Terminology study on vibration-based condition monitoring technique

Konstantinos Chatzitheodorou[1], Vassilios Kappatos[2]
[1]School of Italian Language and Literature, Aristotle University of Thessaloniki, Thessaloniki, GR541 24, Greece
[2]Hellenic Institute of Transport, Centre for Research and Technology Hellas, 6th Km Charilaou Thermi, 60361, Thermi, Thessaloniki, Greece
[2]Corresponding author
E-mail: [1]*chatzik@itl.auth.gr*, [2]*vkappatos@certh.gr*

Received 21 October 2020; accepted 27 October 2020
DOI https://doi.org/10.21595/vp.2020.21758

Abstract. In this paper, we present the initial results of a terminology study on a vibration-based condition monitoring technique to support the research community. We automatically export terminologies from domain specific textual corpora and a team of subject matter experts validates and revises 20 typical terms. Our goal is twofold; (i) to standardize terminology for the domain of vibration since different groups use different definitions of commonly used terms and (ii) to distinguish the specific terminology from the general terminology or from the terminology of other mechanical areas. We intended to serve the needs of a wide range of user groups such as subject matter experts who sometimes need to ascertain the meaning of an unknown term or concept. To date, there is no similar work or initiative that could be integrated, reused or extended.

Keywords: terminology, condition monitoring, vibration, specialized languages, term-base, technical language.

1. Introduction

Technical terms are an essential part of both technical and scientific authoring. Engineers and experts use a vocabulary that relays a variety of specialized concepts by means of technical language. This special vocabulary conveys concentrated meaning that has been built up over the years in a specific domain. The value of terminologies – the lexical components of specialized languages - lies in the way each concept condenses a mass of information into a single-word or a multi-word expression. In our research, we introduce a terminology study on the application of a vibration-based condition monitoring (CM) technique, which is a keytool in the predictive maintenance of any equipment and machines within a wide range of industrial applications such as transport, oil and gas, processing and manufacturing.

CM provides a myriad of benefits, including 24×7 remote monitoring, early warnings of potentials (serious) failures, reduction in unplanned downtime, reduction in maintenance costs, support ongoing reliability and risk reduction with fewer inspections. CM uses measurement of specific equipment parameters, such as vibrations in a machine, its temperature or condition of its oil, taking note of any significant changes that could be indicative of an impending failure. Vibration-based CM refers to the use of in site non-destructive sensing and analysis of system characteristics in the time, frequency or modal domains for the purpose of detecting change, which may indicate damage or degradation.

The area of vibration analysis is filled with different technical terms, jargon, acronyms and concepts. General words that describe some general or vague group or class such as frequency, order, node, etc. lead to semantic confusion in the area of vibration. In fact, vocabularies applied to specific conditions are used interchangeably to denote general dysfunction and vice versa. This confusion may reflect research, which has radically changed over the last decades. To lessen this confusion we aim to provide firm and unequivocal use of terminology of vibration analysis.

While terminology identification, i.e. creation of a list with domain specific vocabulary for a textual corpus, has become widespread nowadays, the notion itself of term is still not clear, both from a pure linguistic and a computational point of view [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".

Quite a bit of work has been done in order to identify terms for a specific domain. Most of them use statistical methods to export terminologies from textual corpora, i.e. collection of documents in the domain. Many of the results are entirely usable. Gerard Salton [5] suggested term frequency-inverse document frequency (TF-IDF). It is a statistical measure that evaluates how relevant a term is to a document in a collection of documents. This is done by taking into consideration: (i) number of times a term appears in a document, and (ii) the inverse document frequency of the term across a set of documents. Later, Kenneth Church and Patrick Hanks [6] proposed the mutual information (MI) which is a concept rooted in information theory. In particular, it measures how much information is communicated, on average, in one random variable about another. Both variables are sampled simultaneously. Ted Dunning [7] proposed a measure, which is based on likelihood ratios and relies on frequency profiling. The method delivers reasonably results and works in both large and small text samples. Recently, researchers have used corpora comparison to capture n-grams (i.e., occurrences of one or more words) 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. In general, though, most methods lead to endless lists of candidate terms.

In this work, we present the initial results of a terminology work which is still in progress. We created a termbase, namely a collection of electronic term records from the vibration-based CM area containing 20 typical terms. Each entry, which is automatically extracted by domain specific textual corpora is validated by a team of subject matter.

2. Methodology

In this research, we extend the methodology proposed by Patrick Drouin [9], where term extraction is performed in domain-specific corpora leveraging information from general corpora. We are applying TF-IDF as well as syntactic information and stop-word lists to reduce the noise in the list of candidate terms by restricting terms that are sub-terms of other terms.

Our algorithm identifies both single-word and multi-word terms and it focuses on extraction on either verb or noun phrases. For verbs, it excludes auxiliary and light verbs and takes their base forms as candidates. For noun phrases, it extracts structures of noun-noun (NN), adjective-nouns (AN) and its combinations, including conjunctions (C) or determiners (D). It also pays attention to unknown words (e.g., words of the domain-specific corpus which are not found in the general corpus) because most technical jargon is not likely to be included in a non-specific dictionary.

2.1. Corpus creation

In order to be able to export terminologies, we first created a textual corpus from scientific publications and websites in the domain of vibration-based CM. We called this analysis corpus (AC). It consists of full texts of 209 Journal papers, excluding references, in the area of CM using vibration technology. Moreover, we crawled data from approx. 30 commercial websites, which offer equipment, software tools and service for vibration-based CM to all kinds of customers for several applications. To validate our results, we then created another corpus, which is called reference corpus (RC). This is a general purpose corpus and it consists of data of the Europarl corpus [10], United Nations Parallel Corpus [11] and ParaCrawl [12]. It's size is 10-times bigger in length than the AC. Both corpora were preprocessed with the Natural Language Toolkit for

python [13] to normalize punctuation, and distinguish it from the words. It is also tokenized on sentence level. Furthermore, the initial letters of each sentence were truecased. In addition, the RC was factored, namely, each word was tagged with its part-of-speech (POS) and its lemma. In more detail, the AC corpus consists of 975,878 tokens, which correspond to roughly 87,785 word forms while the RC consists of 11,243,504 tokens, which correspond to roughly 163,266 word forms. The size of the RC can guarantee that it covers a wide range of subjects and that its content is heterogeneous. In contrast, as previously mentioned, the AC is domain-specific and topic-oriented. Table 1 gives detailed information on the size of each corpus.

As it emerges from Table 1 the corpora are rather small. The size of the AC was determined by the original intent of our research. Since it is a work in progress we decided to use small texts that are representative of the ones mined manually by researchers in the domain. In that regard, each line of our corpus is taken into account as a document during the term extraction by TF-IDF.

Table 1. Statistics of the corpora

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

2.2. Term identification

After the creation of our corpora, the next step was to extract the candidate terms. To do so, our pipeline architecture applies the TF-IDF algorithm into the AC. We set the maximum number of words for each candidate term to seven (7-grams), including hyphens and dashes. After the term extraction is done, we filtered out a few terms using the stop-word list. Though stop-words we refer to the most common words in a language. Since there is no single universal list of stop-words, we used the default stop-word list of NLTK suite. It contains 179 common words such as while, why, further, during, doing, etc.

2.3. Automatic validation

Our technique relies on a frequency calculation observed in domain-specific corpora, hence results expect to contain noise, i.e., general terms. To remove this noise, we validated the candidate terms against the RC. During this step, we also removed the subterms as well as the inflected forms of the terms. Moreover, since the term identification is still an open problem, we used morphological analysis (lemma) and POS tagging in this step. Our architecture selects to export only candidates consisting of NN, AN, C and V as it is described in John Justeson and Slava Katz [14]. Last, the algorithm exports in a separate list all unknown words that are identified as candidate terms in the AC but they don't appear in the RC.

3. Results

The list with the candidate terms consists of approx. 20,000 English entries, after the automatic validation. Table 2 shows the top 40 terms sorted by frequency of their occurrence. The following list does not include units, acronyms, abbreviations and multi-word terms such as vibration signal, vibration analysis, etc. Some examples of candidate terms among their structure are shown in Table 3.

This list contains lexical items that reached the probability threshold the algorithm used during the identification process. Further manual cleansing is needed to reduce the size of the list. For instance, the candidate term study which occurs 360 times will probably be removed from the list if it is not a representative term of the domain of vibration analysis. This is normal taking into consideration that our corpus consists of scientific papers where standard phraseology is used by authors. Hence, all candidate terms with high number of frequency such as example, addition, etc. need further validation for their representativeness in the domain by experts in the vibration-based

CM domain.

Table 2. The top 40 terms sorted by frequency of their occurrence

No	Times	Term	No	Times	Term	No	Times	Term
1	967	Vibration	15	398	Amplitude	29	267	Tool
2	702	Signal	16	397	Analysis	30	265	Process
3	669	Results	17	382	Machine	31	259	Research
4	666	Number	18	380	Faults	32	258	Technology
5	660	Time	19	360	Study	33	247	Example
6	659	Data	20	350	System	34	247	Parameters
7	586	Paper	21	337	Bearing	35	232	Bearings
8	546	Frequency	22	337	Use	36	227	Addition
9	531	Order	23	328	Fault	37	227	Test
10	460	Condition	24	315	Gearbox	38	225	Motor
11	456	Case	25	288	Section	39	224	Information
12	438	Features	26	280	Result	40	223	Conditions
13	428	Method	27	271	Speed			
14	410	Signals	28	269	Value			

Table 3. Examples of candidate terms

Structure	Term
N	Frequency
NN	Frequency spectrum
AN	Rotational frequency
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

The validation is done taking into consideration: (i) the representativeness of the term in 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 unambiguous. Each term is accompanied by metadata; namely, list of contexts, POS, abbreviations, frequency, etc. (see Fig. 1). The experts could either Approve or Reject the candidates based on their knowledge in the area. In addition, they revise the terms by assigning definitions and more metadata or notes, if any.

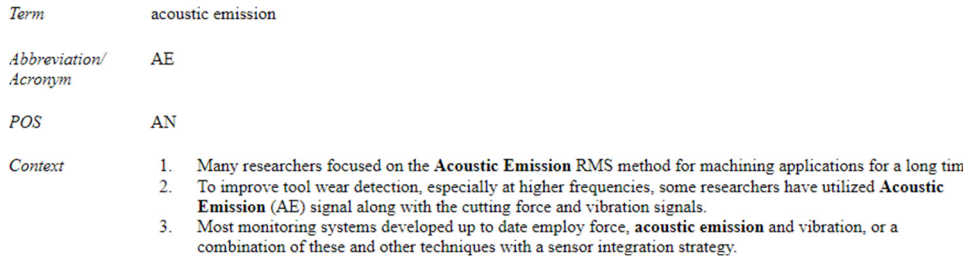

Fig. 1. Metadata displayed for the cacidate term acoustic emission

In this research, we present our findings from the preliminary results. In more detail, we present 20 ambiguous but typical terms that may lead to semantic confusion in the domain of vibration-based CM. These terms are employed in various disciplines including music, physics, acoustics, electronic power transmission, radio technology, and other domains, as well as they have a non-specific meaning. For instance, among others, the term "node" (http://www.electropedia.org/iev/iev.nsf/6d6bdd8667c378f7c12581fa003d80e7?OpenForm&Seq=2), according to IEC's electropedia is used in the domains of acoustics and electroacoustics,

circuit theory, mathematics, telecommunication networks, teletraffic and operation, electric traction. Similarly, 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 in the domains of radiology and radiological physics, oscillations, signals and related devices, electrical and magnetic devices, lighting, etc. It is important to point out that a term from one domain may be borrowed and attributed to a new concept in another domain within the same language. Hence, the meaning of a term may conflict with the general language meaning, or it may confuse the multiplicity of technical meanings.

Under the light of the above, a team of experts in the domain of vibration-based CM revised 20 ambiguous terms (Accessible at: rb.gy/ieb7nh). It is part of the warm-up process of the term validation task given that they don't have previous experience in terminography. For that reason, 2 hours of training was offered aimed to introduce the experts in the terminography. It should be noted that we recognize and acknowledge the valuable terminology work made by the International Organization for Standardization in conjunction with the International Electrotechnical Commission (IEC) [15] as well as by the Vibration Institute (https://www.vi-institute.org/vibration-terminology-project). In our work, we borrow definitions from both contributions and we expand metadata for the terms.

Terms are classified according to their relation to the concepts taking into consideration the principles of terminography; (i) synonyms are grouped together (in the same entry) and (ii) polysemes (i.e., words with the same pronunciation and/or spelling) and homonyms (i.e., (homographs and/or homophones) are presented separately (different entries) because they represent different concepts. Moreover, the term formation shall be concise and as neutral as possible, avoiding connotations, especially negative ones [16].

The terms presented in this work are based on the English language and are not intended to cover all the methods used for English term formation. Fig. 2 shows the term entry "harmonic" in XML format (According to [17]). In more detail, in the subject domain of vibration, the English term "harmonic" was formed as a simple 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 singular, 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 usage of 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 for the cacidate term "harmonic"

4. Conclusions

We introduced a terminology study on vibration-based CM technique aiming to harmonize current definitions and terminologies. To do so, we automatically exported candidate terms from scientific research papers and as well as websites specialized in this domain. In this study we present the findings from the first results which are manually validated by a team of subject matter exerts. In more detail, 20 ambiguous but typical terms that may lead to semantic confusion in the domain of vibration-based CM are handled and documented. Terminologies are compiled in XLM format so that can be read by machines. This choice of the terms provided for the possibility to expand in the near future the temrbase rendering successfully multi-word terms that consist of them. For instance, by documenting the term "frequency", the terms "frequency spectrum", "rotational frequency", "double frequency", will be documented with less effort. The findings of his work will not only help scientists in the domain of vibration-based CM, but also teachers, professors and students who study this domain, as well as professional communication mediators, such as journalists, technical authors and translators [18]. It is also worth noting that this work may help language engineers to solve natural language processing (NLP) tasks such as domain adaptation of machine translation, text classification, search engine optimization, language understanding, etc.

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. Routledge Encyclopedia of Translation Studies. Routledge, London/New York, 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. 232-263.

[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., et al.** 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] **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.

[15] 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.

[16] Terminology Work – Principles and Methods. ISO/DIS Standard No. 704, International Organization for Standardization. 2009, https://www.iso.org/standard/38109.html.

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

[18] **Sager J. C.** Practical Course in Terminology Processing. John Benjamins Publishing, 1990.