Change of biomedical domain terminology over time

Grigonyte, Gintare ; Rinaldi, Fabio ; Volk, Martin

DOI: https://doi.org/10.3233/978-1-61499-133-5-74

Originally published at:
Grigonyte, Gintare; Rinaldi, Fabio; Volk, Martin (2012). Change of biomedical domain terminology over time. In: Human Language Technologies – The Baltic Perspective (Baltic HLT 2012), Tartu, 4 October 2012 - 5 October 2012. I O S Press, 74-81.
DOI: https://doi.org/10.3233/978-1-61499-133-5-74

Human Language Technologies – The Baltic Perspective
A. Tavast et al. (Eds.)

doi:10.3233/978-1-61499-133-5-74

Change of Biomedical Domain Terminology Over Time

Gintarė GRIGONYTĖ[a,1], Fabio RINALDI[a] and Martin VOLK[a]
[a] *The Institute of Computational Linguistics, University of Zürich*

Abstract. Biomedical text processing is relying heavily on terminological resources. Independently of the method used for creating terminologies, either automatically extracted from a domain corpus or human crafted, there is one aspect of which is rarely considered – that terms evolve over time. Terms in the domain literature change due to many factors: new factual evidence, proposing new hypothesis or denying old ones, a shift towards increasing specificity, variation in expression, different people working independently on the same novel phenomenon, etc. This paper reports an experimental investigation carried out on biomedical domain literature capturing how specific domain terminology changes over time.

Keywords. terminology evolution, domain terminology, biomedical domain, PubMed

Introduction

The textual legacy of the biomedical domain is rapidly growing. Currently, large databases like NML PubMed [1], which maintain a record of a large part of medical research literature, could not operate without advanced information retrieval systems, which require terminologies and organized domain knowledge such as ontologies. Independently of the method used for generating terminologies, either automatically extracted from a domain corpus or human crafted, there is one aspect which is rarely considered – that terms evolve over time. Terms in the biomedical literature change due to many factors: new factual evidence, proposing new hypothesis or denying old ones, a shift towards increasing specificity, variation in expression, different people working independently on the same novel phenomenon, etc. Therefore a knowledge on how terms evolve can be useful in several tasks such as search over domain documents, improving named entity recognition, or event and relationship extraction.

[1] Corresponding Author: Gintarė Grigonytė, Institute of Computational Linguistics, Binzmühlestr. 14, 8050 Zurich, Switzerland; Email: gintare@cl.uzh.ch

As a broader implication, the evolution of the terminology indicates the evolving knowledge and thus might even suggest the emerging of new domain fields or changing of the discourse.

Term evolution can be observed as a shift of lexical items and a shift of the semantic meaning. The shift in lexical level includes the disappearing of terms and the emergence of new terms. This type of shift can be pinpointed by the methods of corpus linguistics. The semantic type of shift can be detected by tracking what changes occur within the structure of existing ontologies or taxonomies to which a term can be mapped. In this paper we analyze the change of terminology at the lexical level.

1. Related Research

Lexical change over time is an established research line in philology [2] and in particular in historical linguistics [3], [4], [5]. Very often these studies have to concentrate on a small subset of words, or sometimes even on a single word, as insights about lexical change are gathered from several overlapping fields like historical, social, linguistics, political. Absence of large diachronic corpora is also a serious limitation. Therefore automatic inference about general language on the level of semantics over historical data is a very complex task.

To our knowledge there are several related studies in the field of information retrieval.

For example, [6] describe a method for incorporating so-called *semantically identical temporally altering concepts* (SITAC) like (Hillary Clinton, Hillary Rodham) for query translation. SITACs are automatically extracted like a bundle of a term and words associated with it. All SITACs are pair-wise matched looking for the highest rank in association overlapping. Authors report that using SITACs improves data access by 5% in precision and 10% in recall.

[7], [8] describe an approach based on word-sense-disambiguation, experimenting with the Times archive. The method includes: a) extracting of terms and their co-occuring terms, b) deriving word-senses, i.e., clusters, from clustering term co-occurrence graphs for specific periods of time. Jaccard similarity coefficient is used to compare clusters and track evolution in time.

[9] and [10] use MeSH keywords to track structural changes of MeSH over time.

2. Data and Methods

We used PubMed, the largest database of biomedical literature, indexing over 22 million citation references for biomedical literature from MEDLINE, life science journals, and online books, as a chronological corpus with documents from 1881 to 2012. Over 11 million documents that contain a title and an abstract have been indexed with the Indri IR system [11]. The PharmGKB [12] terminology part on diseases (2010 January release) comprising around 56000 terms has been used for analyzing how the usage of domain terms change over time. The experiments which we describe in this paper concern particularly lexical change of terminology and excludes semantic shift in meaning.

3. Experimental Results

The PubMed citation database is a unique resource spanning over the history of medicine of more than 100 years. Figure 1. shows the distribution of the number of publications (PubMed citations) over the years. The early history of PubMed records hundreds to thousands of citations per year. However starting from around 1975 and then later from around 2000 the amount of citations increases to tens of thousands. Currently the amount of citations reaches around half a million per year.

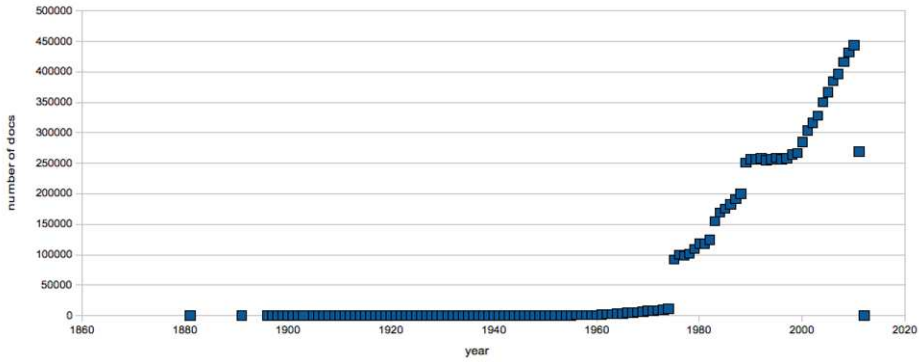

Figure 1. Number of citations registered in NML PubMed per year[2].

Together with the increase of textual data, more occurrences of terminology can be observed. Figure 2. shows the distributional pattern of the diseases terminology. The number of usage of terms of the PharmGKB disease terminology rises together with the increase of citations. However when the number of citations is fairly large, i.e., over 30,000, the number of terms in those documents tends to become stable.

It is important to note that not only a single lexical representation of a term of the PharmGKB Disease database was used, but also its variants, e.g.:

```
    <Entry entryId="C0006826_9" baseForm="malignancy" type="Neoplastic Process"
mlfreq="147570">
    <SourceDC sourceId="UMLSDisease:C0006826"/> <PosDC posName="POS" pos="N"/>
    <Variant writtenForm="malignancies" type="orth1" mlfreq="58659"/>
    <Variant writtenForm="Malignancy" type="orth1" mlfreq="1715"/>
    <Variant writtenForm="malignancy" type="orth1" mlfreq="86278"/>
    <Variant writtenForm="MALIGNANCY" type="orth1" mlfreq="180"/>
    <Variant writtenForm="Malignancies" type="orth1" mlfreq="738"/>
    <DC att="umls_cui" val="C0006826"/>
```

[2]Pubmed version 2012 January contains few documents from 2012 and not all documents from 2011.

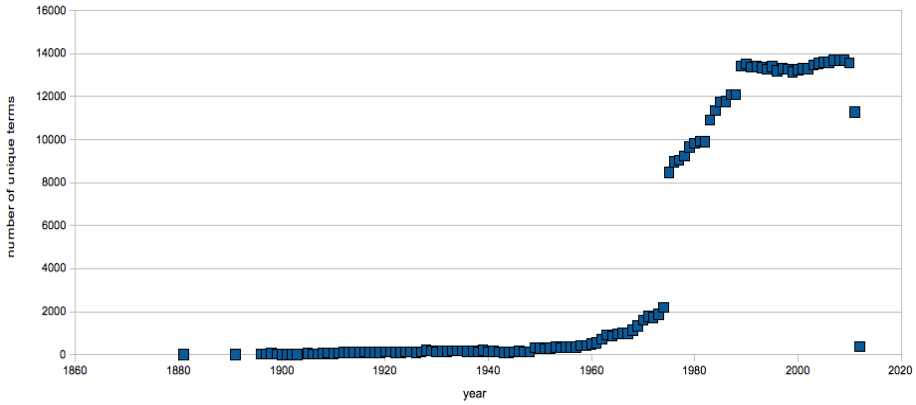

Figure 2. Number of PharmGKB Disease terms observed in NML PubMed citations per year.

The PharmGKB Disease terminology contains over 56,000 terms of which only the average of 14,000 can be found in the PubMed citations each year. The reason for term number appearing to become constant could be that very recent terms are not yet included in the reference terminology which we used for the experiments.

3.1. Terms disappearing

The period of one year is not substantial to judge the situation of usage of disease terms, however longer time frames reveal interesting patterns of how terms appear/disappear in the biomedical literature.

Figure 3. depicts a number of terms from the PharmGKB Disease terminology which have not been used in PubMed citations within the last 100, 50, 40, 30, 20, 10 and 5 years. For instance, there are 510 terms which cannot be found in PubMed literature for the last 20 years. Considering the field of medicine has evolved immensely during the last 20 years, we can be confident that those 510 terms have disappeared or died out. Here are some examples of disease terms which have not been used in the PubMed citations for the last 20 years:

```
fowl coryza
pulmonary distomiasis
infectious bulbar paralysis
acute ascending myelitis
neurolymphomatosi gallinarum
koch week conjunctiviti
fievre boutonneuse
granular lid
subacute yellow atrophy
atypical bronchopneumonia
```

The disease terms listed above are of different granularity: Latin terminology, descriptions of symptoms, animal diseases. Since this particular terminology is edited and updated each year by experts working in the area, we suggest that chronological insight of term usage can be useful in maintaining high quality up-to-date terminologies.

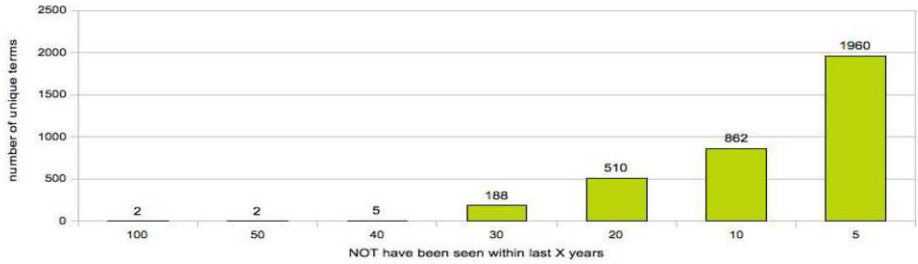

Figure 3. The number of the PharmGKB diseases terms which have not been used in recent years.

3.2. Terms appearing

Along with the terms disappearing new terms are introduced. The number of terms introduced into the biomedical literature per year is depicted in the Figure 4. The lower part of the figure shows a steady number (around 450 on average) of new terms introduced each year, apart from the drop at the tail of the graph – which is due to the time delay for the introduction of new terms into the reference terminology. This results into a linear decline of terms used only during the last years, see the above part of the figure.

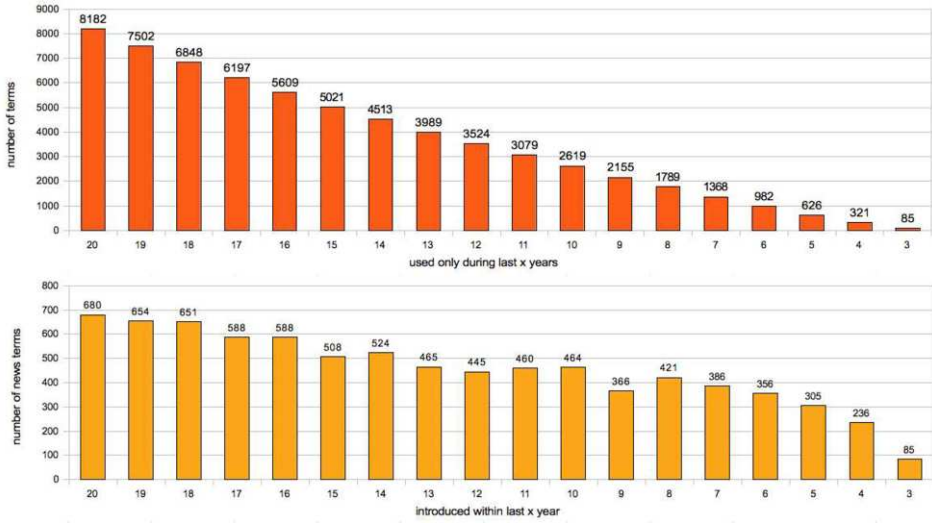

Figure 4. The number of new PharmGKB disease terms which have been introduced/used in recent years.

3.3. Term replacement

One reason why terms disappear in the biomedical domain is that a vaccine or cure to certain diseases is found and they become no longer topical. These phenomena could possibly be pointed out by using chronologically arranged structured domain knowledge like ontologies. The other introspection comes from the field of historical corpus linguistics: according to [13] language changes naturally by lexical borrowing, including more specific and precise terms, synonyms and adopting variants. Thus the

reason for some terms disappearing is the replacement of some terms by their synonyms.

Large data collections are handy for observing a change of a term. One common pattern in terminology change is term replacement: the decline of usage of one term and the increase of usage of its synonym. The following three figures (Figure 5-7) are Google Books N-gram viewer [14] visualizations for the term replacement phenomenon. Figure 5. shows two terms: *hyperpiesia* and its synonym *hypertension*.

Both of these terms are listed in the PharmGKB Disease terminology, however *hyperpiesia* is far less common and the last document including it was published in 1967 (in PubMed). As an example, *hypertension*, the synonym of *hyperpiesia*, has become fully substitute. The second part of the figure reveals the proportional weight of *hypertension* when contrasted to *hyperpiesia*.

A similar case is observed with *pyorrhea alveolaris* disappearing and *periodontal disease* (Figure 6.) being introduced.

Term replacement is not necessarily a pair-wise change. As Figure 7 depicts, *granulopenia* is decreasing in usage and is being replaced by its synonyms *granulocytopenia* and *agranulocytosis*.

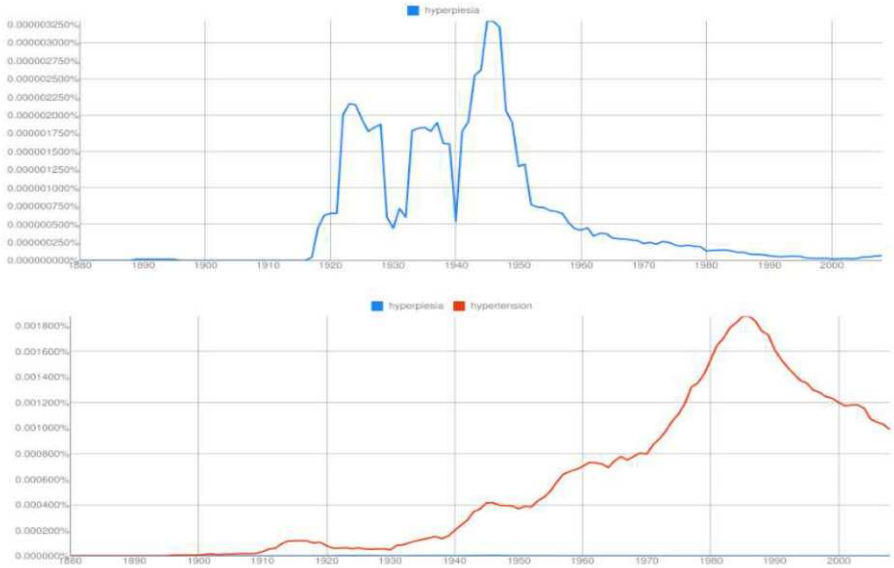

Figure 5. Term change: *hyperpiesia* disappearing, *hypertension* introduced.

Figure 6. Term change: *pyorrhea alveolaris* disappearing, *periodontal disease* introduced.

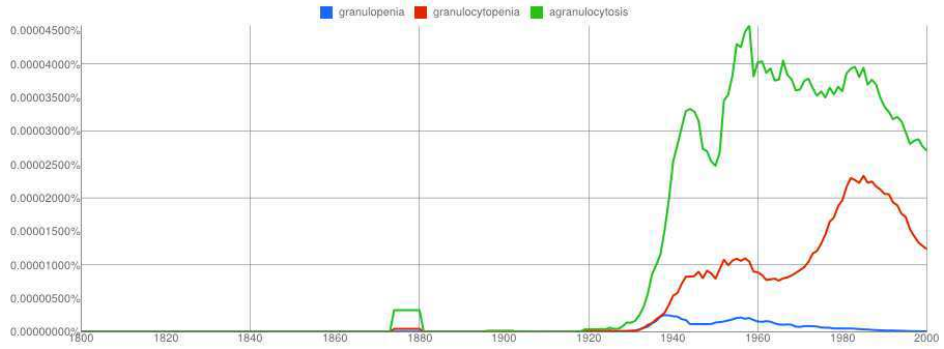

Figure 7. Term change: *granulopenia* disappearing, *granulocytopenia* and *agranulocytosis* introduced.

4. Conclusions

We have presented experimental investigations carried out on a diachronic corpus of the biomedical domain literature capturing how specific diseases terminology changes over time. Terminology change was observed by means of lexical change.

Term usage is dynamic, even in very large volumes of domain texts. i.e. half a million documents within one year, only an average amount of a quarter terms of the entire terminology is being used. On a larger diachronic scale these term occurrence patterns reveal that some terms stop being used and new ones are introduced.

The reason of some terms disappearing is the replacement of a term by its synonyms. A typical example of term replacement is the term *pyorrhea alveolaris* disappearing and *periodontal disease* being introduced. Recognizing such cases of term evolution would be valuable in the process of maintaining a reference terminology such as PharmGKB (where obsolete terms should be marked as such and possibly linked to their modern replacements). Additionally, such information could function as thesaurus enhancement to biomedical IR systems, allowing retrieval of older documents containing obsolete versions of current terms used in a query.

Acknowledgments

This research is funded by the Sciex NMS-CH programme of the Rector's Conference of the Swiss Universities (CRUS). Project Code 11.002.

References

[1] http://www.ncbi.nlm.nih.gov/pubmed/
[2] M. Gorlach. 1991. *Introduction to Early Modern English*. Cambridge University Press.
[3] C. Cowie. 1998. *The Discourse Motivations for Neologising: Action Nominalization in the History of English.* In: Coleman, Julie/ Kay, Christian J. (eds.): Lexicology, Semantics and Lexicography Selected Papers from the Fourth G. L. Brook Symposium, Manchester. 179-206.
[4] I. Taavitsainen and P. Pahta. 2004. *Medical and Scientific Writing in Late Medieval English, Studies in English Language*. Cambridge University Press.
[5] J. Norri. 2004. *Entrances and exits in English medical vocabulary, 1400-1550.* In: Taavitsainen, Irma/Pahta, Paivi (eds.): Medical and Scientific Writing in Late Medieval English. Cambridge University Press, 100-143.
[6] A. Kaluarachchi, A. Varde, S. Bedathur, G. Weikum, J. Peng, A. Feldman. 2010. *Incorporating terminology evolution for query translation in text retrieval with association rules.* In Proceedings of the 19th ACM international conference on Information and knowledge management. 1789-1792.
[7] N. Tahmasebi, G. Zenz, T. Iofciu, T. Risse. 2010. *Terminology Evolution Module for Web Archives in the LiWA Context.* In Proc. of 10th International Web Archiving Workshop (IWAW 2010) in conjunction with iPres 2010, Vienna, Austria.
[8] N. Tahmasebi. 2009. *Automatic Detection of Terminology Evolution.* In Proc. of the Confederated International Workshops and Posters on On the Move to Meaningful Internet Systems, Vilamoura, Portugal. 769-778.
[9] McCray. 2008. *Taxonomic change as a reflection of progress in a scientific discipline*. www.l3s.de/web/upload/talk/mccray-talk.pdf
[10] John Lee. 2009. MEVO: MeSH evolution browser. https://github.com/seouri/mevo
[11] http://www.lemurproject.org/
[12] http://www.pharmgkb.org
[13] Hans H. Hock, Brian D. Joseph. *Language History, Language Change, and Language Relationship.* 1996, Mouton de Gruyter, Berlin.
[14] http://books.google.com/ngrams