Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 95 (2013) 267 – 275

Procedia
Social and Behavioral Sciences

5th International Conference on Corpus Linguistics (CILC2013)

Classifying Different Definitional Styles for Different Users

Luis Espinosa Anke*

Universitat Autònoma de Barcelona, Campus de Bellaterra, Cerdanyola del Vallès, 08193, Spain

Abstract

This paper proposes an approach to classify definitions as they appear in popularizing texts. Following the function theory of lexicography (Bergenholtz and Tarp, 2003), we propose a user-centered classification that breaks down definitions according to the way they are deployed in the text (Westerhout and Monachesi, 2007) and the information they encode (Sierra et al., 2006). We hypothesize that different types of users can benefit from this classification, which covers a range of definitional styles, from the classic *genus et differentia* model to function-oriented definitions. The corpus used for this task consists of 50 transcripts from *The Science Magazine Podcast* (around 400k words), where 570 snippets containing definitional information have been manually annotated and classified.

Keywords: popularizing texts; scientific interviews; lay definitions; function theory of lexicography

1. Introduction

Today, new concepts or new terms are introduced and replace old ones as our perspective of reality changes (Fuertes-Olivera and Nielsen, 2011). This reality becomes evident in scientific contexts where the need to reformulate advancements or new discoveries demands an accurate formalization of knowledge. One of these contexts is specialized communication, often referred to as *Language for Specific Purposes* (Fuertes-Olivera and Arribas Baño, 2008). We focus on popularizing texts, i.e. those where experts and semi-experts communicate for semi-experts and a lay-audience (Diéguez, 2004), and more specifically the genre of scientific interviews, where scientists are interviewed in order to educate the public and engender interest in the scientist's own specialty (DiBella, 1991).

* Corresponding author. Tel.: +34 670585434
E-mail address: luis.espinosa83@gmail.com

doi:10.1016/j.sbspro.2013.10.647

On the other hand, the function theory of lexicography is the theoretical framework under which we express the justification for the appropriateness of different definitional styles for different users. This theory highlights, among other concepts, the notion of *user needs*. These needs are closely linked to the user as well as the social situation where lexicographic needs appear (Bergenholtz and Tarp, 2010). We hypothesize that this theoretical framework can be followed in order to approach the definitional knowledge that appears in scientific interviews, which are a bundle of specialized language and persuasive and rhetoric techniques devoted to seduce and attract the audience, complying with what is expected from texts with a popularizing nature (Suau, 2005). For example, consider the case of a non-expert seeking to understand better what *single nucleotide polymorphisms* are. The definition that he/she would find at Wikipedia would be: "A single-nucleotide polymorphism (SNP, pronounced snip; plural snips) is a DNA sequence variation occurring when a single nucleotide — A, T, C or G — in the genome (or other shared sequence) differs between members of a biological species or paired chromosomes in a human". A non-expert might be interested in a shorter and less dense definition, and a popularizing text would therefore prove useful. See, for example, the information provided about the same term in the *Science Magazine Podcast* (June 27th, 2008): "Well, Rob, what we are talking about are little tiny changes, in fact, just a single change in an entire stretch of DNA. And these little changes are what biologists call single nucleotide polymorphisms."

Assuming a potential of corpora of scientific interviews as repositories for different definitional styles for different users (who, in addition, have different lexicographic needs), we present a first attempt to classify explicit definitional knowledge included in these interviews. The corpus consists of 50 interview transcripts from *The Science Magazine Podcast*[1] fully annotated with linguistic, terminological and definitional information. Table 1 shows basic statistics, such as number of words, sentences, terms and definitions.

The remainder of the paper is structured as follows: Section (2) reviews previous work in definition classification and categorization; Section (3) discusses the steps for compiling and enriching the corpus; Section (4) comments on troublesome cases and the criteria for overcoming them; Section (5) presents and discusses the user-wise classification of definitions according to their degree of specialization and, finally, Section (6) comments on the conclusions drawn and points to potential directions for improvement in future work.

Table 1. Raw counts for our corpus

Unit type	**Count**
Words	3892931
Sentences	15315
Terms	26194
Definitions	570

2. Related Work: Identifying and classifying definitions

Definitions play today a crucial role in the information age. The need to structure the information available on the Internet is obvious as the amount of information out there increases every day. Understanding meaning of words can benefit from the existence of glossaries or ad-hoc dictionaries (Park et al., 2002). From the days of Plato and Aristotle, where the *genus et differentia* model was proposed, research in how to characterize a definition has led to different classifications. We introduce two taxonomies that have been followed during the annotation process. These are (1) a pattern-based classification (Westerhout and Monachesi, 2007) and (2) an information-based classification (Sierra et al., 2006). These two taxonomies seem to be non-overlapping, and appear to resort to different strategies for definitional knowledge building.

[1] www.sciencemag.org/site/multimedia/podcast/index.xhtml

2.1. Sierra et al. (2006): Information-based classification

According to Del Gaudio et al. (2013), this is an extension of the classic Aristotelian classification. This taxonomy considers the kind of information present in a definition for assigning it a class. These are: (1) Analytic or Intensional definitions, which comply with the classic genus et differentia model; (2) Synonymic definitions, where an equivalent term is indicated, (3) Functional definitions, where information about a term is provided by considering its usage or application in a given context, and (4) Extensional definitions, where a term is described by enumerating its components.

2.2. Westerhout and Monachesi (2007): Pattern-based classification

The proposed classification looks at the way definitions spawn in a document, and considers different patterns for their deployment. These are: (1) Is-definitions, definitions introduced by the verb “to be”, (2) Verb-definitions, which are introduced by any verb other than “to be”, (3) Punctuation-definitions, where punctuation marks (commas, colons or brackets) are used to connect term and definition, (4) Pronoun-definitions, where a term is not explicitly mentioned, and instead we have a pronoun. Some kind of anaphora resolution is, then, required to identify the entity to which the pronoun is referring to, (5) Layout-definitions, in which the structure of the document (tables, bulleted lists, font formatting, etc.) is used to identify definitions, and (6) Other-definitions, or unclassifiable definitions.

3. Corpus Compilation and Annotation

We start from the premise that a linguistic corpus is “a collection of texts, of the written or spoken word, which is stored and processed on computer for the purposes of linguistic research" (Renouf, 1987:1). Moreover, it (1) must have a finite size, (2) must be a representative sample of a larger population of texts, (3) must be in machine-readable form, (4) must be a standard reference for the language in question, and finally (5) can be annotated, i.e. it can be enriched with additional information, which usually ranges from part-of-speech or syntactic information to semantic or even discourse-oriented and pragmatic information (McEnery and Wilson, 2001). Figure 1 summarizes the steps followed to compile and annotate the corpus.

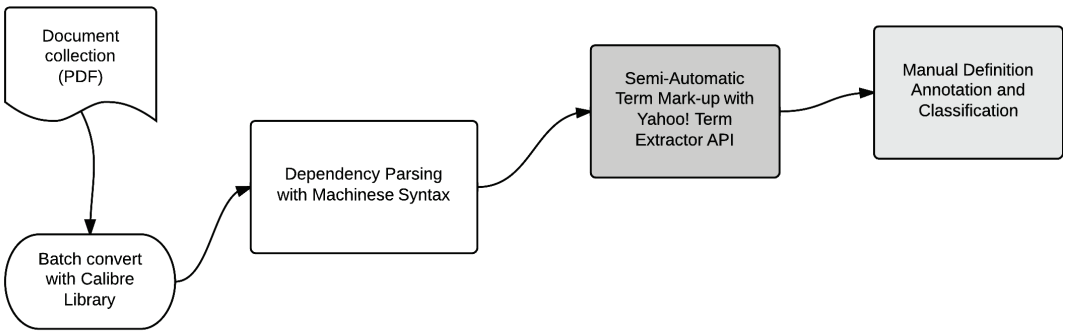

Fig. 1. Summary of the steps involved in the compilation and annotation of the corpus.

3.1. Corpus Pre-Processing

All documents are downloaded manually and converted to plain text files using the software Calibre[2]. Next, documents are parsed using the dependency parser Machinese Syntax (Tapanainen and Järvinen, 1997). This outputs xml documents with extensive linguistic information, such as sentence splitting, word lemmas, Part-of-Speech, Syntactic functions and dependency relation (if any).

3.2. Terminology identification

Once the documents were collected, converted, pre-processed and automatically parsed, the next step was to semi-automatically annotate the terminology. For this, we benefited from an API for Python of the *Yahoo! Term Extractor* (also known as *Yahoo! Content Analysis*[3]). Terms were identified, and *<Term></Term>* tags were inserted to the xml document. Since terms can span multiple words, the *<Term></Term>* tags were introduced as parent nodes of the <token> tags. Figure 2 shows the xml format layout for the term *stable climate change.*

```
<'?xml version= "1.0" encoding="ISO -8859- 1"?>
<Term ID=" ">
  <token id="w1335">
    <text>stable</text>
    <lemma>stable</lemma>
    <depend head="wl336">attr:</depend>
    <tags>
      <syntax>@A>N</syntax>
      <morpho>A ABS</morpho>
    </tags>
  </token>
  <token id="w1336">
    <text>climate</text>
    <lemma>climate</lemma>
    <depend head="w1337">attr:</depend>
    <tags>
      <syntax></syntax>
      <morpho>N NOM SG</morpho>
    </tags>
  </token>
  <token id="w1337">
      <text>change</text>
    <lemma>change</lemma>
    <depend head="w1332">comp:</depend>
    <tags>
      <syntax> @PCOMPL-S \%NH</syntax>
      <morpho>N NOM SG</morpho>
    </tags>
  </token>
</Term>
```

Fig. 2. Xml tree-like structure of a 3-word term.

[2] http://calibre-ebook.com
[3] http://developer.yahoo.com/contentanalysis

When queried, the Term Extractor API yields a list of terms, but its results depend on the size of the input text. This means that each document of the corpus had first to be split in sentences, and then each sentence was queried in order to preserve a high recall. It is worthwhile mentioning that the Yahoo! Term Extractor API has a limit of 5000 queries per IP per day[4].

3.3. Definition Annotation

We argue that in a textual genre like scientific interviews, where a certain degree of specificity and technical jargon is present, a classification that looks at the patterns of the definitions alone, or at their information alone, might prove insufficient to capture the complexity of the way information is presented. Furthermore, and considering the ultimate goal of this research, which is to provide a useful tool for linguistic research as well as for lexicographers or translators, we approach the definition identification as a task where any information that might be worthwhile to include in a glossary of expert, semi-expert or lay definitions (see Muresan and Klavans (2002) for further discussion on lay definitions). Each definition is tagged according to their pattern-based and information-based classification. Table 2 shows the 5 most frequent types of this two-dimensional classification, as well as their count and an example of each.

Table 2. Five most common types of definitions in the scientific interviews corpus.

Type of definition	**Frequency**	**Example**
Pattern type = is def Information type = intensional	135	Clicker's an electronic response device that's keyed to the instructor's computer, so the instructor is getting an answer and can grade it.
Pattern type = verb def Information type = functional	111	Mice develop regulatory T-cells against noninherited maternal alloantigens as a result of fetal exposure.
Pattern type = verb def Information type = extensional	52	Nano-ear is made from a microscopic particle of gold that is trapped by a laser beam.
Pattern type = is def Information type = functional	44	Iridium is not very common on Earth, but it is very common in asteroids.
Pattern type = punct def Information type = synonymic	32	(...) female determinant gene, S-ribonuclease gene.

An overview of the compilation and markup process has been provided. The next section will discuss those problematic cases that have been identified as most prominent. These were cases where discriminating between a definition or a non-definition chunk was difficult, or formal aspects such as cross-sentence definitions.

[4] http://developer.yahoo.com/search/rate.html

4. Problematic Cases

Consider the following passage: "*Well, Rob, what we are talking about are little tiny changes, in fact, just a single change in an entire stretch of DNA. And these little changes are what biologists call single nucleotide polymorphisms*" (*Science Magazine Podcast, June 27th 2008*).

If we were to manually mark up a definition from this passage we would encounter two main problems. First, it is a cross-sentence definition. This means that not all the elements that make up the definition spawn within a single-sentence boundary. In the first sentence, the definition is introduced ("a single change in an entire stretch of DNA"), while in the second sentence we find a pronominal noun-phrase ("these little changes") pointing back to the definition, and then finally the term ("single nucleotide polymorphisms"). Moreover, the term to be defined ("single nucleotide polymorphisms") is preceded by a relative clause ("what biologists call"). This raises the following question: Is this definition context-independent enough to be considered as a legitimate entry in a domain-specific glossary? Does the fact that the name is given by biologists makes any difference in the relevance of the definition? What if instead of "biologists", the passage was "these little changes are what *my father* calls single nucleotide polymorphisms"?. Issues of this nature have been identified as "Term Boundary", "Nested Definitions", "Anaphoric Definitions" and "Factuality" problems.

4.1. Term Boundary

Although technical terminology has no satisfactory definition (Justeson and Katz, 1995), it is agreed that a term should be precise and monosemic (unambiguous), emotionally neutral and stable over time (Gutiérrez Rodilla, 1998). However, one might encounter terms where it is difficult to decide where they begin and end. Taking an example from Justeson and Katz (1995), the term *central processing unit* could be seen as a 3-word term. If this term is extended as *AMD Sempron central processing unit*, the annotator would have to decide whether this is a 5-word term or a 3-word term with a pre-modifying noun phrase. The criterion has been to always keep the longest match.

4.2. Nested Definitions

Whenever two definitions overlap in some way, and due to the restrictions of *xml* mark-up, a well-formed document cannot have a definition opening before the previous one is closed. This might be seen more as a formal or engineering-related problem than a linguistic one, but nevertheless it has been highlighted due to its recurrence. Recall that this corpus reflects spoken language, where the number of pronouns and coreferential items is in general higher and more cryptic than in scientific or technical texts. The criterion followed to solve these cases was to always tag the first definition.

4.3. Anaphoric Definitions

Resolving anaphoras is out of the scope of this work. This means that in those cases where the components of a definition spawn in more than one sentence, these definitions are left out from the mark-up. For example, it is fairly common to have one term appearing in sentence number n, and then the definition in sentence number $n+1$, with a coreferential pronoun connecting both. However, in cases where term and definition appear in the same sentence, definitions are annotated, as in the following case: "*Dopamine is tied to the reward pathway, in our brain, and when our brain releases dopamine, it can reduce, say, the symptoms of chronic pain and depression*" (*Science Magazine Podcast, December 5th 2008*).

4.4. Non-Factual Definitions

Due to the informal register characteristic of these scientific interviews, where there is a degree of popularization of scientific or technological aspects, many apparent definitions are not such. Let us clarify the notion of popularization. According to Alcíbar (2004), experts in a domain, when attempting to disseminate their findings,

should consider popularization as a dramatization of science instead of a mere translation from technical jargon to a vocabulary accessible to laymen. This leads to a fuzzier boundary between what can be considered a definition, and what not. Although we propose a fairly relaxed threshold, there are cases where the context dramatically influences the meaning of a definition. An example has been proposed at the beginning of this chapter, where "*what biologists call"* triggers a *Factuality* problem. Therefore, annotation of definitions is only carried out if these are fully self-sustained, meaningful, and context-independent, whether this context is in the text itself, or in the form of world-knowledge.

5. Specialization-wise Classification

A lexicographical function is defined as the satisfaction of the specific types of lexicographically relevant needs that may arise in a specific type of potential user in a specific type of extra lexicographical situation (Tarp, 2008). We acknowledge different criteria elicited within the function theory of lexicography in terms of user needs, such as their mother tongue or their knowledge experience in translation. However, we focus on criteria that seem to be more relevant for deciding whether a more or less specific definition is chosen, such as a user's general cultural knowledge or his/her knowledge about a specific subject or science. Following the classification by Bergenholtz and Tarp (1995:19), we propose to aim definitions to three types of recipients, as we assume the producer of the definition to be in all cases an expert:

- **Experts**: For example, someone highly acquainted with the scientific language or the discipline to which the definition belongs
- **Semi-experts**: For example, a student of a scientific discipline who has working knowledge of a limited number of topics.
- **Laymen**: Someone who does not necessarily need to have any special knowledge or interest in the topic, e.g. a casual listener of an interview.

Let us highlight what we believe are remarkable facts derived from the classification shown in Table 4. Expert-oriented definitions seem to be safely expressed, in the genre of scientific interviews, with the combination of a non-copular verb and an intensional definition (i.e. a definition that follows the classic genus et differentia model). However, as the degree of specificity becomes lower, the pattern combinatorial increases. Note how semi-expert definitions are either extensional (i.e. enumerate the components of a term or concept) or functional (define something by describing what it does instead of what it is), and in both cases they use a linking verb. Finally, layman-oriented definitions are in general shorter and make a more extensive use of punctuation marks, synonyms and hypernyms, which seem to suggest less information being conveyed in the definition (see Table 3).

Table 3. Classification of definitions according to different levels of specialization.

Degree of Specialization	Type of Definition	Example
Expert	Verb_def + Intensional	The measurement technique that measures the total soluble amyloid-b in the brain is called enzyme-linked immunosorbent assay.
Semiexpert	Verb_def + Extensional	Nano-ear is made from a microscopic particle of gold that is trapped by a laser beam.
	Verb_def + Extensional	Climategate involved hacked e-mails from climate researchers that got released to the public and how this is affecting climate policy, especially in the UK.
	Verb_def + Functional	Homebox genes are normally involved in developmental patterning
	Is_def + Functional	OCA2 is a gene implicated in a variety of eye colors.
Layman	Punct_def + Intensional	Binary star systems – two stars revolve around a mutual center of mass.

Punct_def + Functional	Foja Mountains, home to a range of endemic birds, mammals, butterflies and frogs.
Punct_def + Extensional	Chromosomes – set of genes.
Punct_def + Hypernymic	Drosophila, a fruit fly.
Punct_def + Synonymic	Otoliths (ear bones)
Verb_def + Synonymic	Nitrous oxide is known as the laughing gas.
Is_def + Extensional	Mars atmosphere is composed dominantly of CO2.
Is_def + Intensional	Dark matter is this invisible matter which researches have speculated is responsible for basically holding the universe together.
Is_def + Synonymic	DAMA Collaboration is short for "Dark Matter Collaboration"
Is_def + Hypernymic	Folic acid is a vitamin

6. Conclusions and Future Work

We have discussed the different stages of development of a corpus of scientific interviews. Next, the theoretical framework where this work is included has been briefly pointed out (the function theory of lexicography). This theory emphasizes the importance of user needs for generating contextually relevant definitions. This is better expressed by Fuertes-Olivera and Nielsen (2011): "the levels of user competence indicate the lexicographically relevant user needs, and thus provide lexicographers with a workable basis for selecting the data that fill the lacunae relating to competence and skills". Establishing a perhaps risky boundary between expert, semi-expert and laymen audiences, we have conducted a specialization-wise user-centered classification of definitions according to the surface pattern they follow and how they convey information. We can conclude that, in general, scientific interviews seem to include more layman-oriented definitions than definitions for experts, which is consistent with their target audience. Moreover, shorter definitions which make extensive use of punctuation marks or copular verbs are extensively used in layman-oriented definitions.

To sum up, this proposal has aimed at providing a framework for user-centered definition categorization based on the degree of specialization. For this reason, we expect to extend this research in three key points that we believe are crucial in order to provide a consistent definitional taxonomy. Firstly, deciding on the degree of specialization of a definition that appears in a popularizing text can be controversial, and further research in the field of ESP seems necessary to tailor a classification that gives an answer to any questions that may arise. Secondly, extending the dataset would provide more examples that confirm or reject this classification. Finally, an end-user evaluation, where experts, semi-experts and layman evaluate the definitions would be of interest, as it would provide empiric data on the hypothesized class of a definition.

Acknowledgements

We would like to express gratitude to "Obra Social la Caixa: Programa de Becas 2011" for partially supporting this work. We would also like to thank the students of the "Erasmus Mundus International Masters in Natural Language Processing and Human Language Technology" (Universitat Autònoma de Barcelona, 2013), and the students of the "Máster Universitario en Inglés y Español para Fines Específicos" (University of Alicante, 2012/2013), as well as the "Tecnologías de la Información y el Conocimiento aplicadas al inglés y español para fines específicos" module coordinator, Dr. Borja Navarro Colorado, for his collaboration and feedback.

References

Bergenholtz, H. & Tarp, S. (2003). Two opposing theories: On H.E. Wiegand's recent discovery of lexicographic functions. *Hermes, Journal of Linguistics, 31*, 171-196

Bergenholtz, H., & Tarp, S. (2010). LSP lexicography or terminography? The lexicographer's point of view. In P.A. Fuertes-Olivera (Ed.) *Specialized Dictionaries for Learners* (pp. 27-37). Berlin/New York: De Gruyter.

Bergenholtz, H., Tarp, S., & Duvå, G. (1995). *Manual of Specialised Lexicography: The preparation of specialised dictionaries*. Amsterdam: John Benjamins Publishing Company.

Del Gaudio, R, Batista, G., & Branco, A. (2013). Coping with highly imbalanced datasets: A case study with definition extraction in a multilingual setting. *Natural Language Engineering*, *FirstView*, 1-33.

DiBella, S. M., Ferri, A. J., & Padderud, A. B. (1991). Scientists' reasons for consenting to mass media interviews: A national survey. *Journalism & Mass Communication Quarterly*, *68(4)*, 740-749.

Diéguez, M. I. (2004). El anglicismo léxico en el discurso económico de divulgación científica del español de Chile. *Revista Onomázein*, *10*(2), 117-141.

Fuertes-Olivera, P., & Arribas Baño, A. (2008). *Pedagogical Specialised Lexicography: The representation of meaning in English and Spanish business dictionaries*. Phiiladelphia: John Benjamins Publishing.

Fuertes-Olivera, Pedro A. & S. Nielsen (2011): The dynamics of terms in accounting: what the construction of the accounting dictionaries reveals about metaphorical terms in culture-bound subject fields. In R. Temmerman & M. Van Campenhoudt (Eds.), *The Dynamics of Terms in Specialized Communication. An Interdisciplinary Perspective* (pp. 157-180). Canada: John Benjamins.

Muresan, S. & Klavans, J. (2002). A Method for Automatically Building and Evaluating Dictionary Resources, 3rd International Conference on Language Resources and Evaluation (LREC'02). Las Palmas, Spain, 29-31.

Renouf, A. (1987) Corpus Development. In J. Sinclair (Ed.) *Looking Up*. London: Collins.

Sierra, G., Alarcón, R., Aguilar, C. & Barrón, A. (2006). Towards the building of a corpus of definitional contexts, 12th EURALEX International Congress. Torino, Italy, 229–240.

Suau, F. (2005). Matizadores discursivos frente a elementos apelativos y fáticos o la importancia de gustar a la audiencia: comparación entre artículos de investigación y de divulgación científica en inglés desde el metadiscurso. In G. Aguado de Cea & G. Salom (Eds.) *Revista Española de Lingüística Aplicada (ReSLA), volumen monográfico sobre Lenguas de Especialidad en España*.

Tapanainen, P., & Järvinen, T. (1997). A non-projective dependency parser, 5th Conference on Applied Natural Language Processing. Washington D.C., United States, 64-71.

Tarp, S. (2008). Lexicography in the borderland between knowledge and non-knowledge: General lexicographical theory with particular focus on learner's lexicography. *Lexicographica. Series Maior, 134*. Tübingen: Max Niemeyer.

Westerhout, E. & Monachesi, P. (2007). Extraction of dutch definitory contexts for elearning purposes, Computational Linguistics in the Netherlands (CLIN 2007). Nijmegen, Netherlands, 219–34.