Cyber-Physical Systems for Social Applications

Maya Dimitrova
Bulgarian Academy of Sciences, Bulgaria

Hiroaki Wagatsuma
Kyushu Institute of Technology, Japan

A volume in the Advances in Systems Analysis, Software Engineering, and High Performance Computing (ASASEHPC) Book Series

Published in the United States of America by
IGI Global
Engineering Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA, USA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Names: Dimitrova, Maya, 1961- editor. | Wagatsuma, Hiroaki, editor.
Title: Cyber-physical systems for social applications / Maya Dimitrova and Hiroaki Wagatsuma, editors.
Description: Hershey, PA : Engineering Science Reference, an imprint of IGI Global, [2019] | Includes bibliographical references and index.
Identifiers: LCCN 2018037299| ISBN 9781522578796 (hardcover) | ISBN 9781522578802 (ebook)
Subjects: LCSH: Automation. | Cooperating objects (Computer systems) | Robotics--Social aspects. | Robotics in medicine.
Classification: LCC TJ213 .C8855 2019 | DDC 303.48/3--dc23 LC record available at https://lccn.loc.gov/2018037299

This book is published in the IGI Global book series Advances in Systems Analysis, Software Engineering, and High Performance Computing (ASASEHPC) (ISSN: 2327-3453; eISSN: 2327-3461)

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Chapter 6
Designing an Extensible Domain-Specific Web Corpus for "Layfication": A Case Study in eCare at Home

Marina Santini
RISE Research Institutes of Sweden, Sweden

Marjan Alirezaie
Örebro University, Sweden

Arne Jönsson
RISE Research Institutes of Sweden, Sweden & Linköping University, Sweden

Leili Lind
RISE Research Institutes of Sweden, Sweden & Linköping University, Sweden

Wiktor Strandqvist
RISE Research Institutes of Sweden, Sweden & Linköping University, Sweden

Eva Blomqvist
RISE Research Institutes of Sweden, Sweden & Linköping University, Sweden

Gustav Cederblad
Linköping University, Sweden

Maria Lindén
Mälardalen University, Sweden

Mikael Nyström
RISE Research Institutes of Sweden, Sweden & Linköping University, Sweden

Annica Kristoffersson
Örebro University, Sweden

ABSTRACT

In the era of data-driven science, corpus-based language technology is an essential part of cyber physical systems. In this chapter, the authors describe the design and the development of an extensible domain-specific web corpus to be used in a distributed social application for the care of the elderly at home. The domain of interest is the medical field of chronic diseases. The corpus is conceived as a flexible and extensible textual resource, where additional documents and additional languages will be appended over time. The main purpose of the corpus is to be used for building and training language technology applications for the "layfication" of the specialized medical jargon. "Layfication" refers to the automatic identification of more intuitive linguistic expressions that can help laypeople (e.g., patients, family caregivers, and home care aides) understand medical terms, which often appear opaque. Exploratory experiments are presented and discussed.

DOI: 10.4018/978-1-5225-7879-6.ch006

INTRODUCTION

Cyber-Physical Systems (CPSs) denote an emergent paradigm that combines most advanced technological approaches and computational tools to solve complex tasks. CPSs are domain-independent and have penetrated diversified disciplines, such as healthcare and self-driving vehicles. Corpus-based Language Technology is an essential component of many CPSs, where linguistic knowledge is indispensable to prevent failures or fatal errors due to misunderstandings or poor understanding.

Web corpora are the bedrock underlying modern real-world corpus-based Language Technology applications (henceforth LT applications), such as terminology extraction, ontology learning, text simplification, automatic summarization and machine translation. In this chapter, we describe the design and the development of an extensible domain-specific web corpus to be used in a distributed social application for the care of the elderly at home.

Web corpora are text collections made of documents that have been automatically retrieved and downloaded from the web. Generally speaking, building web corpora is convenient because the whole process of corpus creation is automated, fast and inexpensive. In contrast, the construction of traditional corpora– such as the British National Corpus (BNC) (Burnard, 2007) or the Corpus of Contemporary American English (COCOA) (Davies, 2009) or the recent iWeb corpus[1]– normally spans over several years, relies on considerable amount of human expertise to decide the ideal combination of documents that is worth storing in the corpus and, last but not least, necessitates substantial funding. It goes without saying that the investments in time, financial resources and human knowledge required by traditional corpora are well paid-off because such an effort amounts to high-quality and long-lasting collections, that are extensively used by teachers, students, researchers and system developers. For instance, the Brown corpus created in the 60's (Kucera & Francis, 1979) is still valuable today, especially for monitoring how the language has changed in the last decades (e.g. Malá, 2017).

While traditional corpora are a shrine of hand-crafted qualities, the added value of web corpora is in their malleability. Similar to traditional corpora, web corpora can be general-purpose or specialized (Barbaresi, 2015) and may serve different purposes, such as linguistic studies (e.g. Schäfer & Bildhauer, 2013; Biemann et al., 2007; Lüdeling et al. 2007) and professional uses (Goldhahn et al., 2012; Baroni et al., 2006). However, the unique and unprecedented potential of web corpora is that they can promptly and inexpensively account for virtually any domain, topic, genre, register, sublanguage, style and emotional connotation, since the web itself is a panoply of linguistic and textual varieties. This potential can be profitably exploited for domain-specific projects that require specialized text collections to implement corpus-based LT applications. Examples of these types of LT applications are those implemented in projects like *DigInclude*[2] and *E-care@home*[3] in Sweden or those that have been developed for European projects, such as *SEMANTICMINING*[4] and *SemanticHealthNet*[5] in the semantic interoperability field, as well as *Accurat*[6], *TTC*[7] and *EXPERT*[8] in Natural Language Processing (NLP), Computational Linguistics and Information Retrieval.

Arguably, traditional corpora and web corpora are complementary and allow for a wide spectrum of possible linguistic, empirical and computational studies and experiments.

Since web corpora are often at the core of LT applications, seemingly the design and the quality of web corpora affect the reliability and the performance of final applications. Building a 'clean' corpus with selected documents requires time, careful planning, long-term decision-making and extensive funding. Frequently, in the implementation of LT applications, the corpus is only a single piece (even though an important one) of a complex pipeline, and often the time and financial resources allocated

for corpus creation are limited. For this reason, bootstrapping corpora from the web (either via web crawling or via search engines) has become normal practice. Corpora built from the web are convenient because their creation is fast and inexpensive, although corpus evaluation is not yet fully standardized (cf. Kilgarriff et al. 2011), and it is hard to replicate results or to generalize on the findings, especially when web corpora are domain-specific.

The Whys and Wherefores

The version of the web corpus described in this chapter is known as *eCare_Sv-En_03*. It contains web documents written in English and in Swedish. We propose the construction of an extensible web corpus which should be seen as an ever-changing textual resource, i.e. as a corpus that is constantly in-progress, where web texts can be added when needed and where a light set of metadata keeps track of updates and allows for the extraction of virtual sub-corpora.

The rationale underlying the creation of *eCare_Sv-En_03* stems from the following needs: (1) having publicly-available medical web documents to represent a fine-grained medical domain (e.g. chronic diseases); (2) having a corpus with a design and a structure that allows for expansion with additional documents and languages to account for research, development and commercialization; (3) accounting for very specific technical terms, in our case both specialized and lay medical terms, that can meet the needs of two broad user groups, namely medical professional staff and health consumers, like patients, family caregivers and home-care aides, who are not expected to have any specific medical education.

Our perspective on web corpora is from the point of view of the implementation of corpus-based real-world LT applications in specialized domains. Our ambition is to find ways to build LT applications that are efficient in terms of time and financial resources, and that require the least implementation effort.

Essentially, we take a *minimalist* approach. Our assumption is that not all applications need large and clean corpora, and our ambition is to understand to what extent a corpus can be small and noisy without negatively affecting the performance of an application. More prosaically, we would like to save time and economic resources because building large corpora and cleaning them require time and funding that are not always available in real-world settings.

In practical terms, this means that we try to identify the corpus critical mass for a specific LT application. In this context, critical mass indicates the minimal corpus size that an LT application needs to achieve a "good enough" performance. We also try to understand whether we can build LT applications using noisy documents. In short, we would like to build reliable LT applications using small corpora containing noisy documents.

Our research is somewhat complementary to the current challenge being met by other research lines, which focus on the construction of large-scale web corpora. Examples of this corpus typology include enC^3 (Kristoffersen, 2017), C4Corpus (Habernal et al. 2016), the web corpora created within the COW initiative (Schäfer & Bildhauer, 2012), and those constructed in the WaCky project (Baroni et al. 2009). These large-scale web corpora will certainly help the progress of NLP, as pointed out by Biemann et al. (2013) and Habernal et al. (2016), especially when using neural networks for deep learning or word embedding, since these algorithms require a large quantity of data in order to be effective.

Meeting such a challenge often implies an impressive distributed architecture (such as the Hadhoop MapReduce framework, e.g. see Biemann et al., 2013) that in certain cases is impractical. What is more, large-scale web corpora are "static" (as pointed out in Biemann et al.; 2013, see also Schäfer & Bildhauer, 2013). In this respect, their design is similar to traditional corpora, which are not designed

to be extended (although some of them are available in several releases). These corpora are more of a huge snapshot of the language of the web at a certain point in time. For example, the C4Corpus has been built with a CommonCrawl dating from 2007 to 2015 and has not been updated by adding new texts after 2016-04-14[9]. The static corpus design is certainly beneficial for many empirical studies and NLP tasks. It is be less beneficial for a live real-world LT application that thrives on frequent updates of the underlying corpus to encompass the new terms and the new findings that are constantly being produced by modern science. In a word, the language of static corpora "age" over the years. Even the much welcome CommonCrawl data is affected by this "aging" process, as pointed out by Barbu (2016) who writes: "the way that Common Crawl collects data is not by crawling live sites". Barbu himself uses a list of web urls provided by the defunct searching engine Blekko[10] and downloads the pages corresponding to those links. This means that for some sites there is a huge gap between the content of the site in Common Crawl and the live content of the site." This aging factor may be irrelevant for some tasks (such as morphology, syntax or discourse analysis), while it may not be ideal for some others (e.g. terminology extraction or ontology learning from text).

It is indeed the case that, in some subject fields and for some topics, there is often the need to update a document collection with the most recent texts, containing novel findings, new issues or unprecedented cases, new terms, new medical devices, new medications, as well as the latest discoveries. For this reason, we propose a corpus design and a corpus structure that can accommodate incremental corpus extension over time and when needed, and where documents, languages, metadata and specific topics can be smoothly added or rearranged.

In summary, we need a corpus design that is flexible, replicable and "good enough" to: 1) keep track of diversified textual traits and 2) orderly stratify the successive corpus developments. Depending on the purpose of a specific LT application, a corpus designed in this way will allow for either the use of the corpus as a whole, or of portions (sub-corpora), thus facilitating corpus re-use.

Importantly, the texts in the corpus do not need to be uniformly annotated. For example, a portion of the corpus may be annotated as lay or specialized, while another part may be annotated by readability or genre. What is important is that the subpart of interest can be easily identified and extracted from the whole corpus, thus creating virtual textual collections that serve specific purposes.

To build such a corpus, we were inspired by the Agile methodologies[11] that are based on iterations and incremental developments. To the best of our knowledge, such a corpus design has not been proposed to date. We present the construction of our corpus in Section 5.

A Corpus for Layfication

The medical domain centers upon specialized and technical notions elaborated and usually disseminated by healthcare professionals. These notions often remain opaque and incomprehensible for non-expert users, and especially for patients (Berland et al., 2001). Despite it is acknowledged that understanding what the doctor says has an important influence on the success of treatments, in many cases medical terminology hinders the comprehension of various groups of people (such as non-native speakers, people with low-education, etc.), and has negative effects on the health consumer user group (e.g. patients and caregivers). The main actors of the medical field are physicians and patients. However, also students, pharmacists, managers, biologists, nurses who have different levels of expertise need to interact and understand each other (Tchami & Grabar, 2014). We focus on two broad user groups. The first group (the expert) includes those who use and understand medical specialized terminology, such as healthcare

professionals. The second group encompasses "ordinary people" (the lay), i.e. people without medical education, who struggle to get a grip on the medical jargon. It is true that "ordinary people" are exposed to medical terms through the media (e.g. radio, TV and newspapers) and some of them who suffer from a chronic disease may become experts on their own illness. This knowledge, however, is not reliable, since, as observed in several studies, "ordinary people" might misunderstand medical information in good faith (Claveau et al., 2015; Bigeard et al., 2018).

In this chapter the term "layfication" refers to the automatic identification of more intuitive linguistic expressions that can help laypeople (mostly patients and family, family caregivers and home-care aides) understand medical terms, which often appear neboulous and incomprehensible. Typical examples of the linguistic dichotomy existing in the medical field are words like "anemia" -- also written anaemia -- (a specialized term) vs. "lack of iron" or "iron deficiency" (lay synonyms). Lay synonyms are lexical items that are based on common words, so that an expression like "lack of iron" is more intuitive than the medical term "anemia".

Although medical terms are more precise and less ambiguous than their lay counterpart, it has been widely acknowledged that consumer health information is often inaccessible to healthcare consumers (Miller et al., 2007). When dealing with the lay user group, it becomes apparent that the precision and lack of ambiguity of the medical term does not necessarily benefit the laypeople since it creates a communication gap that entails detrimental consequences for the patient's health due to misunderstandings or partial understanding. It has been repeatedly stressed that it is important that people who receive health care and medical treatments but do not have a medical education (normally patients and caregivers) are helped to fully understand the medical language used by healthcare professionals. Helping laypeople by providing them with lay synonyms (e.g. using "lack of iron"[12] rather than "anemia") or reformulation (e.g. "Anaemia is a lack of red blood cells"[13]) can help prevent unwanted consequences such as the misunderstandings (Claveau et al., 2015) that may cause medication misuses (Bigeard et al., 2018). A better understanding of medical jargon is especially important for elderly people affected by chronic diseases because it facilitates a proactive behaviour and fosters self-empowerment, which has proven to be beneficial for long-term successful treatment (Fotokian et al., 2017).

Nowadays, the creation of medical lay variants is mostly corpus-based (see Section 4). Normally, the corpora for this task are created by going to specific pre-defined web sites and downloading lay and specialized medical texts. Using this approach is theoretically profitable because corpora can be built with the material available. However, it has a reduced applicability in real-world domain-specific LT applications because these websites do not contain all the illnesses but only the most common ones, like "fever" or "allergy". The same is true for user-generated texts, such as those that can be found in forums and blogs, since users mostly talk about general problems or common diseases. Another common approach to build medical corpora has been to focus on journals or, more rarely, on patient record collections but in these cases, there exist copyright, ethical and legal restrictions that limit the shareability and experimental replicability.

For all these reasons, with *eCare_Sv_En_03* we are exploring a different avenue. More specifically, with *eCare_Sv_En_03* the idea is to pre-select some very specific medical terms (not just the most common illnesses) that represent the granularity of domain of interest, use them as seeds in a search engine and download only the pages that are related to the specific terms we focus on. In practice, we aim at building a corpus that contains documents that are related only to specific medical terms that indicate chronic diseases, and that are not always documented in medical websites, such as the Swedish medical information portal called "1177 Vårdguiden[14]".

eCare_Sv_En_03, the current version of the *E-care@home Corpus*, does *not* rely for its annotation on documents coming from specific sources (a method that was also used in Santini, 2006 and referred to as "annotation by objective sources"). Here, we reverse the approach. We start from our topics of interest (i.e. chronic illnesses) and search for the material that is available on the web at a certain point in time. At retrieval time, we make no distinction between lay and specialized web sites. Rather, we follow the approach initiated by Glavas and Stajner (2015) within text simplification. These authors observe that "'simple' words, besides being frequent in simplified text, are also present in abundance in regular text. This would mean that we can find simpler synonyms of complex words in regular corpora, provided that reliable methods for measuring (1) the 'complexity' of the word and (2) semantic similarity of words are available.". Inspired by this remark, we build a web corpus of domain-specific documents retrieved by search engines in the searchable web. Then, we put forward the hypothesis that in this way the corpus include both lay and specialized documents and, consequently, lay and specialized terms. This hypothesis will be tested in Section 5.1.

Research Questions and Objectives

The research questions motivating this work relate to the creation of real-world, domain-specific and corpus-based LT applications. We investigate whether it is possible to:

- Find an agile corpus design that accounts for incremental expansions according to real-word needs that may occur over time (e.g. multilinguality and additional text types);
- Use a minimalist approach to LT applications that ensure good enough performance and easy replicability and/or portability to other domains (application of Occams's razor law as used in the context of machine learning and data science[15]).
- Downplay the effects of noise and corpus size variations.

We investigate possible answers to these research questions by carrying out a number of experiments with clear objectives, namely by:

1. Implementing the design of a web corpus that is conceived as "work-in-progress", i.e. an *extensible,* open-ended and multilingual textual resource, where each stage of the construction is useful to gain insights into some aspects of language and/or language technology (Section 5.1).
2. Automatically classifying texts written for laypeople from those written for the expert and explore the effect of noise and corpus size variations (Section 5.2)
3. Creating a distributional thesaurus by inducing words related to chronic illnesses from a small corpus in Swedish (Section 5.3);
4. Expanding the corpus with documents in English and assessing the domain-specificity, or domain-hood, of the English sub-corpora using well-established language independent statistical measures (Section 5.4).

The experimental investigations presented in this chapter are still exploratory but lay the groundwork for further research and future development.

The chapter is organized as follows: Section 2 briefly describes a distributed CPS where the Internet of Things (IoT) and Language Technology (LT) meet each other to support elderly eCare at home for chronic diseases; in Section 3 the working hypothesis underlying our investigation is set out, and the intrinsic challenges are spelled out; in Section 4, previous research on layfication is summarized; Section 5 subsumes four subsections, each one presenting experiments and discussions; finally in Section 6 conclusions are drawn and future directions are outlined.

THE INTERNET OF THINGS IN E-CARE: TOWARDS SEMANTIC INTEROPERABILITY

Prevention and adaptive support to ageing population is an important objective in today's society. Telemedicine, robotics and the IoT (Internet-of-Things) have made a giant leap forward in providing solutions to overcome the challenge of helping patients who live alone.

Telemedicine is the use of telecommunication and information technology to provide clinical health care from a distance. It has been used to overcome distance barriers and to improve access to medical services that would often not be consistently available in distant rural communities. Telemedicine is a field that is widely developed in geographically extended countries, like the United States and Sweden (for recent advances in this field, see Lilly et al., 2014 and Lind & Karlsson, 2018 respectively).

In addition, robotics has provided intelligent machines that help patients to be more independent. For instance, the EU research project GiraffPlus[16] (Coradeschi et al., 2014; Coradeschi et al., 2013) monitored activities and physiological parameters in the home using a network of sensors. The telerobot Giraff was used to communicate with elderly patients. Recently, also social robots for home (e.g. Jibo and Buddy) have been launched as context-based social artificial companions that verbally interact with humans and help them in several activities (Quintas, 2018).

Extending previous experience in telemedicine and robotics, *E-care@home* (a Swedish research project running from 2015 to 2020), is creating new knowledge and exploring novel avenues for the smooth and robust implementation of eCare for the multimorbid and frail elderly living at home.

E-care@home is a multi-disciplinary project that investigates how to ensure medical care at home and avoid long-term hospitalization in the eldercare (Loutfi et al., 2016). Long hospitalizations are discomforting for elderly patients and expensive for the national healthcare systems. Providing medical care at home to the elderly can be effective by populating the home with electronic devices ("things"), i.e. sensors and actuators, and linking them to the Internet. Creating such an IoT infrastructure is done with the ambition to provide automated information gathering and processing on top of which e-services can be built trough reasoning (Sioutis et al., 2017). The rapid growth of data from sensors can potentially enable a better understanding and awareness of the environment for humans. For example, "[i]n Japan, an estimated 6.24 million people aged 65 or older were living alone in 2015, exceeding the 6 million mark for the first time, according to a welfare ministry survey released in July 2016."[17].

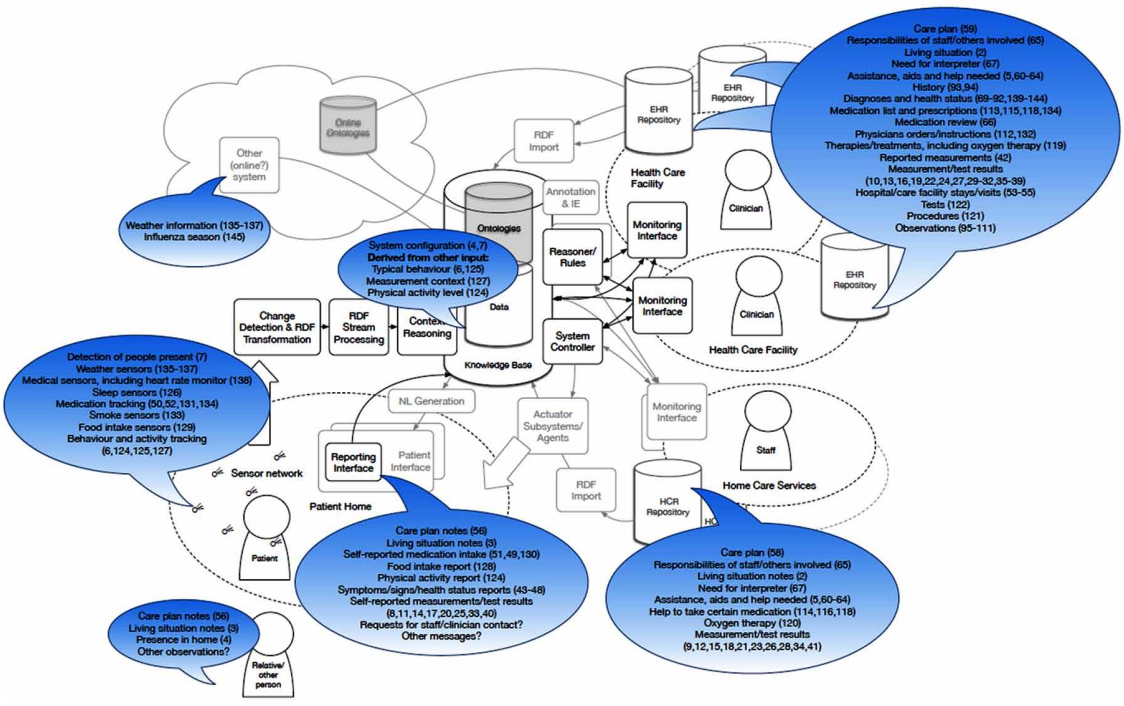
E-care@home: Semantic Interoperability

The interpretation of sensor data needs to be both machine-readable and human understandable. In order to be understandable for humans, interpretation of data may include semantic annotations in the form of context-dependent terms that hold the meaning of numeric data. Information gathered by sensors are

lists of numbers. It is possible however to convert these bare numbers into specialized semantic concepts (Alirezaie, 2015). This conversion complies to one of major objectives of *E-care@home*, i.e. to represent information in a "human consumable way", since the project focuses on technological solutions and uses artificial intelligence for creating a semantic interoperability between sensor data, systems and humans (Kristoffersson and Lindén, 2017). The international challenge of "Patient Empowerment" implies that patients should contribute to their health and include their perspectives for shared decision making with clinicians. Standard international classifications or terminologies are also needed to implement semantic interoperability of the whole system (Cardillo, 2015). This implies using and creating different types of terminologies for different levels of medical expertise and for multiple languages.

A simplified version of the architecture for *E-care@home* semantic interoperability that would allow for all the different data sources to talk to each other is shown in Fig 1. Fig 1 is a conceptual system overview completed by balloons showing where all the data would come from. Data has here been placed as far out towards the sides of the picture as possible, e.g. we imagine that all the sensor data and the reports from the patient would then be stored in the central Knowledge Base (KB) of the home system, but in the picture we show where it entered the system, because that says more about its potential format, how reliable it may be etc. than placing everything at the center. What is placed at the center of the picture is such things that have to be derived from other data that comes in, and hence, actually originates from some of the processing components that would directly operate on the KB content. The semantic interoperability of several data sources has already been implemented in a series of ontologies (Alirezaie et al., 2018a; Alirezaie et al., 2018b). Lay medical vocabulary is also going to be integrated in the whole architecture.

Figure 1. Simplified semantic interoperability architecture



The Contribution of Language Technology

Language Technology is an essential part of an eCare solution, since it empowers patients and other non-professional actors to understand medical information. The focus is on medical terminology that is sorted out based on its explanatory level, either for medical professionals (the expert) or for non-professionals (the lay). The terminology is extracted from documents retrieved from health-related sites on the web. Other applications, like social robots, can benefit from using the methodology for domain-specific web corpus generation in the process of online communication with a person at home, who seeks assistance with monitoring and explaining health related issues.

To date, linguistic understanding of sensor data targets clinicians and other professional staff[18]. To our knowledge, very little research exists on the conversion of sensor data targeted to patients. In *E-care@home*, Language Technology helps enhance patients' self-empowerment. In the project, a lay-specialized textual corpus is being prepared for the automatic extraction of lay terms and paraphrases that match specialized medical terminology used by healthcare professionals. "Lay" means that a document has been written for readers who do not need to have a domain-specific knowledge (e.g. patients, their relatives, home-care aides, etc.). "Specialized" means that a document is written for professional staff (e.g. physicians, nurses, etc.). Research on lay-specialized sublanguages is long-standing and spawn by the need to improve communication between two specific user groups: the layman on one side, and the expert on the other side. A classic example of a specialized term is "varicella", which patients often call "chicken pox". The word "varicella" is a medical term used by healthcare professionals (experts), while "chicken pox" (together with its graphical variant "chickenpox") is a lay paraphrase commonly used by patients (laypeople). Within the *E-care@home* project, the Language Technology group is working to provide methods and tools for the automatic extraction of the lay-specialized linguistic variations.

Converting numbers into concepts expressed in a natural language that experts can understand is certainly a big step forward and it is especially valuable for healthcare professionals, who can use this converted information for timely decision-making. However, since in the *E-care@home* framework patients are empowered and take active part in the management of their illnesses, it is no longer enough to convert sensor data to a medical language that only experts understand. Patients too should be included in the information cycle. There are linguistic hinders, though, as highlighted earlier.

CHALLENGES AND OPEN ISSUES

The research questions and the experiments presented in this chapter contribute to the design and implementation of LT applications for E-care@home. However, several challenges lie on the way. We briefly discuss them below.

Corpus Design: An Extensible Web Corpus

As mentioned above, the purpose of the *E-care@home Corpus* is to be used to build and/or train domain-specific LT applications for eCare and eHealth. We need a corpus whose design is dynamic and flexible, and where additional documents and several languages will be appended over time. Currently, corpus construction practice is still in a stage where a corpus is built as a "static" collection, that is a representative text collection of one or multiple languages of one or several domains at a certain point

in time. Methods have been proposed to expand corpora for specific purposes, e.g. for Statistical Machine Translation (Gao & Vogel, 2011) or for paraphrase generation (Quirk et al., 2004). However, these corpora expansions are made of artificial sentences, generated by algorithms trained on large volumes of sentence pairs, and not by adding running texts. Similarly, the approach used by Zadeh (2016) to study the effect of corpus size on the parameters of a distributional model is ad-hoc and one-off rather than driven by a long-lasting design. The first challenge is then to figure out how to design a dynamic, extensible corpus. As explained in Section 5.1, we propose an agile approach based on iteration and incremental developments to meet the needs when they arise. Therefore, at this stage, *eCare_Sv_En_03* is not incomplete or unfinished: it is at an early stage and has its own validity and usage.

Domain-Granularity

We claim that being focused around specific medical terms, and not common diseases, is important for real-word LT applications that aim at solving very specific problems in our society. Although common medical words are important for many purposes, fine-grained domain granularity plays an important role too. As pointed out by Lippincott et al. (2011) "while variation at a coarser domain level such as between newswire and biomedical text is well-studied and known to affect the portability of NLP systems, there is a need to develop an awareness of subdomain variation when considering the practical use of language processing applications […]". Essentially, we are pushing the limit of domain granularity towards subdomains, and this is our second challenge.

Language Varieties: Lay vs. Specialized

Medicine is a domain where there exists a divide between the language used by healthcare professionals and the language normally used and understood by patients, family caregivers or home-care aides. This is a well-known problem that is extensively researched (see Section 4).

The need of lay synonyms or lay paraphrases that match specialized medical terminology used by healthcare professionals has been the focus of recent research, both in Language Technology (Deléger et al. 2013), and in the clinical community (Seedor et al. 2013). Research on lay-specialized sublanguages is brought about by the need to improve communication between two specific user groups: the layman on one side, and the domain expert on the other side (Miller & Leroy, 2008; Smith & Wicks, 2008; Soergel & Slaughter, 2004). Solid studies show that the gap exists and is detrimental for patients (e.g. Chapman et al., 2003). The importance of matching lay and specialized vocabulary is emphasized by Williams & Odgen (2004) whose study shows that "a doctor's choice of vocabulary affects patient satisfaction immediately after a general practice consultation and that using the same vocabulary as the patient can improve patient outcomes". Thus, the issue of patient empowerment, as well as the development and evaluation of generic methods and tools for assisting patients to better understand their health and healthcare, has been the goal of several EU-funded projects[19]. Unfortunately, while the language and terminology used by professionals are subject to control by continuously evolving standardization, usage of medical terms on the part of laypeople is much more difficult to capture[20].

To date, there is no agreed lexical expression that subsumes concepts such as "lay", "normal", "simplified", "expert", "specialized", "consumer health vocabulary", "consumer terminology", "in plain language", and the like. Researchers use different expressions to indicate these kinds of language varieties, for instance, "different genres (such as specialized and lay texts)" (Deléger et al., 2009); "discourse types

(lay and specialized)" (Deléger et al. 2013); or "registers" (Heppin, 2010). Most commonly, however, researchers do not relate the specialized-lay varieties to any superordinate category (as in Abrahamsson et al., 2014).

Classifying language varieties into categories is a difficult exercise. This is not only the case for the "lay variety" but for any textual dimensions, such as style, genre, domain, register and similar. Dozens of definitions exist for each of these textual varieties (as appropriately pointed out in Lee, 2001), and a common conclusion is that the classification into these textual categories is slippery, since no standard and agreed upon characterization is currently available, but there exist different schools of thought and different needs.

Lay vs. specialized language varieties could go under umbrella terms like "discourse" or "communication" or "language for special purposes", or referred to as "register" or "genre" or "sublanguage", and more. Any of these categories do not fully capture the lay-specialized distinction, and any ontological decision may be either questioned or supported, depending on the researchers' personal stances on textual classification schemes.

Since this long-standing discussion is still ongoing, we contribute to it by suggesting the adoption of the category "sublanguage" to refer to the different language varieties employed by user groups when they talk about topics that belong to specialized disciplines, such as medicine and law.

Normally, a sublanguage refers to a technical language (Kittredge & Lehrberger, 1982; Grishman & Kittredge, 2014) or jargon used in restricted communities (e.g. the jargon used by teenagers stored in the Corpus of London Teenagers (Haslerud and Stenstöm, 1995) or to a very specialized domain-specific communication style (e.g. the "notices to skippers"). Both in linguistics and in computational linguistics, a sublanguage is characterized by domain-specific terms (or word co-occurrences) and syntactic cues that deviate from normal language use (Kittredge, 2003: 437; Basili et al. 1993; O'Brien, 1993 lists several definitions of sublanguage). We can safely say that the medical jargon used by physicians and other healthcare professional staff is a sublanguage. What about the language used and understood by patients when they talk about medical topics? It is not properly speaking "general language", it is not a "register"[21], i.e. a language variety used in special situations or contexts as listed in the standard ISO 12620 on Data Category Registry[22], it is not a genre, and it is not a domain. It is indeed a type of discourse. To be fair, we should call it "layspeech" or "patientspeak" as proposed by Scott & Weiner (1984). Although less restricted than the domain-specific technical sublanguage used by professional staff, the layspeech is also domain-specific. According to Kittredge (2003): "Restricted subsystems of language can arise spontaneously in a subject-matter domain where speech or writing is used for special purposes". Leveraging on this observation, we broaden the definition of sublanguage in order to encompass the non-overlapping language varieties that are commonly used when two or more user groups communicate in specific domains on certain topics. While in previous definitions, the notion of sublanguage indicated either a domain-specific jargon or a community jargon, in the sublanguage definition proposed here we combine the connotation of domain specificity and user group usage. This definition of sublanguage is more flexible and more accurate because it has two attributes, the domain (e.g. medicine, law, etc.) and the user group (e.g. experts, laypeople, novices, learners etc.). It is worth noting that although in the experiments presented here we are using only the lay vs specialized categories, healthcare actors are heterogeneous, including a wide variety of backgrounds, levels of medical literacy and ages.

In this complex landscape, a more flexible characterization of sublanguage allows us to refer to a language variety so that we can use formulations such as: "medical professional and lay sublanguages" or "medical professional, learners' and lay sublanguages", where "medical" refers to the domain, and

"professional", "learners" and "lay" indicate the levels of medical literacy of a user group whose language use is going to be analyzed (cf. Zheng et al., 2002; Miller et al., 2007). This modularity can be easily exported to other domains (e.g. the legal domain, see Heffer et al., 2013 or the business domain[23] or the marketing domain[24]), so we can say "legal lay sublanguage" or "business specialized sublanguage" and so on.

Arguably, this definition of sublanguage is more flexible and applicable to all the domains where the domain-specificity of a jargon causes some kind of "diglossia" or "polyglossia" that causes a gap in human communication. Following the extended definition, we can then say that in the medical domain, two sublanguages normally come in contact, namely the lay sublanguage used by patients and their relatives (the lay) and the specialized sublanguage used by healthcare professionals (the expert).

Normally, lay synonyms are based on everyday language, and are easier to read and to understand than medical terminology, which conversely have highbrow connotation. For ordinary people without a medical education or background, medical terms are often opaque or hard to remember due to the Greek and/or Latin etymology. These terms are called "neoclassical" terms, and, interestingly, recent research shows that also healthcare professionals tend to "normalize" this type of lexicon to everyday language, as in the case of "Swedification" of Latin and Greek affixes in patient records (Grigonyte et al., 2016). Generally speaking, it seems that the layfication of medical language is an extensive phenomenon that affects, in different ways, several user groups. It must be emphasized that the lay sublanguage is not as accurate as the specialized sublanguage. Lay medical terms, when they exist, are indeed more transparent and more easily understood by laypeople. Again, consider the specialized medical term "varicella" and its lay synonym "chickenpox". Both varicella and chickenpox are medical terms, one highbrow and the other one colloquial. The same high-low connotation can be found in the words surrounding the medical terms, e.g. the verb "alleviate" can be rendered by "decrease" in lay texts. Presumably, the lay sublanguage shares similarities across all languages (cf. also Grabar et al. (2007), since it is a phenomenon of text simplification.

Noise

The concept of noise is tightly linked to the concept of quality. Recently, several researchers have investigated this aspect of web texts (e.g. Biemann et al. 2013; Barbaresi, 2015). In particular, Schäfer et al., 2013 have proposed text quality evaluation in the form of the so-called Badness score. That is, a document receives a low Badness score if the most frequent function words of the target language have a high enough frequency in the document. The Badness score is based on research findings in language identification and web document filtering (Grefenstette, 1995; Baroni et al., 2009).

In this chapter, we consider two main forms of noise. The first type of noise (cf. also Versley & Panchenko, 2012) is in the form of misspellings, mis-tokenizations, encoding problems, scattered html tags, residual url chunks and incoherent punctuation caused by boilerplate removal. The second form of noise refers to badly written texts and more precisely noise caused by the presence of automatically translated texts, which have been published on the web without post-editing or proofreading. Since we aim at finding a quick and replicable methodology to compile reliable web corpora with minimum curation, we wish to explore to what extent corpus-based LT applications are tolerant to these kinds of noise. We are aware that certain LT applications require corpora that meet certain quality requirements, for example in Machine Translation, as pointed out by Escartin and Torres (2016). However, our effort is geared towards noise-resistant applications. For example, as presented in Section 5.2, we noticed

that noise become irrelevant and neutralized when using a bag-of-words approach combined with the StringsToVector filter as explained in Section 5.2.

Small (Data) Is Beautiful: The Minimalist Approach

Many recent web corpora have been built using data from the CommonCrawl Foundation, which is the largest crawl in the world (e.g. Kristoffersen, 2017; Habernal et al., 2016; Schäfer, 2016). However, size is not everything. As pointed out by Kistoffersen (2017:1), large corpora are time-consuming. This author reports that it takes some 18 hours just to read a snapshot of web content distributed by the CommonCrawl Foundation. Additionally, Remus & Biemann (2016) highlight that "large-scale data is largely collected without notions of topical interest. If an interest in a particular topic exists, corpora have to undergo extensive document filtering with simple and/or complex text classification methods. This leads to a lot of downloaded data being discarded with lots of computational resources being unnecessarily wasted".

Up to few months ago, the ruling catchphrase was *big data*. Now the opposite concept starts gaining momentum: *small data*. The concept has been created for sales or customer analytics, but now it has expanded not only to healthcare[25] but also to text analytics and corpus construction.

In this work, we wish to strike the balance between corpus size (as small as possible depending on the application), time (as short as possible) and speedy portability (as fast as possible) of LT models to other domains.

Regardless of the current size of the *eCare_Sv_En_03*, small data is an interesting concept in itself. According to current definitions small data is data that has small enough size for human comprehension. As a matter of fact, the small size of *eCare_Sv_01* (one of *eCare_Sv_En_03*'s sub-corpora) has given us the opportunity to detect phenomena such as the noise caused by automatically translated web pages and the inter-rater disagreement due to user group bias. With a much larger corpus, these fine-grained phenomena would go unnoticed or it would have taken much more time to be identified. We argue that for many problems and questions, small data in itself is enough. The challenge of small data is to find the ideal "critical mass" that benefits the application of interest. This critical mass changes from application to applications (see Section 5).

PREVIOUS WORK: LAYFICATION

By layfication, we refer to empirical and computational approaches to the automated identification, extraction, classification of lay and specialized sublanguages. In this section, we summarize research efforts made to characterize, detect or discriminate lay vs. specialized texts in the medical field. Previous work in this area is extensive, although not exhaustive, since more research is still needed.

In this cursory overview, we divide previous work in three broad areas, namely studies focusing on the relationships between readability and lay sublanguage; automatic induction of lay terminology; and finally, automatic lay-specialized text classification. For a more exhaustive overview of previous work in this field, see Åhlfeldt et al. (2006).

As pointed out by Zeng et al. (2007), lay terminology is more challenging to identify than professional health vocabulary and medical terminology. This is because lay terms are more ambiguous and more heterogeneous than medical technical terms. This state of affairs is well-described by Zeng and Tze (2006): "When producing words to describe health-related concepts, a lay person may use terms

such as hair loss and heart failure without knowing their technical definitions or use general language expressions to describe familiar concepts (e.g., loss of appetite for anorexia and pain killer for analgesic). The range of lay expressions seems to vary from general and descriptive (e.g., device to look inside my ear for otoscope) to specific, but colloquial (e.g., sugar for diabetes). Thus, lay discourse on the health-related topics often includes a combination of technical terminology and general language expressions, with many possible interpretations based on individual, contextual, societal, and cultural associations. The challenge is to sort out the different ways consumers communicate within distinct discourse groups and map the common, shared expressions and contexts to the more constrained, specialized language of professionals, when appropriate.". The difficulty is not only about medical expressions per se but also in words that are not technical, but are used as technical terms in the medical jargon, e.g. "alleviate" or "apprehensive" etc. (Scott & Weiner, 1984).

Several researchers have investigated the relation between *readability* and lay/specialized sublanguage (e.g. Ownby, 2005; Zeng-Treitler et al., 2007; Kunz &Osborne, 2010). The general assumption is that the use of specialized vocabulary hinders the comprehension of patients with lower reading skills, thus more "readable" texts are more comprehensible for those who have lower reading proficiency. However, this assumption is challenged by several scholars. For instance, Miller et al. (2007) argue that "traditional readability formulas examine syntactic features like sentence length and number of syllables, ignoring the target audience's grasp of the words themselves". Several studies indicate that standard readability formulas might not be of help when assessing the difficulty of medical texts. Leroy et al. (2008) found that readability differs by topic and source. They proposed metrics different from readability formulas and argued that these metrics were more precise than readability scores. They compared two documents in English for three groups of linguistic metrics and conducted a user study evaluating one of the differentiating metrics, i.e. the percentage of function words in a sentence. Their results showed that this percentage correlates significantly with the level of understanding as indicated by users but not with the readability formula levels. On the same line, Zheng & Yu (2017) found that the correlations of readability predictions and laypeople's perceptions were weak. Their study with English texts explored the relationship between several readability formulas and the laypeople's perceived difficulty on two genres of text: general health information and electronic health record (EHR) notes. Their findings suggested that "the readability formulas' predictions did not align with perceived difficulty in either text genre. The widely used readability formulas were highly correlated with each other but did not show adequate correlation with readers' perceived difficulty. Therefore, they were not appropriate to assess the readability of EHR notes."

The construction of *lay corpora and lay terminology extraction* is advanced for the French language. Several experiments have been carried out by Deléger et al. (2013) based on lay-specialized monolingual comparable corpora which were built using web documents belonging to specific genres from public websites in the medical domain. Grabar & Hamon (2014b) proposed an automatic method based on the morphological analysis of terms and on text mining for finding the paraphrases of technical terms in French. Their approach relies on the analysis of neoclassical medical compounds and for searching their non-technical paraphrases in corpora. Depending on the semantics of the terms, error rate of the extractions ranges between 0 and 59%. Antoine & Grabar (2017) focused on the acquisition of vocabulary by associating technical terms with layman expressions. They proposed exploiting the notion of "reformulation" through two methods: extraction of abbreviations and their extended forms, and of reformulations introduced by markers. Tchami & Grabar (2014) described a method for a contrastive automatic analysis of verbs in French medical corpora, based on the semantic annotation of the verbs'

nominal arguments. The corpora used are specialized in cardiology and distinguished according to their levels of expertise (high and low). The semantic annotation of these corpora was performed using existing medical terminology. The results suggest that the same verbs occurring in the two corpora show different specialization levels, which are indicated by the words with which they co-occur.

Lay terminology extraction methods for the English language were proposed by Elhadad and Sutaria, (2007) who mined a lexicon of medical terms and lay equivalents using abstracts of clinical studies and corresponding news stories written for a lay audience. Their collection is structured as a parallel corpus of documents for clinicians and for consumers. Zeng et al. (2007) explored several term identification methods for the English language, including collaborative human review and automated term recognition methods. The study identified 753 consumer terms and found the logistic regression model to be highly effective for lay term identification. Doing-Harris & Zeng-Treitler (2011) presented the CAU system which consisted of three main parts: a Web crawler and an HTML parser, a candidate term filter that utilizes natural language processing tools including term recognition methods, and a human review interface. In evaluation, the CAU system was applied to the health-related social network website PatientsLikeMe.com. The system's utility was assessed by comparing the candidate term list it generated to a list of valid terms manually extracted from the text of the crawled webpages. Soergel, Tse & Slaughter (2004) proposed an interpretive layer framework for helping consumers find, understand and use medical information. Seedorff et al. (2013) introduced the Mayo Consumer Health Vocabulary (MCV)—a taxonomy of approximately 5,000 consumer health terms and concepts—and developed text-mining techniques to expand its coverage by integrating disease concepts (from UMLS[26]) as well as non-genetic (from deCODEme[27]) and genetic (from GeneWikiPlus[28] and PharmGKB[29]) risk factors to diseases. Jiang and Yang (2013) used co-occurrence analysis to identify terms that co-occur frequently with a set of seed terms. A corpus containing 120,393 discussion messages was used as a dataset and co-occurrence analysis was used to extract the most related consumer expressions. The study presented in Vydiswaran et al. (2014) focused on the linguistic habits of consumers. In their study, the authors empirically evaluate the applicability of their approach using a large data sample consisting of MedLine abstracts as well as posts from a popular online health portal, the MedHelp forum. The "propensity of a term", which is a measure based on the ratio of frequency of occurrence, was used to differentiate lay terms from professional terms.

In Sweden, research on medical language is also strong. Kokkinakis (2006) described efforts to build a Swedish medical corpus, namely the MEDLEX Corpus, where generic named entity and terminology recognition for the detailed annotation of the corpus are combined. Kokkinakis & Gronostaj (2006) carried out a corpus-based, contrastive study of Swedish medical language focusing on the vocabulary used in two types of medical textual material: professional portals and web-based consumer sites within the domain of cardiovascular disorders. Linguistic, statistical and quantitatively based readability studies are considered in order to find the typical language-dependent and language independent characteristics. Heppin (2010) created a unique medical test collection for Information Retrieval to provide the possibility to assess the document relevance to a query according to two user groups, namely patients or physicians. The focus of Abrahamsson et al (2014) was on the simplification of one single genre, namely the medical journal genre. To this purpose, the authors used a subset of a collection built from the journal Svenska Läkartidningen, i.e. the Journal of the Swedish Medical association, that was created by Kokkinakis (2012). Another unique language resource is the Stockholm EPR (Electronic Patient Records) Corpus (Dalianis et al., 2009; and Dalianis et al., 2015), which comprises real data from more than two million patient records. Johansson & Rennes (2016) presented results from using two methods

to automatically extract Swedish synonyms from a corpus of easy-to-read texts. They used two methods, based on distributional semantic models (more specifically word2vec), one inspired by Lin et al. (2003) and the other by Kann and Rosell (2005). The methods were evaluated using an online survey, in which the perceived synonymy of word pairs, extracted by the methods, was graded from "Disagree" (1) to "Totally agree" (4). The results were promising and showed, for example, that the most common grade was "Sometimes" (3) for both methods, indicating that the methods found useful synonyms.

Previous research in *automatic supervised lay-specialized text classification* show that simple methods yield good performance.

As for English, Zheng et al. (2002) addressed the problem of filtering medical news articles for lay and expert audiences. They used two supervised machine learning techniques, Decision Trees and Naive Bayes, to automatically construct classifiers on the basis of a training set, in which news articles have been pre-classified by a medical expert and four other human readers. The goal is to classify the news articles into three groups: non-medical, medical intended for experts, and medical intended for other readers. While the general accuracy of the machine learning approach is around 78% (three classes), the accuracy of distinguishing non-medical articles from medical ones is shown to be approximately 92% (two classes). Miller et al. (2007) created a Naive Bayes classifier for three levels of increasing medical terminology specificity (consumer/patient, novice health learner, medical professional) with a lexicon generated from a representative medical corpus. 96% accuracy in classification was attained. The classifier was then applied to existing consumer health web pages, but only 4% of the pages were classified at a layperson level, regardless of the Flesch reading ease scores, while the remaining pages were at the level of medical professionals. This finding seems to indicate that consumer health web pages are often not using appropriate language for their target audience. In order to recommend health information with appropriate reading level to consumers, Support Vector Machine (SVM) is used to classify consumer health information into easy to read and reading level for the general public by Wang (2006). He used three feature sets: surface linguistic features, word difficulty features, unigrams and their combinations were compared in terms of classification accuracy. Unigram features alone reached an accuracy of 80.71%, and the combination of three feature sets was the most effective in classification with an accuracy of 84.06%. They are significantly better than surface linguistic features, word difficulty features and their combination. Miller & Leroy (2008) created a system that dynamically generates a health topics overview for consumer health web pages that organizes the information into four consumer-preferred categories while displaying topic prevalence through visualization. The system accesses both a consumer health vocabulary and the Unified Medical Language System (UMLS). Overall, precision is 82%, recall is 75%, and F-score is 78%, and precision between sites did not significantly differ.

Multilingual approaches to lay vs. specialized text classification exhibit interesting findings. For example, Porrat et al. (2006) proposed a pipelined system for the automatic classification of medical documents according to their language (English, Spanish and German) and their target user group (medical experts vs. health care consumers). They used a simple n-gram based categorization model and presented promising experimental results for both classification tasks. Seljan et al., (2014)'s research to understand the role of terminology in online resources, was conducted on English and Croatian manuals and Croatian online texts and divided into three interrelated parts: i) comparison of professional and popular terminology use; ii) evaluation of automatic statistically-based terminology extraction on English and Croatian texts; and iii) comparison and evaluation of extracted terminology performed on an English manual using statistical and hybrid approaches. Extracted terminology candidates were evaluated by comparison with three types of reference lists: a list created by a professional medical person, a list of

highly professional vocabulary contained in MeSH and a list created by non-medical persons, made as intersection of 15 lists.

A set of experiments on multilingual lay-specialized medical corpora are presented in Borin et al. (2007). They investigated readability in English, Swedish, Japanese and Russian. They explored variations in readability, lexicon and lexical-semantic relations, grammar, semantic and pragmatics, as well as layout and typography. On the basis of the findings, the authors proposed a set of recommendations per language for adapting expert clinical documents for patients.

On the cross-lingual side, Grabar et al. (2007) put forward the hypothesis that discrimination between lay vs. specialized documents can be done using a small number of features and that such features can be language- and domain-independent. The features used were acquired from a source corpus (Russian language, diabetes topic) and then tested on target (French language, pneumology topic) and source corpora. These cross-language features showed 90% precision and 93% recall with non-expert documents in the source language; and 85% precision and 74% recall with expert documents in the target language.

The medical text collections briefly mentioned above are important language resources, but their construction and usage seem to be contingent to specific experiments, rather than designed for a long-term deployment and continuous enhancement. For this reason, we propose a new kind of design for a domain-specific corpus with the intent to be re-used, easily updatable and hopefully long-lasting.

In the experiments presented in this chapter, we do not compare our results with readability scores. It would be interesting to compare the different readability levels of web documents on chronic diseases. Stable sets of readability assessment features exist both for English and Swedish (i.e. the language included in eCare Sv_En_03). Unfortunately, texts crawled from the web are noisy. For instance, texts may contain informal language (e.g. sv: "nå'n annan som hatar utredningen?" English: "somebody else who hates the investigation"), and unpredictable combinations of English words (e.g. "therapycounseling") are numerous. This means that the automatic extraction of readability assessment features from eCare Sv_En_03 would imply a regularization of the corpus that we have not planned for yet. At this stage, we focus on how to leverage on noisy texts rather than on how to regularize them.

Simple methods based on distributional semantics and automatic lay-specialized text classification are promising and easy to implement. For this reason, we continue along this line (Section 5).

RETHINKING WEB CORPORA: THE WORK-IN-PROGRESS DESIGN

In this section, we describe the *current* implementation of the design of a work-in-progress web corpus. We stress the word "current" because it is our ambition to explore several different approaches that can all be conflated into the same design of a corpus conceived as extensible, updatable and open-ended. The experiments described in this section are based on a version of the corpus that **is not** "unfinished" or "incomplete". Rather, it must be seen as the first iteration of an incremental strategy. The inspiration for this approach comes from the Agile Methodologies used in software development and project management, where the implementation of a plan is based on cycled iterations that ensure a seamless incremental progress. *Agile* is a process that "advocates adaptive planning, evolutionary development, early delivery, and continual improvement, and it encourages rapid and flexible response to change"[30]. This source of inspiration provided a framework for the idea of the work-in-progress corpus. The construction of a corpus is based on iteration which ensures a continual improvement. Since the agile approach is based

on incremental deliveries, each successive version of the corpus is usable, and each version builds upon the previous one. Such a process is adaptive to changes, flexible and open-ended.

For the bootstrapping of *eCare_Sv_En_03*, we followed the general approach initiated by Baroni & Bernardini (2004) and widely used all over the world.

Starting Off: BootCaT-ing the Swedish Corpus About Chronic Diseases

eCare_Sv_01 (see also Santini et al., 2017) is a small text collection bootstrapped from the web. It contains 801 web documents that have been labelled as *lay* or *specialized* by two annotators. In the following subsections, we describe its construction and the actual corpus.

The Seeds

We started off with approximately 1300 term seeds designating chronic diseases in the Swedish SNOMED CT. A qualitative linguistic analysis of the term seeds revealed a wide range of variation as for number of words and syntactic complexity. For instance, multiword terms (n-grams) are much more frequent than single-word terms (unigrams).

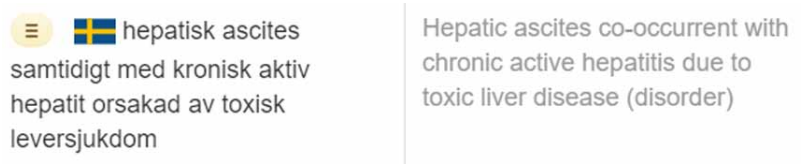
We counted 13 unigrams (i.e. one-word terms) (see Table 1), 215 bigrams (i.e. two-words terms), and the rest of the seeds were characterized by specialized terms and complex syntax, such as: "kronisk inkomplett tetraplegi orsakad av ryggmärgsskada mellan femte och sjunde halskotan" (English: "Chronic incomplete quadriplegia due to spinal cord lesion between fifth and seventh cervical vertebra"). Another example is shown in Figure 2.

To bootstrap this version of the corpus, we used unigrams and bigrams only. This decision was based on the assumptions that (1) unigram- and bigram-terms are more findable on the web than syntactically complex keyword seeds, and (2) complex multiword terms are less likely to have a lay synonym or paraphrase. It should be noticed however that Swedish is a compound language where several words are united in one single graphical unit, thus the distinction between unigrams and bigrams is sometimes blurred.

Table 1. Unigram seeds

Seeds (Swedish)	Translation (English)	SCTID
ansiktstics	Facial tic disorder	230335009
bukangina	Abdominal angina	241154007
chalcosis	Chalcosis	46623005
fluoros	Fluorosis	244183009
kromoblastomykos	Chromoblastomycosis	187079000
lipoidnefros	Minimal change disease	44785005
lungemfysem	Pulmonary emphysema	87433001
mycetom	Mycetoma	410039003
ozena	Ozena	69646003
polyserosit	Polyserositis	123598000
postkardiotomi-syndrom	Postcardiotomy syndrome	78643003
Swimmingpool-dermatit	Swimming pool dermatitis	277784005
trumhinneatelektas	Tympanic atelectasis	232258001

Figure 2. Example of long medical term as it is displayed in the SNOMED CT browser



Pre-Processing and Download

Using regular search engines (like Google, Yahoo or Bing) and term seeds (as queries) to build a corpus is handy, but it also has some caveats that depend on the design or distortion of the underlying search engine (Wong et al., 2011). These caveats affect the content of web corpora since it might happen that irrelevant documents are included in the collection, especially when searching for very specialized terms. For the construction of *ecare_Sv_01,* we decided to use seeds in the following way to have a better insight of the content of the corpus that was going to be built. Each seed was used as a search keyword in *Google.se*, i.e. Google web domain for Sweden. The searches were carried out from within Sweden (namely Stockholm and Örebro). Each of the preselected SNOMED CT terms were used individually, i.e. one seed term per query. This means that we launched 228 queries. For each seed/query, Google returned a number of hits. We limited our analysis to hits on the first page (we extended the visualization of the results to 20 hits per page). We manually opened each snippet to have an idea of the type of web documents that were retrieved. For each search lap, several documents were irrelevant (presumably as an unwanted effect of query expansion) and several were duplicated. 74 keyword seeds were discarded because the retrieved documents were irrelevant or contained passages not written in Swedish.

Unsurprisingly, we also noticed that the number of retrieved pages depends on how common a disease is for web users. For instance, "ansiktstics" (English: "facial tics") had many hits, while "chalcosis" (en: "chalcosis") had very few. As a rule of thumb, we decided to select a maximum of 20 documents for the most common illnesses, and as many as we could for rarer diseases. This observation about common and rare illnesses, is merely based on the number of hits retrieved by the search engine. We do not rely on medical statistics, because the situation may change any time. For example, for some reasons that we are unable to foresee now, an illness like "chalcosis" can become widespread in a couple of years and the web will be inundate by documents about this illness. This is just an example of why a corpus of this kind should be extendible and flexible.

After this preprocessing phase, we applied BootCat[31] (Baroni & Bernardini, 2004) using the advanced settings (i.e. *url seeds*) to create the web corpus.

We handed out documents downloaded with BootCat to two native Swedish speakers (both academics), one lay person (i.e. not working in the medical field) and one specialized person (working with medical-related subjects). They proceeded with the annotation by applying a lay or specialized label to each text in the corpus.

eCare_sv_01 in a Nutshell

eCare_Sv_01 was bootstrapped using 228 terms indicating chronic diseases, namely 13 unigrams and 215 bigrams. The number of unigrams is much lower than multi-word n-grams. This seems to indicate that medical language prefers multiword expressions also in Swedish, which is a compound language.

After the preprocessing, 843 urls (112 for unigrams and 731 for bigrams) were factored out and used as *url-seeds* in BootCat. Some of the urls were automatically discarded by BootCat (e.g. bilingual documents were not included) and some downloaded documents were empty. Finally, 801 documents were successfully bootcat-ed with 155 seeds[32]. Table 2 shows the corpus statistics.

Both annotators pointed out that the quality of the writing was poor in some documents, mainly because they had been machine translated, and not written by humans. Some of the web documents explicitly stated "Översatt från engelska av Microsoft" (English: Translated from English by Microsoft). Out of 801 web documents, 339 have received comments by the lay annotator, e.g. "Machine Translated" or "it is about animals and not about humans", and the expert annotator flagged 23 documents as medically "irrelevant". By 'irrelevant', the annotators meant that these documents contained the seed terms, but the genre was a schedule or a conference program, or the described illnesses was on animals rather than humans.

Essentially, we can observe that the corpus is noisy. The annotators' comments help us understand the different types of noise and emphasize a crucial issue that is underexplored in corpus- and computational linguistics, i.e. the reliability and the quality of corpora bootstrapped from the web. The automatic discrimination of "good" documents from "bad" ones is an important problem, especially in sensitive domains like the medical or legal domains. In the medical domain, recent research shows that the reliability, readability and quality of patient-oriented websites are still open issues for the scientific community. It has been pointed out that this kind of resources are easily accessed by patients, but they might contain information that are less rigorous than the information provided by scientific literature or healthcare practitioner websites and can inconveniently cause "misinformation" in patients (Soobrah and Clark, 2012; Küçükdurmaz et al., 2015). These issues will certainly be explored more extensively in future research. However, in the experiments that we present in this chapter, we took another perspective and investigated to what extent lay-specialized text classification is robust to a number of disturbing factors. Since cleaning or refining a corpus might be prohibitive in many projects, our challenge is to see whether noisy corpora can be used in Language Technology without affecting the performance of LT applications. For this reason, the noisy documents have been left in the corpus, but they are flagged so they can be easily included or bypassed, according to the purpose of the project at hand. Other types of research that can benefit from the inclusion of "disturbing" texts in the corpus include the automatic

Table 2. Corpus statistics

	# initial seeds	# retrieved seeds	# bootcat-ed urls	# urls per seed (mean)	# urls per seed (median)	# urls per seed (st dev)	# words
Unigrams	13	13	112	8.61	9.3	3.37	91 118
Bigrams	215	142	689	4.85	4	3.16	618 491
Total	228	155	801	5.16	5	3.35	709 609

analysis of MT "translationese" (Volansky et al., 2015) and the automatic quality assessment of text writing[33].

Proposed Corpus and Text Metadata

Corpus and text metadata are important elements of the whole corpus design since they characterize the corpus and allow us to extract "virtual sub-corpora" matching some criteria.

In Figure 3, the pseudo-annotation by sublanguage is shown. The tags <sublangs> and </sublangs> surround information about how the lay annotator and the expert annotator labelled each text in the corpus[34]. In the example in Figure 3, the lay- and the expert annotator DO NOT agree about the sublanguage of a text in the corpus.

Having the corpus annotated in this way is convenient not only for the linguistic/textual descriptions of the individual texts, but also to quickly extract datasets for text classification according to the sublanguage textual dimension. As a matter of fact, the datasets that we use for the experiments in Section 8 are extracted from this annotated corpus using a simple R script[35].

Human Annotation and Interrater Agreement

The annotation of documents in the corpus as being *lay* or *specialized* was carried out by two native speakers who participate in the project. They are both academics but operating in different research fields: namely the lay annotator works in Language Technology while the expert annotator works in Health Informatics.

The purpose of the manual annotation is to prove that there is no need to decide beforehand the sources of lay and specialized documents but crawling the web indistinctly will return a mixture of lay and specialized comments. The annotators were to follow their spontaneous linguistic instinct for assessing the language and no training was provided. The "lay" annotator has high education, but no familiarity with medical terminology, either personally or professionally. The expert annotator works with medical terminology. At this stage, we did not involve the potential users of the final system because we first wanted to observe to what extent medical expertise affects the agreement on two highly-educated persons. It is normal that the judgement on specialization/technicality of documents varies with the expertise of the annotators. Other factors like the education level can strongly influence the assessment.

We measured the inter-annotator agreement in two steps, first on approximately 1/3 of the documents, and then on the whole corpus, to see if the size of the corpus creates a bias in the agreement. The sample of 1/3 was random. No consensus step was taken to reach a final agreement between the two annotators,

Figure 3. Metadata describing the sublanguage and the annotators' expertise

```
<?xml version="1.0" encoding="UTF-8"?>
<document>
    <sublangs>
        <sublang>
            <lay-annotator-evaluation>specialised</lay-annotator-evaluation>
            <specialised_annotator-evaluation>lay</specialised_annotator-evaluation>
        </sublang>
    </sublangs>
</document>
```

since at this stage we are not focusing on that. As shown in experiment 3 in Section 5.2, we argue that there is no need to any further step because a regular classifier can establish the consensus itself. This streamlines and accelerates the construction of supervised LT applications.

To have an idea of the agreement between a lay annotator and an expert annotator, we carried out two interrater-agreement calculations. First, we measured the agreement on a random sample (348 out of 801 documents), then on the whole corpus (801 documents) and observed whether the two calculations returned similar coefficients or not. The rationale of this decision was to determine whether the size of the corpus to be annotated plays a role in the agreement.

Several inter-rater agreement measures exist (Artstein and Poesio, 2008). All the inter-rater agreement measures have their strong points and their drawbacks and the use of one over the other depends on the data, the task and the situation. In our case, we wish to measure to what extent two members belonging to two different user groups (i.e. the lay and the expert) spontaneously agree when assessing the "domain-specificity" of medical language. Our expectation is that a lay person tends to label as "specialized" a larger number of medical documents than an expert person, who, conversely, tends to see as "lay" many documents that laypeople would consider to be "specialized". In order to test this assumption, we measured the inter-rater agreement by using percentage (i.e. the proportion of agreed upon documents in relation to the whole without chance correction), the classic unweighted Cohen's *kappa* (Cohen 1960) and Krippendorff's *alpha* (Krippendorff, 1980) to get a straightforward indication of the raters' tendencies. Cohen's *kappa* assumes independence of the two coders and is based on the assumption that "if coders were operating by chance alone, we would get a separate distribution for each coder" (Artstein and Poesio, 2008). This assumption intuitively fits our expectations. Krippendorff's *alpha* is similar to Cohen's *kappa*, but it also takes into account the extent and the degree of disagreement between raters (Artstein and Poesio, 2008).

Table 3 shows the breakdown of the inter-rater agreement on the sample, while Table 4 shows the overall inter-rater agreement on the whole corpus[36]. The breakdown of Table 3 reveals that, interestingly, annotators tend to disagree more on documents harvested with unigrams (Row 1), while they agree more on documents harvested with bigrams (Row 2).

Overall, Table 3 shows that both *kappa* and *alpha* coefficients are approx. 0.5, and both these values indicate a "moderate" agreement according to the magnitude scale for *kappa* (Sim and Wright, 2005), and the *alpha* range (Krippendorff, 2011).

Table 3. Breakdown: inter-rater agreement on the sample (348 documents)

# documents	Percentage	Cohen's *Kappa*	Krippendorff's *Alpha*
112 (unigram seeds)	75.9%	0.52	0.51
238 (bigram seeds)	82.2%	0.60	0.60
348 (total)	80.2%	0.57	0.57

Table 4. Interrater agreement on the whole corpus (801 documents)

Whole Corpus	Percentage	Cohen's *Kappa*	Krippendorff's *Alpha*
801 documents	78.8%	0.51	0.51

The interrater coefficients computed for the whole corpus (see Table 4) are in line with the coefficients calculated on the sample, although the coefficients for the whole corpus are slightly lower than those for the sample. However, all coefficients confirm that the agreement between the lay and the expert annotator is *moderate*.

Table 5 shows the distribution of labels per annotator. We can observe that the lay annotator tends to apply fewer lay labels than the expert annotator, who conversely perceive as "lay" more documents than the lay annotator. Interestingly, the expert annotator is also much more selective than the lay annotator and flags 23 documents as medically "irrelevant".

Discussion

These results seem to endorse our hypothesis that there exists a *user group bias*, which indicates that the annotation may be biased by the annotator's domain expertise. If we contextualize the results, this finding means that patients (who usually have a "lay" perspective) tend to perceive many documents as "specialized", while physicians would assess these documents simply as "normal". This has a linguistic implication that might affect certain LT applications in the eHealth field, and we encourage more in-depth investigation about this topic in the future.

Intrinsic Evaluation: Supervised Lay-Specialized Text Classification

In this section, we validate the reliability of the corpus. We investigate how reliable a small and noisy corpus is for LT applications. We present three experiments based on lay-specialized text classification. We apply fully-supervised machine learning methods to explore how well supervised algorithms learn the labels applied by the lay annotator (Experiment 1 and Experiment 2), and how sensitive a LT task, such as text classification, is to the disagreement between the lay annotator and expert annotator (Experiment 3). Technically speaking, Experiment 1 focuses on corpus scalability, and help us understand whether the size of the corpus has an impact on the classification results. In Experiment 2, we explore to what extent lay-specialized text classification is affected by noisy documents. Both in Experiment 1 and 2, we use the labels applied by the lay annotator because we focus on the patients' perspective on language, which is presumably *lay*. In Experiment 3, we investigate whether text classification is sensitive to the user group bias, i.e. the different ways of labelling documents.

Corpus scalability refers to the capability to identify the critical mass of corpus size that ensures a satisfactory performance of a specific algorithm for a specific task. With supervised machine-learning algorithms, the normal assumption is that classification performance improves when more data is available. However, since we are after a minimalist approach, we wish to identify the minimum data requirement to get a "good enough" performance.

Table 5. Number of labels by annotator

	Lay labels	Specialized Labels	NA	Total
Lay Annotator	246	555	0	801
Expert Annotator	279	499	23	801

In these experiments, we relied on the Weka Machine Learning Workbench, Explorer and Experimenter interfaces (Witten et al., 2016).

Quick-and-Dirty: Features and Noisy Texts

The first question to answer when performing lay-specialized text classification is: which features are most appropriate to represent lay and specialized medical sublanguages? We decided to take a Bag-of-Words approach. Only two attributes were declared, namely *the textual content of the document* defined as "string", and the *sublanguage label* (either "lay" or "specialized") defined as "nominal".

Experiment 1: Lay-Specialized Text Classification and Corpus Scalability

The rationale of this experimental setting is to observe whether and to what extent the performance of the classifiers deteriorates when increasing the corpus size.

We converted four subsets of the whole corpus into four datasets. The first dataset contains 156 documents; the second one 220 documents; the third one 337 documents; the fourth datasets includes the whole corpus and contains 801 documents. The datasets were created randomly and they contain some overlapping data since we wish to simulate the progressive expansion of the corpus over time by appending more documents to the original corpus. Since we did not know in advance which type of machine learning modelling would be more suitable for this kind of data, we applied three standard algorithms that have very different inductive biases, namely *Decision Trees*, *Naive Bayes* and SVM. We took a bag-of-word approach, and construct vectors from strings using a filter (see below).

We used Weka's implementations of the algorithms, namely J*48*, *Naive Bayes* and *SMO*. All the algorithms were run with standard parameters.

We ran each of the algorithms via a metaclassifier[37]. Recent developments in computational learning theory have led to methods that enhance the performance or extend the capabilities of these basic learning schemes. Those learning schemes have been called "meta-learning schemes" or "meta-classifiers" because they operate on the output of other learners or filters.

We selected in turn each of the pre-decided classifiers together with the *StringToWordVector* filter (standard parameters). StringTOWordVector is the filter class in Weka which filters strings into n-grams. Besides just tokenizing, it also provides other functionalities like removing stopwords, weighting words with TFIDF[38], output word count rather than just indicating if a word is present or not, pruning rate, stemming, lowercase conversion of words, etc. Basically, it provides basic functionalities which helps us to fine-tune the training set according to requirements before training.

We applied 10-fold-crossvalidation. Results are shown in Table 6 (values have been truncated to two decimal places).

For the first dataset (156 documents), J48 seems to be less suitable than Naive Bayes and SMO. J48's k statistic is low, indicating that most of the corrected classifications happen by chance. The confusion matrix for J48 shows that lay texts are quite confusing for this classifier (only 48 TP vs 35 misclassified cases), while specialized texts are more clearly set apart (110 TP[39] vs 27 misclassifications). Naive Bayes and SMO do a better job on this dataset: their averaged ROC area[40] values are much higher than 0.5 (0.5 would mean that a classifier is random).

Table 6. Experiment 1: performance on datasets of different sizes

DATASET 1: 156 DOCUMENTS

	k	Acc.	Avg. P	Avg. R	Avg. F	ROC A.	Avg TP	Avg FP
J48	0.14	62.8	0.62	0.62	0.62	0.63	0.62	0.42
NB	0.46	75.6	0.77	0.75	0.76	0.80	0.75	0.26
SMO	0.43	75.6	0.75	0.75	0.75	0.71	0.75	0.32

DATASET 2: 220 DOCUMENTS

	k	Acc.	Avg. P	Avg. R	Avg. F	ROC A.	Avg TP	Avg FP
J48	0.38	71.8	0.71	0.71	0.71	0.69	0.71	0.33
NB	0.45	72.7	0.75	0.72	0.73	0.78	0.72	0.25
SMO	0.36	70.9	0.70	0.70	0.70	0.67	0.70	0.35

DATASET 3: 337 DOCUMENTS

	k	Acc.	Avg. P	Avg. R	Avg. F	ROC A.	Avg TP	Avg FP
J48	0.38	72.1	0.71	0.72	0.71	0.71	0.72	0.33
NB	0.46	73.5	0.76	0.73	0.74	0.80	0.73	0.23
SMO	0.50	77.1	0.77	0.77	0.77	0.75	0.77	0.27

DATASET 4: ALL 801 DOCUMENTS

	k	Acc.	Avg. P	Avg. R	Avg. F	ROC A.	Avg TP	Avg FP
J48	0.38	74.15	0.74	0.74	0.74	0.66	0.74	0.37
NB	0.45	73.9	0.78	0.73	0.74	0.83	0.73	0.23
SMO	0.49	78.6	0.78	0.78	0.78	0.74	0.78	0.29

On the second dataset (220 documents), J48's performance values are equivalent to Naive Bayes's and SMO's. On the third dataset (337 documents), SMO shows better figures. The performance on the fourth dataset is similar to the third dataset.

In order to compare the performance of the three classifiers on the four datasets, we applied the Corrected Paired T-Test (two tailed)[41] provided by Weka's Experimenter interface. Statistical significance was measured on the results of the three classifiers per dataset, and on the performance of each classifier for the four datasets. Statistical significance was measured at significance level of $p < 0.001$ on the weighed averaged F-measure. The test did not detect any statistically significant variation due to the sample[42]. We interpret these findings as a sign of stability since results show the robustness of the models to corpus scalability issues. This experiment supports our claim that a corpus can be extended without causing any deterioration of the performance of LT applications.

Experiment 2: Lay-Specialized Text Classification With and Without Noise

In Experiment 2, we explored whether there exists a performance gap between text classification models trained on a collection containing noisy documents and text classification models trained on a collection containing only noise-less documents.

Results are shown in Table 7. In order to compare the two sets of results, we measured the performance of the same algorithm on the two datasets. As in Experiment 1, statistical significance was measured at significance level of $p < 0.001$ on the weighted averaged F-measure. The test did not detect any statistically significant variation. We interpret these findings as a sign of resistance to noise in the lay-specialized text classification task. This experiment supports our claim that noise does not always negatively affect classification performance.

Table 7. Datasets with and without noise

DATASET 4: ALL 801 DOCUMENTS

	k	Acc.	Avg. P	Avg. R	Avg. F	ROC A.	Avg TP	Avg FP
J48	0.38	74.15	0.74	0.74	0.74	0.66	0.74	0.37
NB	0.45	73.9	0.78	0.73	0.74	0.83	0.73	0.23
SMO	0.49	78.6	0.78	0.78	0.78	0.74	0.78	0.29

DATASET WITHOUT NOISY TEXTS: 462 DOCUMENTS

	k	Acc.	Avg. P	Avg. R	Avg. F	ROC A.	Avg TP	Avg FP
J48	0.36	72.29	0.72	0.72	0.72	0.69	0.72	0.35
NB	0.57	79.22	0.82	0.79	0,79	0.88	0,79	0.16
SMO	0.57	80.95	0.81	0.81	0.81	0.78	0.81	0.23

Experiment 3: Lay vs. Specialized Annotation

The expertise of the annotators does not affect the performance. Essentially, there is no need to reach a high agreement between raters because a moderate agreement caused by different expertise is enough to ensure similar classification results. In practical terms, this means, that we do not to worry so much about the expertise or non-expertise of the annotators in a case like this. This will result in a more streamlined corpus annotation.

In this experiment, we compared the performance of the SMO algorithm on the documents labelled by the lay annotator (801 documents) against the documents labelled by the expert annotator (778 documents. The expert annotator left 23 documents unlabelled since they were considered as medically "irrelevant"). Results on the two datasets (see Table 8) show that the performance is basically the same, since no statistical significance variation has been detected. Apparently, the expertise of the annotators, although their agreement is moderate, does not affect automatic text classification. (To confirm this finding, we are currently cross-testing these classifiers by training a classifier on the corpus labelled by the expert annotator and testing this classification model on the corpus labelled by the lay annotator, and vice versa.)

Table 8. Text classification performance: comparing Lay vs Specialized Annotation

	k	Acc.	Avg. P	Avg. R	Avg. F	ROC A.	Avg. TP	Avg. FP
SMO	0.49	78.6	0.78	0.78	0.78	0.74	0.78	0.29
801 documents labelled by the lay annotator								

	k	Acc.	Avg. P	Avg. R	Avg. F	ROC A.	Avg. TP	Avg. FP
SMO	0.54	79.5	0,79	0,79	0,79	0,77	0,79	0,25
778 documents labelled by the expert annotator								

Discussion

Results of Experiment 1 are in line with previous research. The big advantage with our approach is that we can achieve a competing performance with a noisy corpus and bag-of-words features, certainly the easiest ones to extract automatically.

Experimental results show that lay-specialized classification performance is good (averaged F-measure is above 0.70 in most cases) and stable across classifiers and across datasets of different sizes.

In our view, these results are promising for several reasons. The first reason is *corpus scalability*: Experiment 1 shows that results are essentially equivalent across samples of different sizes since we observe no statistically significant degeneration in the performance when scaling up the size of the corpus. This is reassuring: we can imagine a scenario where we design a dynamic and extensible corpus whose size can be increased over time, and this will not affect the expectation of efficiency and reliability of LT applications when scaling up. We expect SVM to perform better than the other classifiers because this algorithm has been designed to handle large feature sets (which is the case here). Decision Trees (DT) might be disturbed by the presence of many features to build the tree and Naive Bayes (NB) can be negatively affected by the independence assumption. We observe that while the performance of DT is consistently lower than the other two classifiers, the Naive Bayes performs slightly better when the corpus size is smaller, while SVM overperforms NB when the corpus increases. Interestingly, the performance of SVM changes negligibly from 337 to 801 documents, which indicates that a size of about 350 documents can be the "critical mass" to get a reliable performance in this task.

The second reason is resilience to noise: removing noisy documents from a corpus can be prohibitive in some contexts. Arguably, not all LT applications require high quality texts to ensure a good performance and reliable results, as we have shown in Experiment 2.

Another reason is highlighted by Experiment 3 which shows that the disagreement between annotators is flattened out and does not affect the performance of certain LT applications. Essentially, this means that we do not have to worry too much about the agreement or disagreement of different annotators. This finding might have a positive practical consequence since it might contribute to speed up and streamline the construction of certain LT applications.

Additionally, the experiments show that for this kind of text classification, there is no need of linguistically rich features to achieve good results. The word to vector conversion is a ready-made and standard approach that is available in several packages and programming languages, not only in the Weka workbench. It can be easily applied and optimized to achieve better results with little effort and time gain. Again, the process is streamlined.

Whatever the size of the corpus the approach returns good results: the supervised algorithms can detect features which permit to make the distinction between the two categories without any problem. These results indicate that it can be safe to use bag-of-words filtered features and a small corpus with SVM to get a good enough performance.

Arguably these findings are important to streamline automatic lay-specialized text classification, an useful task for document filtering and information retrieval.

The Nitty-Gritty of a Distributional Thesaurus

Term extraction methods based on distributional semantics are very popular and effective (for a thorough evaluation of the offspring of distributional semantic modelling, see Baroni et al., 2014). Meaningful results are returned when the underlying corpora are large, especially when using the most advanced models based on word embeddings.

By contrast, as argued above, our challenge is to find the most effective approaches with the smallest possible corpus. In this experiment, we use a slightly extended version of eCare_Sv_01, known as eCare_Sv_01+, to test the robustness of a simple distributional approach for the extraction of words related to 15 medical terms indicating chronic illnesses and we evaluate the results using an online survey. This approach can be useful to quickly extract synonyms or related terms from a small domain-specific document collection and to enrich existing resources with additional words and to create new resources. Collections of synonyms and of words related to specific terms are useful in LT applications such as, indexing, information extraction, text simplification, question and answering, synonyms expansion, or fine-grained domain-specific ontologies or medical dictionaries.

Expanding the Corpus: eCare_Sv_01+

For expanding the corpus, the software BootCat was used. The corpus was expanded by adding documents related to 10 SNOMED CT medical terms indicating to common ailments (we took inspiration from the 1177 Vårdguiden portal[43]) that were not included in the original list of 155 seed terms. The following 10 term seeds were used to expand the corpus: "*anemi, bronkit, cystit, depression, dermatit, eksem, hepatit, hypotoni, mastit, urtikaria.*" The rationale of such a choice was to foster possible distributional word associations between common diseases and targeted terms (listed below) since a target group of the *E-care@home* project is multimorbid patients.

The search was limited to only .se domains and to sv.wikipedia.org, to avoid foreign and machine translated texts. Furthermore, the domain 1177.se was excluded from the search, to enable evaluation against this site at a later stage. After the corpus was downloaded and cleaned, it was smoothly concatenated with the original corpus, which was increased by 174 new documents. This first corpus expansion proved to be unproblematic (for further details, see Cederblad, 2018).

Methodology

It was decided not to create a full co-occurrence matrix, in order to streamline and accelerate the process. Only context windows were extracted for the desired target terms. For this extraction, a built-in function in the "quanteda" R package, namely KWIC (Key Word In Context), was used. This function provides the possibility to input text, keyword, and window size, getting the desired windows as output.

The 15 target terms chosen for the evaluation are nouns selected from the list of 155 term seeds, namely "*artrit, bronkit, depression, dermatit, emfysem, faryngit, hemicrania, hyperglykemi, hypotoni, pneumoni, prostatit, rinit, schizofreni, sinuit, tonsillit*". For this experiment, all non-nouns were filtered out (the Stagger morphological tagger was used for this task, Östling, 2013) based on the assumption that normally synonyms and related words always belong to the same part-of-speech.

To find synonyms or semantically related words, the textstat_simil function of the Quanteda R package was used to calculate a variety of different similarity measures.

The similarity metrics compared were cosine, Jaccard, eJaccard, Dice, and eDice. For every synonym, the method received two points, and for every related word it received one point. All words were evaluated against 1177.se, to know whether they were synonyms or related. There were, in total, 100 words being evaluated for every method. Results are shown in Table 9.

The settings that have been used for the final results are the following:

- Window size: 10
- Similarity metric: Dice
- Weighting scheme: TFIDF
- Number of output words: 5

In Figure 4, all the steps in extracting the candidate related terms can be seen.

Although two words may have a high statistical similarity, it is not certain that they, in fact, are synonyms or even semantically related. To check whether two words are synonyms or related, a questionnaire was created. The questionnaire was structured in a way similar to that of Kann and Rosell (2005).

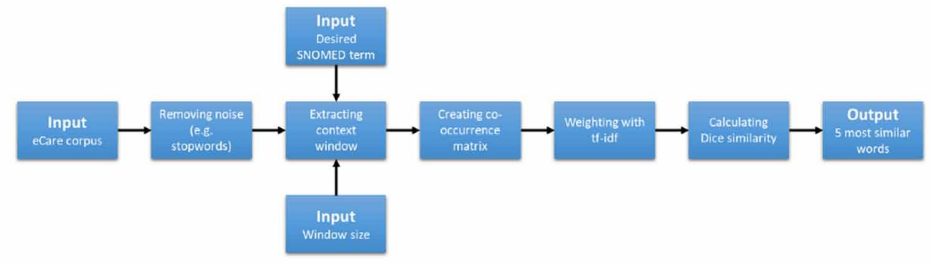
The words in the questionnaire were the output of the system described above. For each of the 15 SNOMED CT terms, five related words were considered reasonable for evaluation. There was also a need for controlling whether the users were guessing and answering without knowing the meaning of the words being evaluated. A participant who consistently answers that the control word has a similar meaning to the word being evaluated might not be reliable. Therefore, a randomly chosen word from the corpus was assigned to each set of five words. It was decided that each rating should have an explicit definition, to make it easier for the participants. Thus, the definitions agreed upon was as follows:

1. The words have entirely different meanings (ex: nästäppa & ryggbesvär, en: nasal congestion & back problem)

Table 9. The comparison of similarity measures

Cosine	Jaccard	eJaccard	Dice	eDice
58	79	67	80	68

Figure 4. A description of the procedure of extracting the candidate related terms



2. The words can be related to each other (ex: nästäppa & symptom; en: nasal congestion & symptom)
3. The words are related and are often used together (ex: nästäppa & förkylning; en: nasal congestion & cold)
4. The words are strongly related and can sometimes replace each other (ex: nästäppa & snuva; en: nasal congestion & rhinitis)
5. The words are synonyms and often replaces each other (ex: nästäppa & nasalobstruktion; en: nasal congestion & nasal obstruction)

The questionnaire was distributed to the participants through an URL published in a Facebook group for nurses in Sweden providing also the possibility to submit the answers through a smartphone. The questionnaire was accessible for seven days.

The SNOMED CT terms chosen for evaluation were considered too specialized for lay people to answer. It has to be stressed that the purpose of the questionnaire is not to find out whether people know the meaning of the terms. Rather, its purpose is to evaluate whether the words extracted by the distributional systems were related to the target terms. This requires some basic knowledge of the medical domain. For this reason, the participants were chosen with respect to their putative medical knowledge. That is, people with occupations exposing them to medical terminology were considered suitable. A control question was placed at the beginning of the questionnaire, to check whether the participants were or had been working in healthcare, or if they were studying or had studied a healthcare related subject. Since the questionnaire was published in a Facebook group for nurses in Sweden, it is reasonable to assume that most of the participants were nurses, although this was not further investigated. The total number of participants was 239, out of which 16 had to be excluded for not having completed the questionnaire.

Knowing the reliability of the raters (i.e. participants in the evaluation) is crucial for comparing their answers with the system output's Dice similarity. For calculating inter-rater reliability, Fleiss Kappa was used. Fleiss kappa is used for measuring the agreement of several raters (Fleiss, 1971). Landis and Koch (1977) propose a way of interpreting Kappa values. The proposal is as in Table 10. It is worth noting that this proposal contains a large portion of arbitrariness and in no way provides exact definitions. Despite this, it can be a useful tool when discussing the results. The agreement among the participants, calculated using Fleiss kappa, was k=0.28. According to the proposed interpretation in Table 10, this is a fair agreement.

The system output for all the 15 SNOMED CT terms chosen for evaluation and their Dice similarity can be seen in Table 11, with the candidate related term in the left column and its corresponding Dice similarity in the right column. The similarity calculated in R was, in general, rather low. As much as

Table 10. Landis and Koch (1977) proposal of Kappa interpretation

Kappa Statistic	Strength of Agreement
< 0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

91% of the words had a Dice similarity that was less than 0.50. The word with the highest similarity was the word pair "hemicrania-continua." The lowest found similarity was 0.17, and three word-pairs got a similarity this low.

There is a difference in how much text the corpus contains for each SNOMED CT term. When extracting the context window for each word, this difference becomes salient and is shown in Table 12. There was a difference of agreement among the categories, shown in Table 13, where the highest agreement was in the lowest rating category. The category with the lowest agreement was the one stating that the words were synonyms.

The most similar words to every SNOMED CT term, based on their average user rating, was extracted and is presented in Table 14 together with the corresponding Dice similarity.

The average rating of all the words, with the control words excluded, was 2.48 (St. Dev. = 0.80).

Table 15 shows the ten word-pairs with the highest number of participants answering "I don't know." Out of these word pairs, five belong to the SNOMED CT term "hemicrania."

Table 11. System output

artrit		bronkit		depression		dermatit	
reumatoid	0.56	luftväg	0.35	liv	0.28	hud	0.52
psoriasisartrit	0.21	kol	0.28	person	0.22	eksem	0.43
knä	0.19	lunga	0.27	ångest	0.18	atopisk	0.37
symptom	0.17	pneumoni	0.25	symtom	0.18	perioral	0.33
barn	0.17	luftvägsinfektion	0.24	episod	0.17	utslag	0.22

emfysem		faryngit		hemicrania		hyperglykemi	
lunga	0.48	hals	0.50	continua	0.90	glukos	0.41
lungsjukdom	0.34	slemhinna	0.27	huvudvärk	0.59	diabetes	0.37
obstruktiv	0.30	sjukdom	0.25	paroxysmal	0.56	insulin	0.32
lungemfysem	0.29	svalg	0.25	info	0.46	mmol	0.27
lungvävnad	0.21	behandling	0.23	indomethacin	0.41	måltid	0.24

hypotoni		pneumoni		prostatit		rinit	
blodtryck	0.61	crp	0.27	prostata	0.59	näsa	0.33
ortostatisk	0.38	bronkit	0.25	prostatakörtel	0.29	nästäppa	0.29
postural	0.28	kol	0.25	besvär	0.27	nässlemhinnan	0.25
hjärtfrekvens	0.24	otit	0.18	bäckenbotten	0.24	allergi	0.21
blodvolym	0.21	pneumoniae	0.18	bakterie	0.23	polyp	0.19

schizofreni		sinuit		tonsillit	
psykos	0.28	bihåleinflammation	0.29	tonsill	0.53
beteende	0.20	otit	0.29	halsfluss	0.38
tonår	0.18	öli	0.26	halsont	0.18
hjärnvolym	0.18	polyp	0.23	gas	0.18
mathalon	0.18	karies	0.22	polyp	0.18

Table 12. The size of the extracted context windows

Word	Number of words in window
artrit	5542
bronkit	5174
depression	11544
dermatit	7377
emfysem	3667
faryngit	8270
hemicrania	1101
hyperglykemi	2180
hypotoni	758
pneumoni	2465
prostatit	3482
rinit	2551
schizofreni	1153
sinuit	1402
tonsillit	2426

Table 13. The agreement rate of each category

Category	Agreement
The words have entirely different meanings	0.29
The words can be related to each other	0.24
The words are related and are often used together	0.25
The words are strongly related and can sometimes replace each other	0.08
The words are synonyms and often replaces each other	0.06
I don't know	0.08

For correlation, Kendall's tau was used. This choice is motivated by the data not being normally distributed. Kendalls's tau uses the difference between the number of concordant and discordant pairs. This is divided by the total number of pairs (Kendall, 1938). There was a significant relationship between the average rating of the words and their Dice similarity. A Kendall's tau coefficient test showed the following: tau=.28, p<.001.

Discussion

The correlation between the average human rating and the average Dice similarity was significant, but rather weak. What this correlation means is that the higher Dice similarity a word gets, the higher aver-

Table 14.

SNOMED term	Candidate related term	Average user rating	Dice similarity
artrit	psoriasisartrit	3.28	0.21
bronkit	luftvägsinfektion	3.84	0.24
depression	ångest	2.94	0.18
dermatit	eksem	3.51	0.43
emfysem	lungemfysem	4.35	0.29
faryngit	svalg	2.93	0.25
hemicrania	huvudvärk	3.26	0.59
hyperglykemi	diabetes	3.22	0.37
hypotoni	blodtryck	3.32	0.61
pneumoni	pneumoniae	3.95	0.18
prostatit	prostatakörtel	2.85	0.29
rinit	nästäppa	3.47	0.29
schizofreni	psykos	3.09	0.28
sinuit	bihåleinflammation	4.72	0.29
tonsillit	halsfluss	4.39	0.38

Table 15. Percentage of "I don't know" answers

SNOMED term	Candidate related term	Percentage of "I don't know" answers
hemicrania	indomethacin	72%
hemicrania	continua	68%
schizofreni	mathalon	61%
hemicrania	paroxysmal	52%
hemicrania	info	45%
hemicrania	huvudvärk	40%
hypotoni	postural	33%
dermatit	perioral	14%
sinuit	öli	9%
sinuit	polyp	8%

age human rating it gets. However, much can be said about these results. Firstly, the Dice similarity was in general rather low (91% of all word pairs got a Dice similarity less than 0.50, see Table 17). Words that are synonyms, or at least by the participants considered closely related should have a Dice similarity closer to 1 than 0. The overall quite low Dice similarity may be a reason for the weak correlation. An example of the low Dice similarity is for the word pair "emfysem-lungemfysem" (en: emphysema-pulmonary emphysema), where the average user rating is 4.35, whereas the Dice similarity was only 0.29. What this rating means it that the participants perceived the similarity as somewhere in between "The words are strongly related and can sometimes replace each other", and "The words are synonyms and often replaces each other". The word "lungemfysem" is a compound of the two words; "lunga" and "emfysem." Thus "lungemfysem" is a certain kind of emphysema, which makes it clear that they have a strong semantic relationship. It is difficult to decide on whether they are synonyms or just strongly related, but what is clear is that the participants were more accurate on rating these words than the system was.

Table 14 shows the highest rated candidate related term for every SNOMED CT term. Among these, the word pair which was least similar according to the participants, with an average rating of 2.85, was "prostatit-prostatakörtel" (en: prostatitis-prostate gland). The word "prostatit" means inflammation of the prostate gland and has the synonym "blåshalskörtelinflammation". In a best-case scenario, the system would have found this synonym. However, the word "blåshalskörtelinflammation" only occurs twice in the context window extracted around "prostatit". The word "prostatakörtel" on the other hand occurs 116 times. A comparison with the highest rated word pair, "sinuit-bihåleinflammation" (en: sinusitis-sinus infection) emphasizes how the word occurrence affects the similarity. "Bihåleinflammation" occurs 68 times together with "sinuit". A reason for "blåshalskörtelinflammation" rarely occurring might be that it is not, in general, being used that often. It might be that "prostatit" more often is described with an expression containing several words, and by limiting the analysis to unigrams, this definition is impossible to catch. Section 5.3, Table 12 shows that there was a considerable size difference of the context windows, where the window for "depression" was more than 15 times the size of that of "hypotoni." The fewer words in a context window, the more impact each word has on the similarity measure. Preferably, there would have been a more even distribution of texts among all the SNOMED CT terms.

Limiting the system only to handle unigrams may have decreased the overall performance of the system. Some of the terms might have to be described by more than one word. An example of this is "hyperglykemi" (en: hyperglycemia) which does not have a synonym. Instead, it is best described as "onormalt hög blodsockernivå" (en: abnormally high blood glucose level). This is a trigram and would therefore not be found by the system in this work. In the case of "onormalt hög blodsockernivå," the words "onormalt" and "hög" would have been removed in the noise removal stage, which is reasonable since they, taken out of context, have no apparent relation to "hyperglykemi." The fact that there is no synonym to "hyperglykemi" explains why none of the candidate related terms got a higher human rating than 3.22. There are other medical terms to which there are no synonyms, only correctly described by multiple-word expressions. This lack of synonyms emphasizes the need for a model that can manage n-grams larger than unigrams.

It was decided only to keep the nouns. However, some of the more complicated medical adjectives were left in the corpus. An example of this is the word "atopisk" (en: atopic) that often occur together with "dermatit" (en: dermatitis) as the expression "atopisk dermatit." Stagger did not manage to tag "atopisk" correctly. All 88 occurrences of this word were tagged with "NN," making it mistakenly included in the analysis. This inaccurate tagging was also the case with a few other words. As mentioned earlier, synonyms always belong to the same part of speech. If an adjective is included in the results, one can be certain that this is not a synonym. Had these words been excluded from the analysis, it is entirely possible that the results would have been more accurate.

Asking humans to put a rating on something always entails some arbitrariness. It is not certain that they know the meaning of the words, nor that they perceive the similarity rating in the same way. For each rating in the rating scale, an example was provided. These examples, however, was evidently not enough for the participants to be sure of what to answer. The highest, on average, rated word pair was "sinuit-bihåleinflammation" with an average rating of 4.72 (St. Dev. = 0.66). Although this is a rather clear case of synonymy, 49 participants did not rate them as synonyms. This mean that as much as a fifth of the participants, either did not know the meaning of the word or how to interpret the questionnaire.

The participants' knowledge in Swedish was not investigated. Since a majority of the participants are nurses in Sweden, it is possible to assume that their knowledge in Swedish is sufficient for performing their work. However, it would have been good to include a question about the participants' native language.

Table 15 shows that some words tended to get a lot of "I don't know" answers. Among the top ten, word pairs with "hemicrania" occur five times. The fact that "hemicrania" occurs so many times makes it reasonable to conclude that the participants, in general, had trouble with understanding the meaning of this word. The opaque meaning of the term "hemicrania" even for healthcare professional was confirmed by our domain expert. This is somewhat surprising since even the Swedish Mesh indicates that "huvudvärk" (en: "headache") and "hemicrania" are synonyms[44].

Generally speaking, according to our domain expert (who is a nurse specialized in health informatics), some answers from the participants were quite unexpected. Arguably, the cause was in the interpretation of the task. Looking directly at the question - "To what extent do the following words have a linguistic meaning similar to XXX" - one might be quite "hard" in his/her assessment. But if one takes into account the initial text that not only "capturing synonyms, but also words used in similar contexts", one seems to have made another assessment. Therefore, it might be possible that the questionnaire contained some unforeseen or veiled ambiguities in the formulation. However, what is relatively problematic at this stage is indeed informative, since we can build on these findings when we proceed on to the next step of development.

The purpose of this work was to extract synonyms and related words from a medical corpus. It was indeed possible to extract related words from the given E-care corpus. The evaluation showed that although the nitty-gritty distributional thesaurus needs improvements, it was able to extract words that are related to each other. The average human rating of the extracted words was 2.48, i.e. somewhere in between "The words can be related to each other" and "The words are related and are often used together". Enhancements can certainly include bigrams or multi-grams and not only unigrams since many medical terms (also in Swedish) are expressed as multi-words. Future work aims at a more in-depth investigation about the medical knowledge and the language proficiency of the evaluation study participants, as well as additional distributional semantic models and several corpus sizes. Although this experiment was not focused on the lay-ness of term, we can certainly observe that some of the related words are indeed lay, such as "hud" (en: "skin") extracted with "dermatit" (from Ancient Greek "derma", (en: skin) +ˈ-itis)[45], or "luftväg" (en: "windpipe") (from Ancient Greek ("brónkhos" (= "windpipe") +ˈ-itis.)[46]. Considering the small size and the noisiness of the corpus and the open issues about the human evaluation, these results of these experiments are certainly promising.

Growing Up: Creation and Evaluation of eCare_En_02

In this experiment (see also Strandqvist et al. 2018), we extend *the E-care@home Corpus* by adding web documents in English. For this expansion we used a two-step approach, namely:

1. Automatic extraction and evaluation of term seeds from *use cases, personas/scenarios*;
2. Creation and validation of specialized and domain-specific web corpora bootstrapped with term seeds automatically extracted in step 1.

In the first step, we build a term extraction algorithm that can automatically identify term candidates in project-specific *use cases, personas/scenarios*. These texts are narratives that describe a "system's behavior under various conditions as the system responds to requests from stakeholders" (Cockburn, 2000) and are nowadays normally included in many language technology projects (e.g. see Henkel et al. 2015). *Use cases, personas/scenarios* are relatively short texts - only a few dozen pages (see Press-

mann, 2005:657) - normally written by domain experts who know how to correctly use terms in their own domain. For this reason, we argue that they are a convenient textual resource (when available) to automatically extract term seeds to bootstrap specialized web corpora, thus overriding any tedious and sometimes controversial or arbitrary process normally required to compile term lists (e.g. see Vivaldi et al., 2007; Loginova et al., 2012). In our study, we focus on the medical terms that occur in use cases, personas/scenarios written in English for *E-care@home*. We complete this step with the evaluation (Precision and Recall metrics) of the term extractor against a gold standard made of SNOMED CT terms. The challenge of this phase is to create an accurate term extractor based on a relatively small textual resource, a task that is still under-investigated since most of existing term extractors are based on large corpora (e.g. Park et al., 2002; Nazarenko and Zargayouna, 2009).

In the second step, we use the term seeds extracted in the previous step to *bootcat* (Baroni and Bernardini, 2004) a medical web corpus and evaluate the quality of the corpus. Leveraging on the web as a textual source for language technology applications is a well-established idea (e.g. Kilgarrif et al. 2010) and many general- or special-purpose corpora have already been created. While bootstrapping a web corpus is common practice (many tools exist, either based on crawling or on search engine queries), the validation of web corpora is still a grey area. Currently, there is little research available on this topic (among the few who address the issue, see: Ciaramita & Baroni, 2006; Eckart et al., 2012; Schäfer et al. 2013; Kilgarriff, 2014). It follows that approaches are not standardized; thus it is not possible to compare results. In our study, we analyze and test several corpus profiling measures (e.g. Rayson and Garside, 2000; Oakes, 2008; Nanas and De Roeck, 2008, Rayson, 2008) and propose answers to the following questions: What is meant by "quality" of a web corpus? How can we assess the quality of a corpus automatically bootstrapped from the web? What if a bootstrapped web corpus contains documents that are *not* relevant to the target domain? Can we measure the domain-specificity of a corpus?

eCare Term Extractor

Generally speaking, automatic term recognition (ATR) deals with the extraction of domain-specific lexical units from text. Normally, the input of ATR is a large collection of documents, i.e., a special corpus, and the output is a vocabulary that is used for communicating specialized knowledge (L'Homme, 2014). Terms, extracted by an ATR system, represent a broad spectrum of concepts that exist in a domain knowledge. (Zadeh, 2016). In contrast, keyword extraction focuses on the extraction of topical words from individual documents for indexing (Hasan & Ng, 2014).

In this experiment, we conflate the two foci of ATR and keyword extraction and implement a term extractor from individual documents. The challenge of this step is to create a "good enough" term extractor based on a relatively small textual resource, a task that is still under-investigated since most of existing term extractors are based on large corpora (e.g. see Nazarenko and Zargayouna, 2009).

Arguably, the use of personas and use cases/scenarios, when available, is a good starting point to automatize the manual process of term seeds selection. The E-care term extractor developed for this purpose includes three main components. The first component (terminology extractor) uses a shallow syntactic analysis of the text to extract candidate terms. The second component (terminology validator) compares each of the candidate terms and their variations to SNOMED CT to produce candidate terms. The third component is a seed validator.

The terminology extractor uses the Stanford Tagger (Toutanova, Klein, Manning, & Singer, 2003) to assign a part-of-speech (POS) tag to each word in the texts. The tagged text is then searched sequentially with each of the syntactic patterns (Pazienza, Pennacchiotti, & Zanzotto, 2005) presented in Table 16.

The terminology validator takes the candidate terms produced in the previous step and matches them against SNOMED CT. If an exact match is not found, each word is stemmed. The stemmed words are permuted, and each permutation is then matched against SNOMED CT once again, this time using wildcards between the word, to allow for spelling variations. Matches are then ranked by TFIDF scores (cutoff = 200).

The seed generator generates three terms (i.e. triples) from the cutoff list when they occurred in the same document.

The E-care term extractor performance is summarized in Table 17. The terminology extractor has an extraction recall of 81.25% on the development set. When evaluated, the terminology validator achieves the following performance: Precision = 34.2%, Recall = 71%, F1 = 46.2%. These results are promising if compared with the state of the art of keyword extraction methods, but are moderate if compared with term-extractor based on large corpora.

Extrinsic Evaluation: Assessing Domainhood

Intrinsic evaluation is when the quality of a system is assessed by direct analysis of the system's characteristics, and how they relate to some given criteria, as shown above. Often, the hope is that good results in an intrinsic evaluation will be telling of a system's quality and of its aptness for further use in downstream tasks (however, this assumption might not always be true).

Unlike intrinsic evaluation, extrinsic evaluation does not assess or inspect the system directly. Rather, the system is assessed by using its output as input of another downstream task. The results of this downstream, task is then indicative of the quality of the original system. For instance, Kilgarriff et al., (2014) describe a method for extrinsic evaluation of web corpora by extracting collocations for lexicography, while Biemann et al. (2013) evaluate web corpora on two collocation identification tasks that focus on

Table 16. Syntactic patterns used for term recognition

Patterns
(noun)+
(adjective)(noun)+
(noun)(prep)(noun)+

Table 17. Current performance of E-care term extractor

	Metrics	**%**
Term candidate extraction	Extraction recall	81
Term validation	Precision	34.2
	Recall	71
	F1	46.2

different aspects of multiword expressions and different types of data. It must be stressed, however, that good results in one task do not necessarily imply good results in another task.

In this experiment, we evaluate how good the performance of the eCare term extractor is to bootstrap a web corpus based on the domain of the use cases. We measure the domainhood (or domain-specificity) against a reference corpus representing general language (see also Santini et al., 2018).

For corpus evaluation, we use metrics based on word frequency lists, namely rank correlation coefficients (Kendall and Spearman), KL divergence, log-likelihood. It makes sense to use domain-specific terms for both bootstrapping the web corpus and for evaluating its domainhood because the terms used as seeds (source terms) should be found in non-trivial proportions in the final corpus to be sure that the corpus is domain-representative. Estimating domainhood may be a useful preliminary check for domain adaptation (e.g. see Cuong et al., 2016; Hoang & Sima'an 2014; McClosky, 2010), ie when porting NLP applications from one domain to another.

1. **Correlation Coefficients:** The Kendall correlation coefficient (Tau) and Spearman correlation test (Rho) are non-parametric tests. They both measure how similar the order of two ranks is. (We used the R function "cor.test()" with method="kendall, spearman" to calculate the tests).
2. **Kullback–Leibler (KL) Divergence:** (a.k.a. relative entropy) is a measure of the "distance" between two distributions. The KL divergence quantifies how far-off an estimation of a certain distribution is from the true distribution. The KL divergence is non-negative and equal to zero if the two distributions are identical. In our context, the closer the value is to 0, the more similar two corpora are. (We used the R package "entropy", function "KL.empirical()" to compute KL divergence).
3. **Log-Likelihood (LL-G2):** LL-G2 (Dunning, 1993) has been used for corpus profiling (Rayson and Garside, 2000). The words that have the largest LL-G2 scores show the most significant word-frequency difference in two corpora. LL-G2 is not affected by corpus size variation.

For the evaluation, we use three web corpora, namely:

- **ukWaCsample (872 565 words):** A random subset of ukWaC, a general- purpose web corpus (Ferraresi et al., 2008).
- **Gold (544 677 words):** A domain-specific web corpus collected with hand-picked term seeds from the E-Care personas and use cases/scenarios.
- **Auto (492 479 words):** A domain-specific web corpus collected with automatically extracted term seeds from the E-Care personas and use cases/scenarios.

Measuring Rank Correlation

We computed the normalized frequencies of the three corpora (words per million) and ranked them (with ties). The plots of the first 1000 top frequencies of the three corpora are shown in Figure 5. From the plots, we can see that UkwaCsample has very little in common with both Gold and Auto (boxes 1 and 2), while Gold and Auto (box 3) are similar.

When testing the rank correlation (Kendall and Spearman), we observe a statistically significant positive rank correlation between Gold and Auto (see Figure 6, box 3; Figure 3, box 3), which means that words in Gold and in Auto tend to have similar ranks. Conversely, the correlation between ukWaCsample

Figure 5. Plotting 1000 top ranks: (from left to right): ukWaCsample and Gold (box 1), ukWaCsample and Auto (box2), and Gold and Auto (box 3)

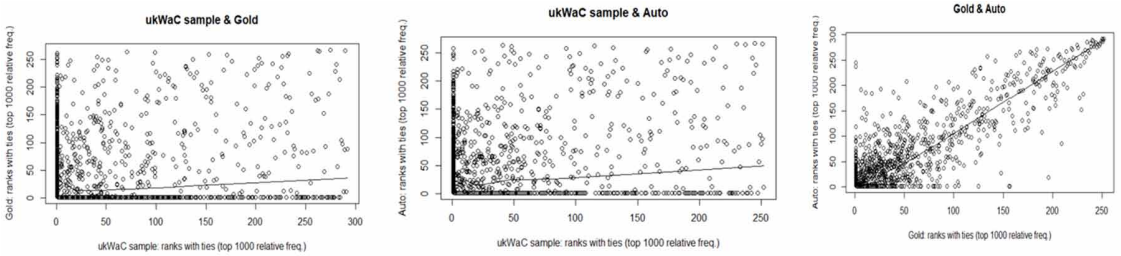
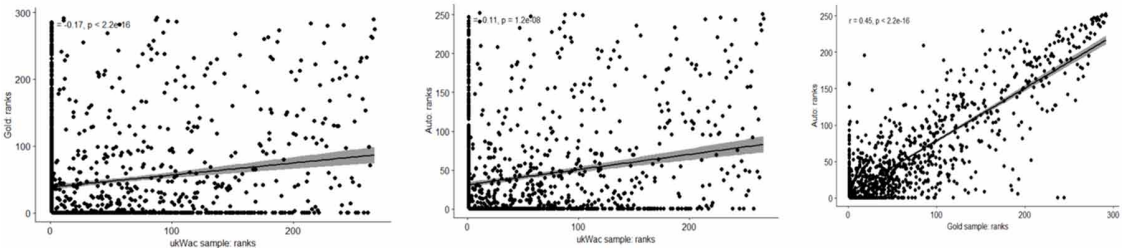


Figure 6. Kendall Tau: (from left to right):): ukWaCsample and Gold (box 1), ukWaCsample and Auto (box 2), and Gold and Auto (box 3)



and Gold and ukWaCsample and Auto is weak (see Figure 7, box 1 and box 2; Figure 6, box 1 and box 2), which essentially means that their ranks follow different distributions.

Kullback-Leibler Divergence

Before calculating KL divergence, a smoothing value of 0.01 was been added to the normalized frequencies. Results are shown in Table 18. The scores returned by KL distance for ukWacSample vs Gold (row 1) and ukWacSample vs Auto (row 2) – 7.544118 and 6.519677, respectively – are (unsurprisingly) large and indicate a wide divergence between the general-purpose ukWacSample and the domain-specific Gold and Auto. On the contrary, the KL score of 1.843863 indicates that Gold vs Auto (row 3) are similar to each other.

Figure 7. Spearman Rho: (from left to right):): ukWaCsample and Gold (box 1), ukWaCsample and Auto (box 2), and Gold and Auto (box 3)

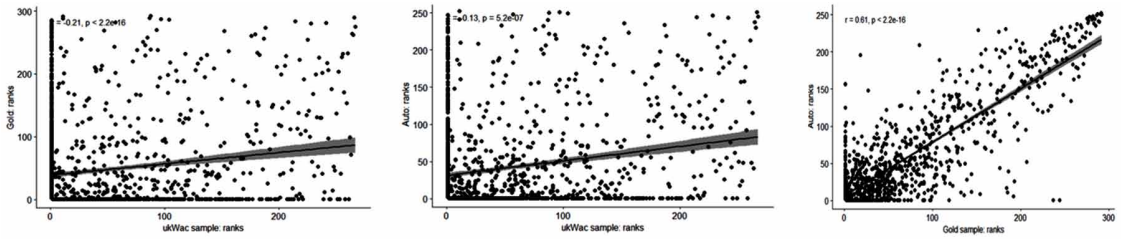


Table 18. KL scores

Corpora	KL Scores
ukWacSample vs Gold	7.544118
ukWacSample vs Auto	6.519677
Gold vs Auto	1.843863

Log-Likelihood (LL-G2)

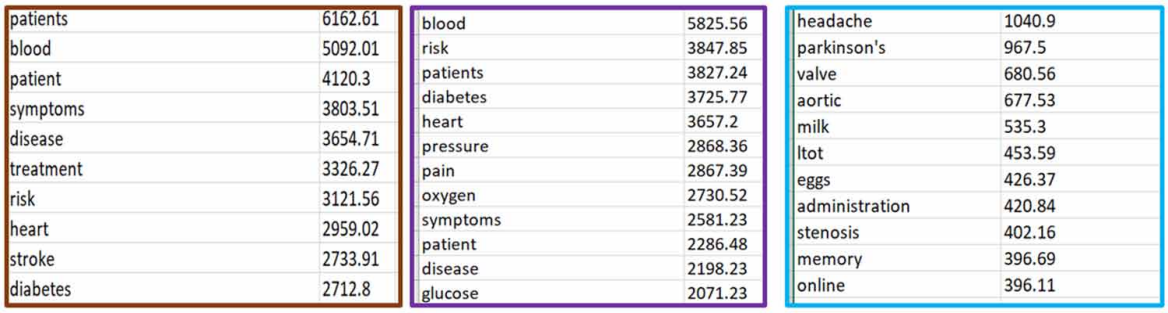
We computed LL-G2 scores on smoothed word frequencies. The total LL-G2 scores for the three web corpora (top 1000 words) are shown in Table 19. The larger the LL-G2 score of a word, the more different its distribution in two corpora. The large LL-G2 scores for ukWaCsample vs Gold (453 441.6) and for ukWaCsample vs Auto (393 705.9) indicate that these corpora are remarkably dissimilar if compared to the much smaller LL-G2 score returned for Gold vs Auto (114 694.2), which suggests that Gold and Auto are more similar to each other.

It is also possible to assess the statistical significance of the individual LL-G2 scores. Normally, a LL-G2 score of 3.8415 or higher is significant at the level of $p < 0.05$ and a LL-G2 score of 10.8276 is significant at the level of $p < 0.001$ (Desagulier, 2017). Figure 8 shows the breakdown of the top-ranked LL-G2 scores of three corpora. We take 3.8415 ($p < 0.05$) as a threshold and observe that ukWaCsample vs Gold (box 1) differs very much in the use of words such as "patient" or "patients" and "blood", and in ukWaCsample vs Auto (box 2) these words have a similar behavior. Conversely, these words are not in the top list of Gold vs Auto (box 3). Additionally, the LL-G2 scores in box 3 are much smaller in magnitude, which indicates that the difference between words is less pronounced.

Table 19. LL G2 scores of the three corpora

Corpora	Total LL-G2 Scores
ukWacSample vs Gold	453 441.6
ukWacSample vs Auto	393 705.9
Gold vs Auto	114 694.2

Figure 8. Top-ranked LL-G2 scores (from left to right): ukWaCsample and Gold (box 1), ukWaCsample and Auto (box 2), and Gold and Auto (box 3)



Discussion

Interestingly, this performance of the current version of the E-care term extractor did not affect detrimentally the quality of the resulting web corpus. This means that our approach is effective and help create a domain-specific corpus without any manual intervention. The corpus bootstrapped with the candidate terms returned by the term extractor is indeed domain-specific and it contains both lay and specialized terminology in the previous figures.

We have shown that it is possible to create a fairly accurate term extractor for relatively short texts written by domain experts. When used to bootstrap a web corpus, the automatically extracted term seeds create a corpus whose domain-specificity quality is similar to a corpus bootstrapped with hand-picked term seeds. This is an added value because corpus construction can be accelerated and standardized.

We have seen that well-established measures– such as rank correlation, KL divergence and log-likelihood (LL-G2 scores)– *do* give a coarse but grounded idea of domain-specificity. Essentially, they allow for an evaluation of the quality of a domain-specific web corpus and can also be used to pre-assess the portability of NLP tools from one domain-specific corpus to a different corpus belonging to another domain. Similar experiments have also been carried out on Swedish corpora with much the same results (Santini et al., 2018), showing that our approach may become a language-independent standardized step in corpus evaluation practice (intrinsic evaluation metrics).

We can now provide some empirical answers, namely: (1) in these experiments, "quality" means high density of medical terms related to certain illnesses described in the personas and use cases/scenarios; (2) we can assess the quality of a corpus automatically bootstrapped from the web by using metrics that are well-established and easily replicable; (3) we can get a coarse but robust indication of the similarities across corpora; (4) we can measure the domain-specificity of a corpus and assess whether it is satisfactorily domain-specific or whether the corpus needs some amends before being used for LT applications.

CONCLUSION AND FUTURE DIRECTIONS

In this chapter, the design, the creation and current use of the *E-care@home Corpus* was presented. More than a single corpus, it has been conceived to be a family of domain-specific sub-corpora. The domain of interest is the medical field of chronic diseases. The current version of the corpus, *eCare_Sv-En_03,* has been bootstrapped from the web via search engines using a selection of SNOMED CT medical terms. The corpus includes web texts written in Swedish (namely those included in *eCare_Sv_01* and *eCare_Sv_01+*) and in English (*eCare_En_02*). A further expansion is currently in progress in both languages.

The *E-care@home Corpus* proposes a new corpus design. This design is centered upon the idea of a flexible and extensible textual resource, where additional documents and additional languages will be appended over time. The main purpose of the corpus is to be used for building and training LT applications for the "layfication" of specialized sublanguages, namely the medical jargon used by professional staff working in eHealth and eCare. Although a case study based on the ongoing project *E-care@home* is presented here, it is claimed that this design is applicable to any kind of corpora and domains.

Exploratory experiments that leverage on subparts of *the E-care@home Corpus* were presented. Namely, supervised "lay-specialized" classification of Swedish web documents (subcorpus *eCare_Sv_01*), automatic extraction of words semantically related to medical terms (*eCare_Sv_01+*) and the assessment of the domain specificity of a corpus (*eCare_En_02*).

In these experiments, the focus was on the development of corpus-based LT applications that are simple but robust enough to disturbing factors, such as noise and corpus size variations.

It was argued that, for layfication tasks, the creation of a parallel or comparable corpus of specialized and lay documents is not needed. Aligning corpora, documents, sentences or words is a daunting task that takes time and engineering. The proposed approach was aimed to investigate whether a "regular" corpus bootstrapped with technical terms as seeds (Section 5.1) can also be used for layfication tasks. The answer is positive since off-the-shelf classifiers built with easy to extract bag-of-word features show promising performance (Section 5.2) and a simple model based on distributional semantics can extract sensible set of related terms (Section 5.3). The corpus can be easily expanded by adding more documents and additional languages (Section 5.4). Each increment of the corpus is annotated with metadata, so it is easy to extract sub-corpora based on specific attributes. It was also shown that the domain-specificity of a corpus bootstrapped with term seeds automatically extracted from use cases (i.e. single documents rather than a corpus) is equivalent to the domain-specificity of a corpus built with hand-picked seeds. This means that the seed selection phase (which is usually a delicate task that needs the knowledge of a domain expert and several iterations) can be streamlined and accelerated.

As stressed several times along the chapter, the corpus is conceived as work-in-progress, and future expansions are scheduled. For instance, to expand the Swedish part, a new corpus expansion has been bootstrapped by translating use case seeds (originally in English) into Swedish. To expand the English part, we used the SNOMED CT Swedish translations of chronic diseases.

Future experiments will include bilingual lay terminology induction, ontology creation from text, text simplification, or simplified summarization, i.e. tasks that all involve a layfication of technical terminology.

ACKNOWLEDGMENT

This research was supported by *E-care@home*, a "SIDUS – Strong Distributed Research Environment" project, funded by the Swedish Knowledge Foundation [kk-stiftelsen, Diarienr: 20140217]. Project website: <http://ecareathome.se/>. *E-care@home Corpus* website: <http://santini.se/eCareCorpus/home.htm>.

REFERENCES

Abrahamsson, E., Forni, T., Skeppstedt, M., & Kvist, M. (2014). Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language. *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@ EACL*, 57–65. 10.3115/v1/W14-1207

Åhlfeldt, H., Borin, L., Daumke, P., Grabar, N., Hallett, C., Hardcastle, D., . . . Willis, A. (2006). *Literature review on patient-friendly documentation systems?* Technical Report 2006/04, Department of Computing, Faculty of Mathematics and Computing, The Open University, Milton Keynes, UK.

Alirezaie, M. (2015). *Bridging the semantic gap between sensor data and ontological knowledge* (Ph.D. dissertation). Örebro University.

Alirezaie, M., Hammar, K., & Blomqvist, E. (2018a). SmartEnv as a network of ontology patterns. *Semantic Web*, *9*(6), 903–918.

Alirezaie, M., Hammar, K., Blomqvist, E., Nyström, M., & Ivanova, V. (2018b). *SmartEnv Ontology in E-care@home*. 9th Workshop on Semantic Sensor Networks (SSN) held in conjunction with ISWC 2018, Monterey, CA.

Antoine, E., & Grabar, N. (2017). Acquisition of Expert/Non-Expert Vocabulary from Reformulations. *Studies in Health Technology and Informatics*, *235*, 521–525. PMID:28423847

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596. doi:10.1162/coli.07-034-R2

Band, J., & Gerafi, J. (2015). *Fair Use/Fair Dealing Handbook*. Retrieved from http://infojustice.org/wp-content/uploads/2015/03/fair-use-handbook-march-2015.pdf

Barbaresi, A. (2015). Ad hoc and general-purpose corpus construction from web sources (PhD thesis). École Normale Supérieure de Lyon (Université de Lyon), France.

Barbaresi, A. (2016). Efficient construction of metadata-enhanced web corpora. In *10th Web as Corpus Workshop* (pp. 7-16). Academic Press. 10.18653/v1/W16-2602

Barbu, E. (2016). *D3.2: Multilingual Corpus. Project Deliverable. EXPERT (Exploiting Empirical appRoaches to Translation), Project funded by the People Programme*. Marie Curie Actions.

Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. *LREC 2004 - Fourth International Conference On Language Resources And Evaluation*.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*(3), 209–226. doi:10.100710579-009-9081-4

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (*Vol. 1*, pp. 238-247). Academic Press. 10.3115/v1/P14-1023

Baroni, M., Kilgarriff, A., Pomikálek, J., & Rychlý, P. (2006). WebBootCaT: instant domain-specific corpora to support human translators. In *Proceedings of EAMT* (pp. 247-252). Academic Press.

Basili, R., Pazienza, M. T., & Velardi, P. (1993). Acquisition of selectional patterns in sublanguages. *Machine Translation*, *8*(3), 175–201. doi:10.1007/BF00982638

Biemann, C., Bildhauer, F., Evert, S., Goldhahn, D., Quasthoff, U., Schäfer, R., ... Zesch, T. (2013). Scalable Construction of High-Quality Web Corpora. *JLCL*, *28*(2), 23–59.

Biemann, C., Heyer, G., Quasthoff, U., & Richter, M. (2007). The Leipzig Corpora Collection-monolingual corpora of standard size. *Proceedings of Corpus Linguistic*.

Bigeard, E., Grabar, N., & Thiessard, F. (2018). Detection and Analysis of Drug Misuses. A Study Based on Social Media Messages. *Frontiers in Pharmacology*, *9*, 2018. doi:10.3389/fphar.2018.00791 PMID:30140224

Borin, L., Grabar, N., Gronostaj, M. T., Hallett, C., Hardcastle, D., Kokkinakis, D., . . . Willis, A. (2007). Semantic Mining Deliverable D27. 2: Empowering the patient with language technology. Technical Report Semantic Mining, NOE 507505), 1–75. Göteborg: Göteborg University.

Branco, A., Cohen, K. B., Vossen, P., Ide, N., & Calzolari, N. (2017). Replicability and reproducibility of research results for human language technology: Introducing an LRE special section. *Language Resources and Evaluation*, *51*.

Burnard, L. (2007). *Reference guide for the British National Corpus (XML edition).* Retrieved from http://www.natcorp.ox.ac.uk/XMLedition/URG/

Cardillo, E. (2015). *Medical terminologies for patients. Annex 1 to SHN Work Package 3 Deliverable D3.3*. Retrieved from http://www.semantichealthnet.eu/index.cfm/deliverables/

Cardillo, E., Tamilin, A., & Serafini, L. (2011). A Methodology for Knowledge Acquisition in Consumer-Oriented Healthcare. *Knowledge Discovery, Knowledge Engineering and Knowledge Management, 249*.

Cederblad, G. (2018). *Finding Synonyms in Medical Texts – Creating a system for automatic synonym extraction from medical texts* (Bachelor thesis). Linköping University, Department of Computer Science.

Chapman, K., Abraham, C., Jenkins, V., & Fallow, L. (2003). Lay understanding of terms used in cancer consultations. *Psycho-Oncology*, *12*(6), 557–566. doi:10.1002/pon.673 PMID:12923796

Chmielik, J., & Grabar, N. (2009) Comparative study between expert and non expert biomedical writings: their morphology and semantics. In *Medical Informatics in a United and Healthy Europe: Proceedings of MIE 2009, the XXII International Congress of the European Federation for Medical Informatics* (Vol. 150, p. 359). IOS Press.

Ciaramita, M., & Baroni, M. (2006). A Figure of Merit for the Evaluation of Web-Corpus Randomness. *EACL 2006 - 11th Conference of the European Chapter of the Association for Computational Linguistics*.

Claveau, V., Hamon, T., Maguer, S. L., & Grabar, N. (2015). Health consumer-oriented information retrieval. *Studies in Health Technology and Informatics*, *210*, 80–84. PMID:25991106

Cockburn, A. (2000). *Writing effective use cases, The crystal collection for software professionals.* Addison-Wesley Professional Reading.

Cocoru, D., & Boehm, M. (2016). An analytical review of text and data mining practices and approaches in Europe Policy recommendations in view of the upcoming copyright legislative proposal. *OpenForumEurope*. Retrieved from http://www.openforumeurope.org/wp-content/uploads/2016/05/TDM-Paper-Diana-Cocoru-and-Mirko-Boehm.pdf

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. doi:10.1177/001316446002000104

Coradeschi, S., Cesta, A., Cortellessa, G., Coraci, L., Gonzalez, J., Karlsson, L., . . . Pecora, F. (2013). Giraffplus: Combining social interaction and long term monitoring for promoting independent living. In *Human system interaction (HSI), 2013 the 6th international conference on* (pp. 578-585). IEEE.

Coradeschi, S., Cesta, A., Cortellessa, G., Coraci, L., Galindo, C., Gonzalez, J., ... Loutfi, A. (2014). GiraffPlus: a system for monitoring activities and physiological parameters and promoting social interaction for elderly. In *Human-Computer Systems Interaction: Backgrounds and Applications 3* (pp. 261–271). Cham: Springer.

Costa, Â., Castillo, J. C., Novais, P., Fernández-Caballero, A., & Simoes, R. (2012). Sensor-driven agenda for intelligent home care of the elderly. *Expert Systems with Applications*, *39*(15), 12192–12204. doi:10.1016/j.eswa.2012.04.058

Cuong, H., Sima'an, K., & Titov, I. (2016). Adapting to all domains at once: Rewarding domain invariance in SMT. *Transactions of the Association for Computational Linguistics*, *4*, 99–112. doi:10.1162/tacl_a_00086

Dalianis, Henriksson, Kvist, Velupillai, & Weegar. (2015). Health bank-a workbench for data science applications in healthcare. *CAiSE Industry Track*, 1–18.

Dalianis, H., Hassel, M., & Velupillai, S. (2009). The stockholm epr corpus–characteristics and some initial findings. *Women*, *219*(906), 54.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*(2), 159–190. doi:10.1075/ijcl.14.2.02dav

Deléger, L., Cartoni, B., & Zweigenbaum, P. (2013). *Paraphrase detection in monolingual specialized/lay corpora*. Building and Using Comparable Corpora. doi:10.1007/978-3-642-20128-8_12

Deléger, L., & Zweigenbaum, P. (2009). Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*. Association for Computational Linguistics. 10.3115/1690339.1690343

Desagulier, G. (2017). *Corpus Linguistics and Statistics with R*. Springer. doi:10.1007/978-3-319-64572-8

Doing-Harris, K. M., & Zeng-Treitler, Q. (2011). Computer-assisted update of a consumer health vocabulary through mining of social network data. *Journal of Medical Internet Research*, *13*(2), e37. doi:10.2196/jmir.1636 PMID:21586386

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1).

Eckart, T., Quasthoff, U., & Goldhahn, D. (2012). The Influence of Corpus Quality on Statistical Measurements on Language Resources. In LREC (pp. 2318-2321). Academic Press.

Elhadad, N., & Sutaria, K. (2007). Mining a lexicon of technical terms and lay equivalents. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*. Association for Computational Linguistics. 10.3115/1572392.1572402

Escartin, C. P., & Torres, L. S. (2016). *D6. 3: Improved Corpus-based Approaches. Project Deliverable. EXPERT (Exploiting Empirical appRoaches to Translation), Project funded by the People Programme*. Marie Curie Actions.

Falkenjack, J., Mühlenbock, K. H., & Jönsson, A. (2013). Features indicating readability in swedish text. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*. Linköping University Electronic Press.

Ferraresi, A., Zanchetta, E., Bernardini, S., & Baroni, M. (2008). Introducing and evaluating ukWaC, a very large Web-derived corpus of English. *Proceedings of the 4th Web as Corpus Workshop (WAC-4) "Can we beat Google?"*

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382. doi:10.1037/h0031619

Fotokian, Z., Mohammadi Shahboulaghi, F., Fallahi-Khoshknab, M., & Pourhabib, A. (2017). The empowerment of elderly patients with chronic obstructive pulmonary disease: Managing life with the disease. *PLoS One*, *12*(4), e0174028. doi:10.1371/journal.pone.0174028 PMID:28369069

Gao, Q., & Vogel, S. (2011). Corpus expansion for statistical machine translation with semantic role label substitution rules. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers* (vol. 2, pp. 294-298). Association for Computational Linguistics.

Glavaš, G., & Štajner, S. (2015). Simplifying lexical simplification: do we need simplified corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (*Vol. 2*, pp. 63-68). Academic Press. 10.3115/v1/P15-2011

Goldberg, L., Lide, B., Lowry, S., Massett, H. A., O'Connell, T., Preece, J., ... Shneiderman, B. (2011). Usability and accessibility in consumer health informatics: Current trends and future challenges. *American Journal of Preventive Medicine*, *40*(5), S187–S197. doi:10.1016/j.amepre.2011.01.009 PMID:21521594

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In LREC (Vol. 29, pp. 31-43). Academic Press.

Grabar, N., Chauveau-Thoumelin, P., & Dumonet, L. (2015). *Study of Subjectivity in the Medical Discourse: Uncertainty and Emotions Advances in Knowledge Discovery and Management* (Vol. 5). Springer.

Grabar, N., & Hamon, T. (2014). Automatic extraction of layman names for technical medical terms. In *Healthcare Informatics (ICHI), 2014 IEEE International Conference on* (pp. 310-319). IEEE. 10.1109/ICHI.2014.49

Grabar, N., & Hamon, T. (2014). Unsupervised method for the acquisition of general language paraphrases for medical compounds. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)* (pp. 94-103). Academic Press. 10.3115/v1/W14-4812

Grabar, N., & Hamon, T. (2017). Understanding of unknown medical words. In *Proceedings of the Biomedical NLP Workshop associated with RANLP* (pp. 32-41). Academic Press. 10.26615/978-954-452-044-1_005

Grabar, N., Hamon, T., & Amiot, D. (2014). Automatic diagnosis of understanding of medical words. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)* (pp. 11-20). Academic Press. 10.3115/v1/W14-1202

Grabar, N., Krivine, S., & Jaulent, M. C. (2007). Classification of health webpages as expert and non expert with a reduced set of cross-language features. *AMIA ... Annual Symposium Proceedings - AMIA Symposium. AMIA Symposium*, *2007*, 284. PMID:18693843

Grabar, N., van Zyl, I., de la Harpe, R., & Hamon, T. (2014). The Comprehension of Medical Words. In *Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies* (vol. 5, pp. 334-342). SCITEPRESS-Science and Technology Publications, Lda.

Greenes, R. A. (2001). eCare and eHealth: The Internet meets health care. The Journal of medical practice management. *MPM*, *17*(2), 106–108. PMID:11680134

Gries, S. Th. (2013). Elementary statistical testing with R. In M. Krug & J. Schlüter (Eds.), *Research methods in language variation and change*. Cambridge University Press. doi:10.1017/CBO9780511792519.024

Grifoni-Waterman, R. (2016). *International fair use developments: Is fair use going global?* Retrieved July 2018. https://www.authorsalliance.org/2016/02/25/international-fair-use-developments-is-fair-use-going-global/

Grigonyte, G., Kvist, M., Wirén, M., Velupillai, S., & Henriksson, A. (2016). Swedification patterns of latin and greek affixes in clinical text. *Nordic Journal of Linguistics*, *39*(01), 5–37. doi:10.1017/S0332586515000293

Grishman, R., & Kittredge, R. (2014). *Analyzing language in restricted domains: sublanguage description and processing*. Psychology Press.

Habernal, I., Zayed, O., & Gurevych, I. (2016). C4Corpus: Multilingual Web-size corpus with free license. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 914–922). Portorož, Slovenia: European Language Resources Association (ELRA).

Handke, C., Guibault, L., & Vallbé, J. J. (2015). *Is Europe falling behind in data mining? Copyright's impact on data mining in academic research*. Academic Press.

Hasan, K. S., & Ng, V. (2014). Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (*Vol. 1*, pp. 1262-1273). Academic Press. 10.3115/v1/P14-1119

Haslerud, V., & Stenström, A.-B. (1995). The bergen corpus of london teenager language (colt). Spoken English on Computer, 235–42.

Heffer, C., Rock, F., & Conley, J. (Eds.). (2013). *Legal-Lay Communication: Textual Travels in the Law*. Oxford University Press. doi:10.1093/acprof:oso/9780199746842.001.0001

Henkel, M., Perjons, E., & Sneiders, E. (2015). Supporting Workflow and Adaptive Case Management with Language Technologies. In WorldCIST (pp. 543-552). Academic Press. doi:10.1007/978-3-319-16486-1_53

Heppin, K. F. (2010). *Resolving Power of Search keys in Medeval a Swedish Medical Test Collection with User Groups: Doctors and patients* (Ph.D. dissertation). University of Gothenburg.

Hoang, C., & Sima'an, K. (2014). Latent domain translation models in mix-of-domains haystack. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 1928-1939). Academic Press.

Hole, W. T., & Srinivasan, S. (2000). Discovering missed synonymy in a large concept-oriented Metathesaurus. In *Proceedings of the AMIA Symposium* (p. 354). American Medical Informatics Association.

Humphreys, B. L., McCray, A. T., & Cheh, M. L. (1997). Evaluating the coverage of controlled health data terminologies: Report on the results of the NLM/AHCPR large scale vocabulary test. *Journal of the American Medical Informatics Association*, *4*(6), 484–500. doi:10.1136/jamia.1997.0040484 PMID:9391936

Huo, H., Xu, Y., Yan, H., Mubeen, S., & Zhang, H. (2009, June). An elderly health care system using wireless sensor networks at home. In *Sensor Technologies and Applications, 2009. SENSORCOMM'09. Third International Conference on* (pp. 158-163). IEEE. 10.1109/SENSORCOMM.2009.32

Jackson, G. L., Powers, B. J., Chatterjee, R., Bettger, J. P., Kemper, A. R., Hasselblad, V., ... Gray, R. (2013). The patient-centered medical home: A systematic review. *Annals of Internal Medicine*, *158*(3), 169–178. doi:10.7326/0003-4819-158-3-201302050-00579 PMID:24779044

Jiang, L., & Yang, C. C. (2013). Using co-occurrence analysis to expand consumer health vocabularies from social media data. *2013 IEEE International Conference on Healthcare Informatics (ICHI)*, 74-81. 10.1109/ICHI.2013.16

Jimison, H., Gorman, P., Woods, S., Nygren, P., Walker, M., Norris, S., & Hersh, W. (2008). Barriers and drivers of health information technology use for the elderly, chronically ill, and underserved. *Evidence Report/technology Assessment*, *175*, 1–1422. PMID:19408968

Johansson, V., & Rennes, E. (2016). Automatic extraction of synonyms from an easy-to-read corpus. *Proceedings of the Sixth Swedish Language Technology Conference (SLTC-16)*.

Kann, V., & Rosell, M. (2005). Free construction of a free Swedish dictionary of synonyms. In *Proceedings of the 15th Nordic Conference of Computational Linguistics (NODALIDA 2005)* (pp. 105-110). Academic Press.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, *6*(1), 97–133. doi:10.1075/ijcl.6.1.05kil

Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory*, *1*(2), 263–276. doi:10.1515/cllt.2005.1.2.263

Kilgarriff, A. (2007). Googleology is bad science. *Computational Linguistics*, *33*(1), 147–151. doi:10.1162/coli.2007.33.1.147

Kilgarriff, A. (2014). *How to evaluate a corpus. Slides. 15th lecture of the Fred Jelinek Seminar series*. Institute of Formal and Applied Linguistics Charles University, Czech Republic Faculty of Mathematics and Physics.

Kilgarriff, A., Avinesh, P. V. S., & Pomikálek, J. (2011). BootCatting comparable corpora. In *9th International Conference on Terminology and Artificial Intelligence* (p. 123). Academic Press.

Kilgarriff, A., Reddy, S., Pomikálek, J., & Pvs, A. (2010). *A corpus factory for many languages*. In LREC workshop on Web Services and Processing Pipelines, Malta.

Kilgarriff, A., Reddy, S., & Pomikálek, J. and PVS A. (2010). A corpus factory for many languages. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Kim, J., Joo, J., & Shin, Y. (2009). An exploratory study on the health information terms for the development of the consumer health vocabulary system. *Studies in Health Technology and Informatics, 146*, 785.

Kittredge, R. (2003). Sublanguages and controlled languages. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (pp. 430–447). Oxford, UK: Oxford University Press.

Kittredge, R., & Lehrberger, J. (Eds.). (1982). *Sublanguage: Studies of language in restricted semantic domains*. Walter de Gruyter. doi:10.1515/9783110844818

Koch, S., & Hägglund, M. (2009). Health informatics and the delivery of care to older people. *Maturitas*, *63*(3), 195–199. doi:10.1016/j.maturitas.2009.03.023 PMID:19487092

Kokkinakis, D. (2006). Developing resources for Swedish Bio-Medical text mining. *Proceedings of the 2nd International Symposium on Semantic Mining in Biomedicine (SMBM)*.

Kokkinakis, D. (2012). *The Journal of the Swedish Medical Association - A Corpus Resource for Biomedical Text Mining in Swedish*. In The Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM), an LREC Workshop.

Kokkinakis, D., & Gronostaj, M. T. (2006). Comparing lay and professional language in cardiovascular disorders corpora. *WSEAS Transactions on Biology and Biomedicine, 3*(6), 429.

Krippendorff, K. (1980). *Content analysis. Beverley Hills* (Vol. 7). Sage Publications.

Krippendorff, K. (2011). *Computing Krippendorff's alpha-reliability*. Available: http://repository.upenn.edu/asc_papers/43

Kristoffersen, K. B. (2017). *Common Crawled Web Corpora: Constructing corpora from large amounts of web data. Thesis submitted for the degree of Master in Informatics: Programming and Networks (Language Technology group).* Department of Informatics, Faculty of mathematics and natural sciences, University of Oslo. Retrieved from http://urn.nb.no/URN:NBN:no-60569

Kristoffersson, A., & Lindén, M. (2017). *Understanding users of a future E-care@home system*. Retrieved from http://oru.diva-portal.org/smash/get/diva2:1073710/FULLTEXT01.pdf

Kucera, H., & Francis, W. (1979). *A standard corpus of present-day edited American English, for use with digital computers* (revised and amplified from 1967 version). Academic Press.

Küçükdurmaz, F., Gomez, M. M., Secrist, E., & Parvizi, J. (2015). Reliability, Readability and Quality of Online Information about Femoracetabular Impingement. *Archives of Bone and Joint Surgery*, *3*(3), 163–168. PMID:26213699

Kunz, M., & Osborne, P. (2010). A Preliminary Examination of the Readability of Consumer Pharmaceutical Web Pages. *Journal of Marketing Development and Competitiveness*, *5*(1), 33–41.

L'Homme, M. C. (2014). Terminologies and taxonomies. In J. R. Taylor (Ed.), The Oxford Handbook of the Word. Oxford University Press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

Lee, D. W. Y. (2001). Genres, Registers, Text Types, Domains, And Styles: Clarifying The Concepts And Navigating A Path Through The Bnc Jungle. *Language Learning & Technology*, *5*(3), 37–72.

Leroy, G., Helmreich, S., Cowie, J. R., Miller, T., & Zheng, W. (2008). Evaluating online health information: Beyond readability formulas. *AMIA ... Annual Symposium Proceedings - AMIA Symposium. AMIA Symposium*, *2008*, 394. PMID:18998902

Lewin, S., Munabi-Babigumira, S., Glenton, C., Daniels, K., Bosch-Capblanch, X., van Wyk, B. E., ... Scheel, I. B. (2010). Lay health workers in primary and community health care for maternal and child health and the management of infectious diseases. *Cochrane Database of Systematic Reviews*, (3): CD004015. PMID:20238326

Lilly, C. M., Zubrow, M. T., Kempner, K. M., Reynolds, H. N., Subramanian, S., Eriksson, E. A., ... Cowboy, E. R. (2014). Critical care telemedicine: Evolution and state of the art. *Critical Care Medicine*, *42*(11), 2429–2436. doi:10.1097/CCM.0000000000000539 PMID:25080052

Lin, D., Zhao, S., Qin, L., & Zhou, M. (2003). Identifying synonyms among distributionally similar words. IJCAI, 3, 1492-1493.

Lind, L., & Karlsson, D. (2018). The eHealth Diary – tailoring a solution for elderly, multimorbid homecare patients. Presented at Medical Information Europe (MIE2018), Gothenburg, Sweden.

Lindberg, B., Nilsson, C., Zotterman, D., Söderberg, S., & Skär, L. (2013). Using Information and Communication Technology in Home Care for Communication between Patients, Family Members, and Healthcare Professionals: A Systematic Review. *International Journal of Telemedicine and Applications*, *2013*, 1–31. doi:10.1155/2013/461829 PMID:23690763

Lippincott, T., Séaghdha, D. Ó., & Korhonen, A. (2011). Exploring subdomain variation in biomedical language. *BMC Bioinformatics*, *12*(1), 212. doi:10.1186/1471-2105-12-212 PMID:21619603

Loginova, E., Gojun, A., Blancafort, H., Guégan, M., Gornostay, T., & Heid, U. (2012). Reference lists for the evaluation of term extraction tools. *Proceedings of the Terminology and Knowledge Engineering Conference (TKE'2012)*.

Loutfi, Jönsson, Karlsson, Lind, Lindén, Pecora, & Voigt. (2016). *Ecare@Home: A Distributed Research Environment on Semantic Interoperability*. Presented at the 3rd EAI International Conference on IoT Technologies for HealthCare (HealhtyIoT 2016), Västerås, Sweden.

Lüdeling, A., Evert, S., & Baroni, M. (2007). Using web data for linguistic purposes. *Language and Computers*, *59*, 7.

Malá, M. (2017). A Corpus-Based Diachronic Study of a Change in the Use of Non-Finite Clauses in Written English. *Prague Journal of English Studies*, *6*(1), 151–166. doi:10.1515/pjes-2017-0009

McClosky, D. (2010). *Any domain parsing: automatic domain adaptation for natural language parsing* (PhD dissertation). Department of Computer Science at Brown University, Providence, RI. Retrieved from https://pdfs.semanticscholar.org/1c6e/e895c202a91a808de59445e3dbde2f4cda0e.pdf

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Taylor & Francis.

McGregor, B. (2005, January). Constructing a concise medical taxonomy. *Journal of the Medical Library Association: JMLA*, *93*(1), 121–123. PMID:15685285

Merilampi, S., & Sirkka, A. (2016). *Introduction to smart eHealth and eCare technologies*. CRC Press. doi:10.1201/9781315368818

Messai, R., Zeng, Q., Mousseau, M., & Simonet, M. (2006). *Building a bilingual French-English patient-oriented terminology for breast cancer*. Toronto, Canada: MedNet.

Miller, T., & Leroy, G. (2008). Dynamic generation of a Health Topics Overview from consumer health information documents. *International Journal of Biomedical Engineering and Technology*, *1*(4), 395–414. doi:10.1504/IJBET.2008.020069

Miller, T., Leroy, G., Chatterjee, S., Fan, J., & Thoms, B. (2007). A Classifier to Evaluate Language Specificity in Medical Documents. *Hawaii International Conference on System Sciences*. 10.1109/HICSS.2007.6

Nanas, N., & De Roeck, A. (2008). Corpus profiling with Nootropia. In *Proceedings of Workshop on Corpus Profiling for Information Retrieval and Natural Language Profiling*. London: BCS-IRSG.

Nazarenko, A., & Zargayouna, H. (2009). Evaluating term extraction. In *International Conference Recent Advances in Natural Language Processing (RANLP'09)* (pp. 299-304). Academic Press.

Nilsson, C., Öhman, M., & Söderberg, S. (2006). Information and communication technology in supporting people with serious chronic illness living at home–an intervention study. *Journal of Telemedicine and Telecare*, *12*(4), 198–202. doi:10.1258/135763306777488807 PMID:16774702

Norman, G. R., Arfai, B., Gupta, A., Brooks, L. R., & Eva, K. W. (2003). The privileged status of prestigious terminology: Impact of "medicalese" on clinical judgments. *Academic Medicine*, *78*(10Supplement), S82–S84. doi:10.1097/00001888-200310001-00026 PMID:14557104

Nyström, M., Merkel, M., Ahrenberg, L., Zweigenbaum, P., Petersson, H., & Åhlfeldt, H. (2006). Creating a medical english-swedish dictionary using interactive word alignment. *BMC Medical Informatics and Decision Making*, *6*(1), 35. doi:10.1186/1472-6947-6-35 PMID:17034649

Nyström, M., Merkel, M., Petersson, H., & Åhlfeldt, H. (2007). Creating a medical dictionary using word alignment: The influence of sources and resources. *BMC Medical Informatics and Decision Making*, *7*(1), 37. doi:10.1186/1472-6947-7-37 PMID:18036221

O'Brien, S. (1993). *Sublanguage, text type and machine translation* (Doctoral dissertation). Dublin City University.

Oakes, M. P. (2008). Statistical measures for corpus profiling. *Proceedings of the Open University Workshop on Corpus Profiling*.

Östling, R. (2013). Stagger: An open-source part of speech tagger for Swedish. *Northern European Journal of Language Technology*, *3*, 1–18. doi:10.3384/nejlt.2000-1533.1331

Ownby, R. L. (2005). Influence of vocabulary and sentence complexity and passive voice on the readability of consumer-oriented mental health information on the internet. In *AMIA 2005 Symposium Proceedings*. AMIA.

Park, Y., Byrd, R. J., & Boguraev, B. K. (2002). Automatic glossary extraction: beyond terminology identification. In *Proceedings of the 19th international conference on Computational linguistics* (vol. 1, pp. 1-7). Association for Computational Linguistics.

Pazienza, M., Pennacchiotti, M., & Zanzotto, F. (2005). Terminology extraction: an analysis of linguistic and statistical approaches. In *Terminology Extraction: An Analysis of Linguistic and Statistical Approaches*. Springer. doi:10.1007/3-540-32394-5_20

Pearson, J. (1998). *Terms in context* (Vol. 1). John Benjamins Publishing. doi:10.1075cl.1

Pieterse, A. H., Jager, N. A., Smets, E. M., & Henselmans, I. (2013). Lay understanding of common medical terminology in oncology. *Psycho-Oncology*, *22*(5), 1186–1191. doi:10.1002/pon.3096 PMID:22573405

Pomikálek, J., Rychlý, P., & Kilgarriff, A. (2009). Scaling to billion-plus word corpora. *Advances in Computational Linguistics*, *41*, 3–13.

Poprat, M., Markó, K., & Hahn, U. (2006). A language classifier that automatically divides medical documents for experts and health care consumers. *Studies in Health Technology and Informatics*, *124*, 503. PMID:17108568

Pressman, R. S. (2005). *Software engineering: a practitioner's approach*. Palgrave Macmillan.

Quintas, J. M. L. (2018). *Context-based Human-Machine Interaction Framework for Arti ficial Social Companions* (Doctoral dissertation). Universidade de Coimbra.

Quirk, C., Brockett, C., & Dolan, B. (2004). Monolingual machine translation for paraphrase generation. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004*.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, *13*(4), 519–549. doi:10.1075/ijcl.13.4.06ray

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing Corpora* (pp. 1-6). Association for Computational Linguistics.

Remus, S., & Biemann, C. (2016). Domain-Specific Corpus Expansion with Focused Webcrawling. LREC.

Rudd, R. E. (2013). Needed action in health literacy. *Journal of Health Psychology*, *18*(8), 1004–1010. doi:10.1177/1359105312470128 PMID:23349399

Santini, M. (2006). *Common Criteria for Genre Classification: Annotation and Granularity*. Workshop on Text-based Information Retrieval (TIR-06), In Conjunction with ECAI 2006, Riva del Garda, Italy.

Santini, M., Jönsson, A., Nyström, M., & Alirezai, M. (2017). A Web Corpus for eCare: Collection, Lay Annotation and Learning - First Results. Workshop LTA'17 (Language Technology Applications 2017) co-located with FedCSIS 2017, Prague. In M. Ganzha, L. Maciaszek, & M. Paprzycki (Eds.), *Position Papers of the 2017 Federated Conference on Computer Science and Information Systems, Proceedings*, (Vol. 12, pp. 71-78). Academic Press.

Santini, M., Strandqvist, W., Nyström, M., Alirezai, M., & Jönsson, A. (2018). *Can we Quantify Domain-hood? Exploring Measures to Assess Domain-Specificity in Web Corpora*. TIR 2018 - 15th International Workshop on Technologies for Information Retrieval.

Schäfer, R. (2016). CommonCOW: Massively Huge Web Corpora from CommonCrawl Data and a Method to Distribute them Freely under Restrictive EU Copyright Laws. LREC.

Schäfer, R., Barbaresi, A., & Bildhauer, F. (2013). The good, the bad, and the hazy: Design decisions in web corpus construction. *Proceedings of the 8th Web as Corpus Workshop*.

Schäfer, R., Barbaresi, A., & Bildhauer, F. (2014). Focused web corpus crawling. In *Proceedings of the 9th Web as Corpus workshop (WAC-9)* (pp. 9-15). Academic Press. 10.3115/v1/W14-0402

Schäfer, R., & Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. In LREC (pp. 486-493). Academic Press.

Schäfer, R., & Bildhauer, F. (2013). Web corpus construction. *Synthesis Lectures on Human Language Technologies*, *6*(4), 1–145. doi:10.2200/S00508ED1V01Y201305HLT022

Scott, N., & Weiner, M. F. (1984). "Patientspeak": An exercise in communication. *Journal of Medical Education*. PMID:6492107

Seedor, M., Peterson, K. J., Nelsen, L. A., Cocos, C., McCormick, J. B., Chute, C. G., & Pathak, J. (2013). Incorporating expert terminology and disease risk factors into consumer health vocabularies. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*. NIH Public Access.

Seedorff, M., Peterson, K. J., Nelsen, L. A., Cocos, C., McCormick, J. B., Chute, C. G., & Pathak, J. (2013). Incorporating expert terminology and disease risk factors into consumer health vocabularies. In Biocomputing 2013 (pp. 421-432). Academic Press.

Seljan, S., Baretić, M., & Kučiš, V. (2014). Information Retrieval and Terminology Extraction in Online Resources for Patients with Diabetes. *Collegium Antropologicum*, *38*(2), 705–710. PMID:25145011

Sharoff, S., Rapp, R., Zweigenbaum, P., & Fung, P. (Eds.). (2013). *Building and using comparable corpora*. Springer. doi:10.1007/978-3-642-20128-8

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, *85*(3), 257. PMID:15733050

Sioutis, M., Alirezaie, M., Renoux, J., & Loutfi, A. (2017). *Towards a Synergy of Qualitative Spatio-Temporal Reasoning and Smart Environments for Assisting the Elderly at Home*. 30th International Workshop on Qualitative Reasoning (held in conjunction with IJCAI 2017), Melbourne, Australia.

Smith, C. A., & Wicks, P. J. (2008). PatientsLikeMe: Consumer health vocabulary as a folksonomy. *AMIA ... Annual Symposium Proceedings - AMIA Symposium. AMIA Symposium*, *2008*, 682. PMID:18999004

Soergel, D., Tse, T., & Slaughter, L. A. (2004). Helping healthcare consumers understand: an "interpretive layer" for finding and making sense of medical information. In Medinfo (pp. 931-935). Academic Press.

Soobrah, R., & Clark, S. K. (2012). Your patient information website: How good is it? *Colorectal Disease*, *14*(3), e90–e94. doi:10.1111/j.1463-1318.2011.02792.x PMID:21883807

Strandqvist, W., Santini, M., Lind, L., & Jönsson, A. (2018). Towards a Quality Assessment of Web Corpora for Language Technology Applications. In *Proceedings of TISLID18: Languages For Digital Lives And Cultures*. Ghent University.

Suryadevara, N. K., Gaddam, A., Rayudu, R. K., & Mukhopadhyay, S. C. (2012). Wireless sensors network based safe home to care elderly people: Behaviour detection. *Sensors and Actuators. A, Physical*, *186*, 277–283. doi:10.1016/j.sna.2012.03.020

Tchami, O. W., & Grabar, N. (2014). Towards automatic distinction between specialized and non-specialized occurrences of verbs in medical corpora. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)* (pp. 114-124). Academic Press. 10.3115/v1/W14-4814

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 1. 10.3115/1073445.1073478

Versley, Y., & Panchenko, Y. (2012). Not just bigger: Towards better-quality Web corpora. In *Proceedings of the seventh Web as Corpus Workshop (WAC7)* (pp. 44-52). Academic Press.

Vivaldi, J., & Rodríguez, H. (2007). Evaluation of terms and term extraction systems: A practical approach. Terminology. *International Journal of Theoretical and Applied Issues in Specialized Communication*, *13*(2), 225–248.

Volansky, V., Ordan, N., & Wintner, S. (2015). On the features of translationese. *Digital Scholarship in the Humanities*, *30*(1), 98–118. doi:10.1093/llc/fqt031

Vydiswaran, V. V., Mei, Q., Hanauer, D. A., & Zheng, K. (2014). Mining consumer health vocabulary from community-generated text. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association.

Wang, Y. (n.d.). Automatic Recognition of Text Difficulty from Consumers Health Information. *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*.

Williams, N., & Ogden, J. (2004). The impact of matching the patient's vocabulary: A randomized control trial. *Family Practice*, *21*(6), 630–635. doi:10.1093/fampra/cmh610 PMID:15520032

Wong, W., Liu, W., & Bennamoun, M. (2011). Constructing specialized corpora through analysing domain representativeness of websites. *Language Resources and Evaluation*, *45*(2), 209–241. doi:10.100710579-011-9141-4

Zadeh, B. Q. A. (2016). Study on the Interplay Between the Corpus Size and Parameters of a Distributional Model for Term Classification. In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)* (pp. 62-72). Academic Press.

Zeng, Q., Kim, E., Crowell, J., & Tse, T. (2005). A text corpora-based estimation of the familiarity of health terminology. In *International Symposium on Biological and Medical Data Analysis* (pp. 184-192). Springer. 10.1007/11573067_19

Zeng, Q., Tse, T., Divita, G., Keselman, A., Crowell, J., Browne, A. C., … Ngo, L. (2015). Term Identification Methods for Consumer Health Vocabulary Development. *Journal of Medical Internet Research, 9*(1), 4.

Zeng, Q. T., & Tse, T. (2006). Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association*, *13*(1), 24–29. doi:10.1197/jamia.M1761 PMID:16221948

Zeng-Treitler, Q., Kim, H., Goryachev, S., Keselman, A., Slaughter, L., & Smith, C. A. (2007). Text characteristics of clinical reports and their implications for the readability of personal health records. In Medinfo. MEDINFO (2nd ed.; vol. 12, pp. 1117-1121). Academic Press.

Zheng, J., & Yu, H. (2017). Readability formulas and user perceptions of electronic health records difficulty: A corpus study. *Journal of Medical Internet Research*, *19*(3), e59. doi:10.2196/jmir.6962 PMID:28254738

Zheng, W., Milios, E., & Watters, C. (2002). Filtering for medical news items using a machine learning approach. In *Proceedings of the AMIA Symposium* (p. 949). American Medical Informatics Association.

ENDNOTES

1 iWeb: The Intelligent Web-based Corpus has been released in May 2018. For further details, see <<https://corpus.byu.edu/iweb/help/iweb_overview.pdf>> and the list of corpora available at BYU (Brigham Young University, Utah, USA) <<https://corpus.byu.edu/>>. URLs retrieved July 2018.

2 See <<https://www.sics.se/projects/digital-inkludering-i-det-uppkopplade-samhallet-for-grupper-med-speciella-behov>>. In Swedish. Retrieved July 2018.

3 <<http://ecareathome.se/>>.

4 SEMANTICMINING: Semantic Interoperability and Data Mining in Biomedicine <<https://cordis.europa.eu/project/rcn/71155_en.html>>. Retrieved July 2018.

5 <<http://www.semantichealthnet.eu/>>. Retrieved July 2018. This project has contributed to give to the establishment of the European Institute for Innovation through Health Data (i~HD) <<https://www.i-hd.eu/>>, a permanent organization aimed to "develop and promote best practices in the governance, quality, semantic interoperability and uses of health data, including its reuse for research".

6 One of Accurat's main objectives was to "To develop methods and tools for automatic acquisition of comparable corpora from the Web" <<http://www.accurat-project.eu/index.php-p=about.htm>>. Retrieved July 2018.

7 TTC (Terminology Extraction, Translation Tools and Comparable Corpora) developed and adapted tools for "gathering and managing comparable corpora, collected from the web, and managing terminologies" <<https://cordis.europa.eu/docs/projects/cnect/5/248005/080/publishing/readmore/TTC-public-annual-report-2012.pdf>>. Retrieved July 2018.

8 "EXPERT (EXPloiting Empirical appRoaches to Translation) aims to train young researchers, namely Early Stage Researchers (ESRs) and Experienced Researchers (ERs), to promote the research, development and use of hybrid language translation technologies. The overall objective of EXPERT is to provide innovative research and training in the field of Translation memory and Machine Translation Technologies". <<http://expert-itn.eu/?q=content/deliverables>>. Retrieved July 2018.

9 <<https://dkpro.github.io/dkpro-c4corpus/>>. Retrieved July 2018.

10 <<https://en.wikipedia.org/wiki/Blekko >>. Retrieved July 2018.

11 <<http://agilemanifesto.org/ >> Retrieved July 2018.

12 <<https://www.nhs.uk/conditions/iron-deficiency-anaemia/ >>. Retrieved July 2018.

13 <<https://www.health24.com/Medical/Anaemia/Anaemia-20130216-3>>. Retrieved July 2018.

14 <<https://www.1177.se/>>. Retrieved July 2018.

15 "Applied in the context of machine learning, this means that if two algorithms have broadly similar performance for the criteria identified as the most important for a particular project — accuracy and stability, say — we should always prefer the "simpler" one." <<https://www.teradata.com/Blogs/Occam%E2%80%99s-razor-and-machine-learning#/>>. Retrieved July 2018.

16 Review <<https://telepresencerobots.com/reviews/giraff/giraff-plus-revolutionizing-remote-patient-care>>. Retrieved July 2018.

17 "Healthcare Is Coming Home With Sensors And Algorithms" in The Medical Futurist <<http://medicalfuturist.com/healthcare-is-coming-home/ >>. Retrieved July 2018.

18 For example, see the recently funded project (2016-2019) led by M. Popescu: "*Linguistic Summarization of Sensor Data for Early Illness Recognition in Eldercare*" <<https://www.eldertech.missouri.edu/projects/linguistic-summarization-of-in-home-sensor-data-trends/ >>. Retrieved July 2018.

19 For instance, CARRE (Personalized patient empowerment and shared decision support for cardiorenal disease and comorbidities) <<https://www.carre-project.eu/>>; Designing for People with Dementia: designing for mindful self-empowerment and social engagement <<https://cordis.europa.eu/project/rcn/199934_en.html>>. Retrieved July 2018.

20 Several initiative exist at present. For instance, the Open Access, Collaborative Consumer Health Vocabulary Initiative<<http://consumerhealthvocab.chpc.utah.edu/CHVwiki/>>; Flashcards on consumer health terminology <<https://quizlet.com/70547741/chapter-1-consumer-health-vocab-flash-cards/>>. Retrieved July 2018.

21 The concept of "register" has numberless definitions and characterizations in linguistics and related fields. <<https://en.wikipedia.org/wiki/Register_(sociolinguistics)>>. Retrieved July 2018.

22 This is a registry for registering linguistic terms used in various fields of translation, computational linguistics and natural language processing and defining mappings both between different terms and the same terms used in different systems. The registers identified are: bench-level register,

dialect register, facetious register, formal register, in house register, ironic register, neutral register, slang register, taboo register, technical register, vulgar register, <<https://en.wikipedia.org/wiki/Register_(sociolinguistics)>>. Retrieved July 2018.

[23] <<https://plainlanguage.gov/guidelines/words/avoid-jargon/ >>. Retrieved July 2018.

[24] <<https://skillcrush.com/2015/03/26/99-tech-terms/>>. Retrieved July 2018.

[25] For instance, see "Small Data and Big Health Benefits" <<https://research.cornell.edu/news-features/small-data-and- big-health-benefits >>. Retrieved July 2018.

[26] The UMLS integrates and distributes key terminology, classification and coding standards, and associated resources to promote creation of more effective and interoperable biomedical information systems and services, including electronic health records. <<https://www.nlm.nih.gov/research/umls/>>. Retrieved July 2018.

[27] "The deCODEme service was discontinued in January 2013 and deCODE genetics stopped selling personal genetic tests altogether." <<https://en.wikipedia.org/wiki/DeCODE_genetics#deCODEme>>. Retrieved July 2018.

[28] "GeneWiki+ is a mirror of the Gene Wiki project on Wikipedia, running on top of the Semantic Mediawiki framework. Content from Wikipedia is mirrored in near-real time by our server and modified to work with Semantic Mediawiki's special semantic links. This allows you, the user, to ask simple queries that exploit the huge amount information in Wikipedia." <<https://archive.is/4Aqhp#selection-255.0-267.253>>. Retrieved July 2018.

[29] "PharmGKB is a comprehensive resource that curates knowledge about the impact of genetic variation on drug response for clinicians and researchers." <<https://www.pharmgkb.org/ >>. Retrieved July 2018.

[30] <<https://en.wikipedia.org/wiki/Agile_software_development>>. The principles of the Agile strategy is explained in the Agile Manifesto <<http://agilemanifesto.org/>>. Retrieved July 2018.

[31] "The BootCaT front-end is a graphical interface for the BootCaT toolkit (Baroni & Bernardini 2004). It automates the process of finding reference texts on the web and collating them in a single corpus". Read more here: <<https://bootcat.dipintra.it/>>. Retrieved July 2018.

[32] 228 initial URL seeds; 155 retrieved URLs; 73 URLs discarded either because they were empty documents or containing non-Swedish text. The list of 155 url seeds are available on the corpus website.

[33] Arguably, also the project "Text-based measures of information quality in online health information", recently funded in the UK, will address some of these issues.

[34] The annotated corpus has a xml format and is available here <http://santini.se/eCareCorpus/home.htm>. Also accessible from <<http://ecareathome.se/>>.

[35] See: <<http://santini.se/eCareCorpus/home.htm>>. Also accessible from <<http://ecareathome.se/>>.

[36] All calculations of intercoder/interrater reliability coefficients for lay-specialized labels (nominal data) coded by two annotators have been computed using the ReCal2 online calculator <<http://dfreelon.org/utils/recalfront/recal2/>>. Retrieved July 2018.

[37] Options: Classify - Meta – FilteredClassifiers. See: Weka Explorer Interface. <<http://weka.sourceforge.net/doc.dev/weka/classifiers/meta/FilteredClassifier.html>>. Retrieved July 2018.

[38] "In information retrieval, tf–idf or TFIDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus". <<https://en.wikipedia.org/wiki/Tf%E2%80%93idf>>. Retrieved July 2018.

[39] "TP / True Positive: case was positive and predicted positive". <<https://www.kdnuggets.com/faq/precision-recall.html >>. Retrieved July 2018.

[40] "A receiver operating characteristic curve, i.e. ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.". <<https://en.wikipedia.org/wiki/Receiver_operating_characteristic>>. Retrieved July 2018.

[41] Corrected Paired t-test is a statistical hypothesis testing method for comparing the result of measuring one group twice (here with two different classifiers). By taking into account both mean and variance of the differences between these two measures over several runs, it calculates a t-value. Using this value and desired significance level (normally 5%), the probability that these two measurements are significantly different can be obtained by looking it up from a t-distribution table. Consequently, one can say that these classifiers with a certain degree of confidence (100 - significance level) are significantly different or not. Paired t-test wrongly assumes that these differences between accuracy of two classifiers are independent and therefore has normal distribution (which is in fact not true because test sets and training sets overlap). The corrected resampled paired t-test boosts the paired t-test by entering the fraction of data used for testing and training into t-calculation formula (see <<http://imagej.net/Trainable_Weka_Segmentation_-_How_to_compare_classifiers >>. Retrieved July 2018.; Witten et al., 2016).

[42] "Statistical significance helps quantify whether a result is likely due to chance or to some factor of interest,". When a finding is significant, it simply means you can feel confident that it is real, not that you just got lucky (or unlucky) in choosing the sample.

[43] <<https://www.1177.se/ >>. Retrieved July 2018.

[44] <<https://mesh.kib.ki.se/Mesh/search/?searchterm=hemicrania >>. Retrieved July 2018.

[45] <<https://en.wiktionary.org/wiki/dermatitis>>. Retrieved July 2018.

[46] <<https://en.wiktionary.org/wiki/bronchitis>>. Retrieved July 2018.