n - 1)th overtone.</termNote>
        <termNote type="deprecated">overtone</termNote>
      </termGrp>
    </ntig>
    <ntig>
      <termGrp>
        <term>overtone</term>
        <termNote type="partOfSpeech">noun</termNote>
        <termNote type="grammaticalNumber">singlural</termNote>
        <termNote type="grammaticalGender">masculine</termNote>
        <termNote type="administrativeStatus">deprecated</termNote>
      </termGrp>
    </ntig>
  </langSet>
</termEntry>
```

Fig. 2. Metadata displayed in XML format for the candidate term *harmonic*

As it is mentioned above, experts with strong knowledge of the domain of the vibration-based CM validate manually the term candidates. At the same time, we perform an error analysis so as to better understand the behavior of our method. For that purpose, we took 5 % random examples (400 candidate terms) from each n-gram category and manually analyzed each of the errors made by the algorithms. Our analysis first reveals that the rejected terms make up 4.5 % of the sample.

As for the actual errors, we observe that nearly a third of them correspond to out-of-domain. For instance, terms like *network error*, *limited α-frequency range*, *angle asymmetries*, etc. are not relevant to vibration-based CM domain. We also observe a few errors that are related to nested terms. In our work we filter out candidate substrings of a candidate term taking into consideration its frequency by itself in the corpus (i.e., within different surrounding words) [14]. However, not all cases are covered. For instance, in the term candidate list there are examples like *wind turbine vibration* and *wind turbine vibration signal*/*wind turbine vibration data*, *transient elastic wave* and *transient elastic wave generation*. For the remaining errors, terms that are syntactically interrupted by [A] are rejected by the experts. For instance, *angular frequency*, *angular samplic frequency* and *angular supply frequency*.

All in all, our error analysis reveals that our method performs well and based on the feedback the algorithm will be able to filter out the extracted list. Moreover, it shows that the quality of the multi-word term candidates is much better than what the statistical approaches encouraging the extension of the algorithm in other applications or the integration of other techniques.

4. Conclusions

In this work, we propose a method for domain-specific terminology extraction of groups of words that commonly occur together. We use an advanced filtering strategy to exclude out-of-domain term candidates and/or nested terms. Thy hybrid methodology we designed, benefits from both statistical and linguistic approaches in order to provide a more accurate and precise term selection. In addition, we use two different types of corpora, (domain-specific vs. general-purpose corpus) to remove the noise. Our algorithm emphasizes in identification of verb or noun phrases multi-word terms. For verb phrases, it excludes auxiliary and light verbs and takes their base forms as candidates. For noun phrases, it extracts structures of [N N], [A N] [A A], and

[V] or [V N] or combinations, including [C] or [D]. In addition, it takes into consideration the identification of neologisms and technical jargons.

Our study based on a vibration-based CM domain corpus shows that our method can export high quality domain-specific terminology from a small corpus. Through the error analysis results, we can see that most extracted terms have nice correspondence to the domain of CM concepts and notions. Hence, our method reduces the human effort as well as helps in a wide-range and representative term selection for a given domain.

References

[1] **Pazienza M. T., Pennacchiotti M., Zanzotto F. M.** Terminology extraction: an analysis of linguistic and statistical approaches. Knowledge Mining, Springer, Berlin, Heidelberg, 2005, p. 255-279.

[2] **Sager J.** Terminology: Theory. BAKER, M. (ed.). Routledge Encyclopedia of Translation Studies. London/New York: Routledge, 1998, p. 258-262.

[3] **Jacquemin C.** Variation terminologique: Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. Mémoire d'Habilitation à Diriger des Recherches en informatique fondamentale, Universitè de Nantes, France, 1997, (in French).

[4] **Pazienza M. T.** A domain specific terminology extraction system. International Journal of Terminology, Vol. 5, Issue 2, 1999, p. 183-201.

[5] **Salton G.** Automatic Text Processing: the Transformation, Analysis and Retrieval of Information by Computer. Addison Wesley, 1989.

[6] **Church K., Hanks P.** Word association norms, mutual information, and lexicography. Computational Linguistics, Vol. 16, Issue 1, 1990, p. 22-29.

[7] **Dunning T.** Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, Vol. 19, Issue 1, 1993, p. 61-74.

[8] **Kilgarriff A.** Comparing corpora. International Journal of Corpus Linguistics, Vol. 6, Issue 1, 2003, p. 97-133.

[9] **Drouin P.** Term extraction using non-technical corpora as a point of leverage. Terminology, Vol. 9, Issue 1, 2003, p. 99-115.

[10] **Koehn P.** Europarl: A parallel corpus for statistical machine translation. MT summit, Vol. 5, 2005, p. 79-86.

[11] **Ziemski M., Junczys Dowmunt M., Pouliquen B.** The united nations parallel corpus. Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 2016.

[12] **Banón M., Chen P., Haddow B., Heafield K., Hoang H., Espla Gomis M., Forcada M., Kamran A., Kirefu F., Koehn F., Ortiz Rojas S., Pla Sempere L., Ramírez Sánchez G., Sarrías, E., Strelec M., Thompson B., Waites W., Wiggins D., Zaragoza J.** ParaCrawl: Web-scale acquisition of parallel corpora. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, p. 4555-4567.

[13] **Loper E., Bird S.** NLTK: the natural language toolkit. Proceedings of the ACL Interactive Poster and Demonstration Sessions, 2004, p. 214-217.

[14] **Frantzi K., Ananiadou S., Mima H.** Automatic recognition of multi-word terms: the C-value/NC-value method. International Journal on Digital Libraries, Vol. 3, 2000, p. 115-130.

[15] **Justeson J. S., Katz S. M.** Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering, Vol. 1, Issue 1, 2005, p. 9-27.

[16] Mechanical vibration, shock and condition monitoring - Vocabulary (ISO/DIS Standard No. 2041). International Organization for Standardization, 2018, https://www.iso.org/obp/ui/#iso:std:iso:2041:ed-4:v1:en.

[17] Terminology work — Principles and methods (ISO/DIS Standard No. 704). International Organization for Standardization, 2009, https://www.iso.org/standard/38109.html.

[18] **Chatzitheodorou K., Kappatos V.** Terminology study on vibration-based condition monitoring technique. Vibroengineering Procedia, Vol. 34, 2020, p. 20-26.

[19] Management of Terminology Resources – TermBase eXchange (TBX) (ISO/DIS Standard No. 30042). International Organization for Standardization, 2019, https://www.iso.org/standard/62510.html.

Konstantinos Chatzitheodorou received his Ph.D. in applied translation studies and computational linguistics from the Aristotle University of Thessaloniki. He holds a B.A. in Italian language and literature from the School of Italian Language and Literature, Aristotle University of Thessaloniki and an M.Sc. in informatics in humanities from the Department of Informatics, Ionian University. He is also an ECQA Certified Terminology Manager – Engineering. He is currently working in the private sector as computational linguist assisting organizations to use language data to gain strategic insights.

Vassilios Kappatos obtained Ph.D. degree in the area of non destructive evaluation (NDE) and also holds the Diploma of mechanical and aeronautical engineering (MEng) since 2002. Now, he is the Head of Constructural and Infrastructure Research in Maritime and Air Transport Laboratory at Hellenic Institute of Transport (HIT), Center for Research and Technology Hellas (CERTH), Greece. His research areas are NDE, structural health monitoring, condition monitoring, structural integrity, pattern recognition and signal processing. His research has been supported by the European Commission and other organizations.