SCAN: A Swedish Clinical Abbreviation Normalizer

Further Development and Adaptation to Radiology

Maria Kvist[1,2] and Sumithra Velupillai[1]

[1] Dept. of Computer and Systems Sciences (DSV)
Stockholm University, Forum 100, SE-164 40 Kista, Sweden
[2] Department of Learning, Informatics, Management and Ethics (LIME)
Karolinska Institutet, Sweden
sumithra@dsv.su.se, maria.kvist@karolinska.se

Abstract. Abbreviations pose a challenge for information extraction systems. In clinical text, abbreviations are abundant, as this type of documentation is written under time-pressure. We report work on characterizing abbreviations in Swedish clinical text and the development of SCAN: a Swedish Clinical Abbreviation Normalizer, which is built for the purpose of improving information access systems in the clinical domain. The clinical domain includes several subdomains with differing vocabularies depending on the nature of the specialist work, and adaption of NLP-tools may consequently be necessary. We extend and adapt SCAN, and evaluate on two different clinical subdomains: emergency department (ED) and radiology (X-ray). Overall final results are 85% (ED) and 83% (X-ray) F1-measure on the task of abbreviation identification. We also evaluate coverage of abbreviation expansion candidates in existing lexical resources, and create two new, freely available, lexicons with abbreviations and their possible expansions for the two clinical subdomains.

1 Introduction

Access to information is crucial in the clinical domain. In health care, the main form of written communication is in narrative form. Today, most clinical texts are written in Electronic Health Records (EHRs). Accessing information from this type of text requires automated solutions, for instance by Natural Language Processing (NLP) tools.

Clinical text is often written as short telegraphic messages under time-pressure, as memory notes for the healthcare team. Subjects, verbs, and content words are frequently omitted, but the text has a high proportion of technical terms [8]. However, there are more formal parts of the records, such as discharge letters and radiology reports, that are communications to another physician. These parts of the EHR may be written with more complete sentences. Abbreviations and acronyms are frequently used in both the formal and informally written parts of the EHR.

E. Kanoulas et al. (Eds.): CLEF 2014, LNCS 8685, pp. 62–73, 2014.

For information extraction, it is necessary to normalize abbreviations by a multistep procedure of detecting an abbreviation, expanding it to its long form and, when necessary, disambiguate.

1.1 Related Work

Abbreviation detection in the clinical domain is associated with special difficulties as many of the normal standards for abbreviation creation are set apart and the full form of the word or expression is rarely present or explained.

Abbreviations in Clinical Text. Abbreviations and acronyms in EHRs are often domain specific but can also belong to general language use [13,26]. There are established standard acronyms that can be found in medical terminologies, but often abbreviations are created ad hoc, not following standards, and may be ambiguous. An abbreviation can be used with a number of different meanings depending on context [13,18,14]. For example, the abbreviation RA can represent more than 20 concepts, e.g. renal artery, right atrium, refractory anemia, radioactive, right arm, and rheumatoid arthritis [18]. In the Unified Medical Language System (UMLS[1]), 33% of abbreviations had multiple meanings [13]. Furthermore, a certain word or expression can be shortened in several different ways, some of which mimic ordinary words [13]. These meanings can depend on specialty or profession [14].

Clinical texts differ between specialties, as the vocabulary reflects the nature of diagnoses, examinations and the type of work performed, as well as the temperament of the speciality. Hence, an NLP-tool developed in one subdomain may drop in performance when applied on text from another subdomain. The clinical text in radiology reports has been characterized and differences between this sublanguage, the language in physicians' daily notes and general Swedish have been studied [11,22]. Text from the clinical domain contained more abbreviations than general Swedish, both clinical subdomains contained around 8% abbreviations. Moreover, other higher-level language aspects pose challenges for adapting NLP-tools for the clinical domain, e.g. 63% of all sentences in Swedish radiology reports lack a main predicate (verb) [22].

It has been noted that abbreviations are more common for frequently used expressions and multiword expressions, and it has been found that 14% of diagnostic expressions in Swedish clinical text are abbreviated [21]. Some attempts have been made to capture the full form of words and pair with their abbreviations with distributional semantics in Swedish clinical texts [9,23].

Terminological Resources. Although there exist terminologies like the UMLS for English, that also covers medical and clinical abbreviations, there are currently no terminologies or lexicons that have full coverage of clinical abbreviations found in clinical notes, and their possible expansions [24]. Similarly

[1] `http://www.nlm.nih.gov/research/umls/`

for Swedish, there is one comprehensive lexicon of medical abbreviations and acronyms [4], as well as scattered online resources - but no resources that handle a majority of the abbreviation variants that could be found in clinical notes.

Abbreviation Normalizing Tools. For English, several tools for clinical NLP, including abbreviation handling, have been developed, such as MetaMap [1,2], MedLee [7], and cTakes [20]. However, a study by Wu et al. [24] showed that these systems did not perform well on the abbreviation detection and expansion tasks when applied to new data. Moreover, results from a recent shared task on abbreviation normalization [15,16] for English clinical text showed that automatic mapping of abbreviations to concept identifiers is not trivial. Hence, this is still a challenging task for improved information extraction and access in the clinical domain. Furthermore, most previous work is done for English. To our knowledge, no tools exist for Swedish clinical text.

This work is an extension of our earlier work [10] with a system for identification and expansion of abbreviations in clinical texts called Swedish Clinical Abbreviation Normalizer (SCAN). The system is rule-, heuristics- and lexicon-based, inspired by previous approaches taken for English and Swedish [26,27,19,12,6]. SCAN relies on word character lengths, heuristics for handling patterns such as hyphenation, and lexicons of known common words, abbreviations and medical terms for identifying abbreviation candidates in a corpus. The system was initially evaluated for the task of *identifying* abbreviations in Swedish clinical assessment entries from an emergency department. It's best performance was reported as 79% F1-measure (76% recall, 81% precision), a considerable improvement over a baseline where words were checked against lexicons only (51% F1-measure). In this work, we extend the evaluation to another clinical subdomain (radiology), and initiate evaluation on abbreviation expansion.

1.2 Aim and Objective

Our aim is to improve automated information access and information extraction from clinical text in EHR, for e.g. decision support systems and patients' access to reading their own records. The objective of this study is to characterise abbreviations in Swedish clinical text, improve the performance of SCAN, adapt it for the clinical sublanguage of radiologic reports, and to advance work on abbreviation expansion for Swedish clinical text by creating new lexical resources.

2 Method

This study consisted of three main steps: 1) data collection, analysis and characterization of abbreviation types and reference standard creation, 2) iterative development of SCAN, and 3) evaluation of system outputs, coverage analysis of expansion candidates in existing lexical resources along with new lexicon creation. These steps are further described below.

2.1 Data and Content Analysis

For each iteration (n=3) in the development of SCAN, we used subsets from the Stockholm EPR Corpus[2] [5]: 3×10 000 words from randomly selected assessment entries from an emergency department (ED), and 3×10 000 words from randomly selected radiology reports (X-ray). All notes are written in Swedish, and written by physicians. Entire notes with preserved context were used. Each subset was manually annotated for abbreviations by a domain expert (a clinician, MK), resulting in 3×2 reference standards (ED, X-ray). Furthermore, we performed a content analysis on abbreviations found in both text subtypes, resulting in a characterisation of types of abbreviations found in Swedish ED and X-ray notes.

2.2 SCAN: Iterative Development

We employed an iterative development of SCAN; an error analysis of the output from the first SCAN version (SCAN 1.0) was performed in order to identify common error types. New versions were subsequently developed based on the identified modification needs found through this error analysis, ending in a final version (SCAN 2.0). Three SCAN versions were evaluated on the created reference standards.

2.3 Evaluation: System Results, Expansion Coverage and Lexicon Creation

System results were evaluated with precision, recall and F1-measure as the main outcome measures on the held-out datasets, to approximate system performance on unseen data. We also evaluated the coverage (%) of abbreviation expansion candidates in the provided lexicons, and produced a lexicon with abbreviations and expansions for each clinical subtype (ED, X-ray) based on the actual abbreviations found in the datasets. Furthermore, we performed an extensive error analysis on the results from the first and the final iteration of SCAN, performed by a physician.

3 Results

We report the results from the content analysis and characterisation of the types of abbreviations found in the studied clinical subtypes ED and X-ray. The error analysis of SCAN 1.0 and the iterative development of SCAN 2.0 is described, followed by the abbreviation identification results from the three versions of SCAN. Finally, we report the coverage analysis of abbreviation expansion candidates from the provided lexicons, and describe the resulting lexicons.

3.1 Content Analysis and Characterization

The types of abbreviations are shown in Table 1. Of the abbreviated words, a fraction of 12% were part of compound words, consistently for both text types.

[2] This research was approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2012/2028-31/5.

Table 1. Characterization of abbreviations in EHR assessment fields from an emergency department (ED) and from radiology reports (X-ray). Numbers were calculated as averages of three datasets with 10 000 words in each set.

Abbreviation type	**ED (%)**	**X-ray (%)**
Abbreviations, total	11	7,1
Of these abbreviations:		
Acronyms	37	62
Shortened words or contractions	63	38
Compounds with abbreviation	12	12

For compound words, it was very common that both parts of the word were abbreviated, e.g. *jnlant* (*journalanteckning* eng: record note). The abbreviated words were more often of the type shortened words (*pat* for *patient*) or contractions (*ssk* for *sjuksköterska*, eng: *nurse*) than acronyms (*ECG* for *electrocardiogram*) for the texts from the emergency department, with the proportions 63:37. For the radiology reports, the reverse was consistently seen as average in three sets of 10 000 words, with the proportions 38:62.

The content analysis (Figure 1) reflects the different tasks for physicians at the two respective departments. At the emergency department, the text is a narrative of events such as examinations, blood sampling and resulting lab results, prescribing medication and consultations with other physicians of various specialities. Administrative words denote different hospitals or wards. For radiologists, the task when writing is more descriptive of examinations and findings in the resulting images. A distinctive difference is that the patients are mentioned as a subject (abbr *pat*) at the emergency department whereas the patient is not mentioned as the subject in the radiology reports. In the emergency department assessment entries, the abbreviation of *pat*, *pats* (plural) is so extensive that this singular abbreviation make up 22% of the total number of abbreviations. Abbreviations for medications were for doses (ml=milliliter) or injections (i.v.= intravenous) but rarely of the medicine names or chemical compounds. Of all abbreviations, diagnostic expressions made up 5,2%. On the other hand, 14% of all diagnostic expressions were abbreviated. This is consistent with the findings of Skeppstedt et al. (2012) [21], where another subset of emergency department assessment entries from the Stockholm EPR Corpus was studied.

3.2 Error Analysis of SCAN 1.0 and Development of SCAN 2.0

The error analysis of SCAN 1.0 revealed that names, missing terminology and tokenization were the main sources of errors (Table 2). Names (people and locations) constituted the majority of errors (54%) as these were identified as unknown words and hence the names shorter than 6 characters were identified as abbreviations. Terms that were missing in lexicons were also mistakenly labeled as abbreviations (21%). Incorrect tokenization, making abbreviations

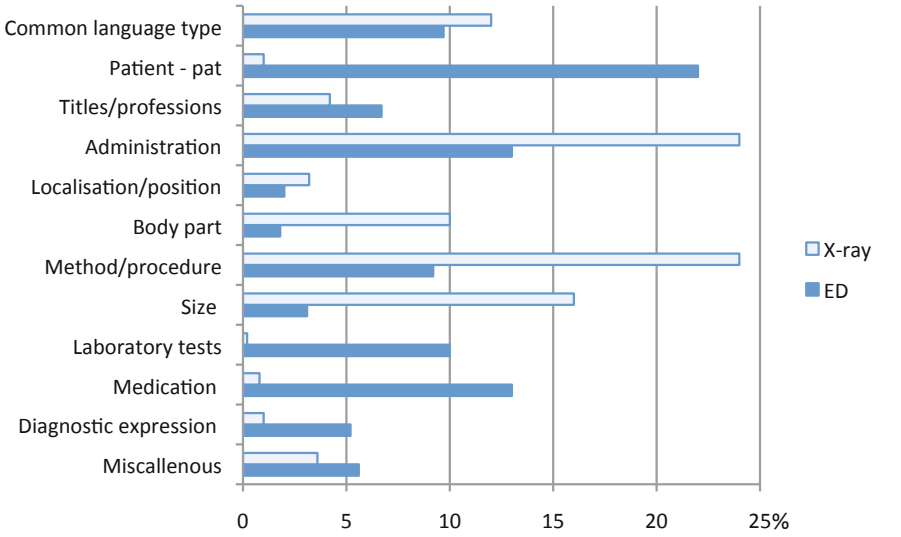

Fig. 1. Content analysis for abbreviations in Swedish emergency department assessment entries (ED) and radiology notes (X-ray). Numbers were calculated as averages of 3 datasets with 10 000 words in each set. For the category "Administration", during the iterations in this work, it was found that some administrative information in the end of the radiology texts was redundant and was subsequently removed (proportion dropped from 31% to 10%).

undetectable by e.g. breaking up an abbreviation into two different tokens after a punctuation character, made up 19% of the errors. Only 6% of the errors were due to ambiguity with common words, e.g. *hö*, abbreviation for *höger* (right) can also mean "hay", and was therefore not identified as an abbreviation.

When developing SCAN 2.0, we took these issues into consideration by modifying the following parts:

1. Tokenization: instead of tokenising with in-built heuristics and regular expressions, an existing tool for tokenization and Part-of-Speech (PoS) tagging developed for general Swedish was used (Stagger [17]). In order to better handle domain specific clinical abbreviations, the tokenization rules in Stagger were extended and modified to better handle domain-specific abbreviations found in the error analysis, e.g. handling different instantiations of the abbreviation *vb* (*vid behov* eng: when needed).
2. Names and missing terminology: we added several lexicons for better coverage. Lists of names (first and last) were added from Carlsson & Dalianis [3], totalling 404,899 lexicon entries. Furthermore, in addition to the lexicons used for SCAN 1.0, freely available online lexical resources were added to

handle a) medical terms[3] (17,380 entries in total), b) known abbreviations[4] (7,455 entries in total), and common words[5] (122,847 entries in total).

Moreover, in addition to the abbreviation detection heuristics from SCAN 1.0, we added heuristics to exploit the PoS tags produced by Stagger, e.g. punctuation PoS-tags were used to exclude tokens in the analysis and words PoS-tagged as abbreviations were marked as such in the output. We also set the length of a candidate word to six characters, based on the finding that setting the length too short (i.e. three characters) decreases precision, and setting it too long decreases recall (i.e. eight characters) [10].

Table 2. Error analysis: SCAN 1.0 output on Swedish radiology reports, characterization of false positives

Error type	%
Names (people and locations)	54
Missing terminology in lexicons	21
Tokenization	19
Common words	6
$\sum$	100

3.3 Abbreviation Identification

The abbreviation identification results from three versions of SCAN (1.0, 1.5 and 2.0) are shown in Table 3. SCAN 1.0 is the original version of SCAN. The new tokenization as well as added and modified heuristics are used in SCAN 1.5 and 2.0. SCAN 1.5 uses the same lexicons as SCAN 1.0, while SCAN 2.0 also uses additional lexicons. The tokenization changes clearly leads to performance improvements, in particular for the X-ray data (from 61% to 83% F1-measure). Precision results are best when using SCAN 2.0, with the largest improvement observed for X-ray (from 66% using SCAN 1.5 to 78% using SCAN 2.0). Adding new lexicons does not improve recall (83% vs. 80% for ED, 92% vs. 89% for X-ray), but overall results are improved with the added lexicons (SCAN 2.0).

The false positives produced by SCAN 2.0 include medical terminology, e.g. *flavum*, misspellings, e.g. *västka* (*vätska*, eng: fluid), and unusual person names.

[3] Downloaded from: `anatomin.se`, `neuro.ki.se smittskyddsinstitutet.se`, `medicinskordbok.se`

[4] Resulting abbreviations from the error analysis, along with entries downloaded from `sv.wikipedia.org/wiki/Lista_över_förkortningar`, `karolinska.se/Karolinska-Universitetslaboratoriet/Sidor-om-PTA/Analysindex-alla-enheter/Forkortningar/`

[5] Downloaded from `runeberg.org`, `g3.spraakdata.gu.se/saob`,

Table 3. Performance of SCAN 1.0, SCAN 1.5 and SCAN 2.0, evaluated with precision, recall and F1-measure. SCAN 1.0 = original SCAN, as reported in [10]. SCAN 1.5 = improved SCAN (new tokenization, added and modified heuristics), but using the same lexicons as SCAN 1.0. SCAN 2.0 = improved SCAN plus added lexicons.

	SCAN 1.0		SCAN 1.5		SCAN 2.0	
	ED	X-ray	ED	X-ray	ED	X-ray
recall	0.79	0.83	**0.83**	**0.92**	0.80	0.89
precision	0.81	0.48	0.85	0.66	**0.92**	**0.78**
F1-measure	0.80	0.61	0.84	0.77	**0.85**	**0.83**

False negatives include compounds, e.g. *lungrtgbilder* (*lung-röntgen-bilder*, eng: lung x-ray images) and ambiguous words, e.g. *sin* (could mean "his" or "her" as well as "sinister", Latin for left side).

3.4 Abbreviation Expansion Coverage Analysis and Lexicons

For both record types, a majority of the correct expansions are present in the lexicons[6] (Table 4): 79% for emergency department assessment entries (ED), and 60% for radiology notes (X-ray). However, for radiology, there were more cases where there were no suggestions for expansions in the lexicons (32%). Moreover, in many cases where the correct expansion was present in lexicons, there were many possible expansion candidates. As a result of this analysis, a comprehensive lexicon of abbreviations and their correct expansion in its context was created for each clinical subset (ED, X-ray). Some abbreviations were found in several different typographic variants, e.g. *ua*, *u.a.*, *u a*, *u. a.* (*utan anmärkning*, eng: without remark). Moreover, abbreviations could in some cases be expanded to an inflected form, e.g. *us - undersökning* (examination) or *undersökningen* (the examination). The resulting lexicons from this analysis include all typographic variants and expansion inflections found in the data[7].

Table 4. Coverage analysis, abbreviation expansions in lexicons. Results from the evaluation of SCAN 2.0 on 10 000 words each of the two datasets are shown.

Coverage type	ED (%)	X-ray (%)
Correct expansion in lexicons	79	60
Missing the correct expansion in lexicons	8	8
No suggestion for expansion in lexicons	13	32

[6] Note that one of the lexicons is the result of the analysis in one of the SCAN development iterations.

[7] *All* possible inflections for an abbreviation expansion are *not* included in the lexicons.

4 Analysis and Discussion

We characterized the abbreviations in two subsets of Swedish clinical text and further developed an abbreviation normalizer for Swedish clinical text, SCAN. We also adapted SCAN to the new sublanguage radiologic reports (X-ray) in addition to the previous development for emergency department assessment entries (ED). Our characterisation analysis shows that some abbreviations are from the general language but around 90% of the abbreviations are unique for the domain. The type of abbreviations differ between the subdomains; ED notes contain more references to the patient, medications and laboratory tests, while radiology reports contain abbreviations about methods/procedures and sizes. Acronyms are more prevalent in radiology reports, while shortened forms and contractions are more common in emergency department notes. This information could be informative features in future abbreviation detection systems.

Overall results of SCAN 2.0 are improved on ED data when compared to SCAN 1.0: 0.85 F1-measure as opposed to the initial 0.79. This was mainly due to high precision (0.92) with more extensive lexicons and improved tokenization. On X-ray data, both precision and recall was improved with the largest improvement seen on precision (from 0.48 to 0.78) and we obtained 0.83 F1-measure for SCAN 2.0. Compared to results for English clinical text, our system still has room for improvement. Excellent results (95.7% F1-measure) have been reached by Wu et al. [25] with a combination of machine learning techniques. In the 2013 ShARe/CLEFeHealth task 2 for abbreviation normalization [15], the top-performing system resulted in an Accuracy of 0.72. However, the task did not include the abbreviation detection part, i.e. this was only for normalizing a given abbreviation to its UMLS concept identifier. To our knowledge, there are no available abbreviation detection and/or normalizing tools for Swedish clinical text to which we could compare our results. For Swedish biomedical scientific text, there are results on acronym identification reaching 98% recall and 94% precision [6]. However, as mentioned previously, clinical text differs greatly from other types of texts, in particular the way abbreviations are used and created.

The coverage analysis revealed that existing Swedish lexical resources contain the majority of correct expansions (79%/61% ED/X-ray), but clearly more comprehensive resources are needed. Most importantly, many abbreviations found in the data are missing altogether in existing lexicons.

Part of our aim was to produce new resources. We have created two reference standards with abbreviation annotations that can be used for further studies on abbreviation detection in Swedish clinical text[8]. Moreover, we have created two lexical resources with abbreviations and their expansions as found in the clinical data (ED and X-ray), that are freely available[9].

[8] These datasets are available for research purposes upon request, although constrained by obtaining appropriate ethical approval and signing confidentiality agreements.

[9] Please contact the authors for access to the lexical resources.

4.1 Limitations and Future Work

This study has some limitations. Although we have created reference standards, they were only annotated by one annotator. SCAN is rule-based and relies on lexicons and heuristics. We depend on existing lexical resources, but do not claim to have included all possible available resources. SCAN was only evaluated with a word length of six characters. We intend to utilise the reference standards to build machine learning classifiers which we believe will lead to improved results for abbreviation detection. Moreover, for abbreviation normalization, we have only evaluated coverage in existing lexicons, and created new lexicons manually. We intend to extend our work also to include abbreviation normalization. Finally, a limitation with our rule-based approach is disambiguation, which is one of the sources for false negatives. Our plan is to extend our work to larger datasets and develop probabilistic classifiers for disambiguating words that could be either an abbreviation or a common word, as well as disambiguating abbreviation expansions when there are multiple candidates. Furthermore, we intend to map Swedish abbreviations to existing terminologies such as SNOMED CT, for future cross-language interoperability.

4.2 Significance of Study

To our knowledge, this is the first in-depth study on automatic detection of abbreviations in Swedish clinical text, which also covers two sublanguages (emergency department notes, radiology notes). Our tool, SCAN, is freely available upon request. The created lexicons are being made available online. These resources are a significant contribution to the research community, as they will enable other researchers to work on abbreviation detection and normalization in the Swedish clinical domain. Also, abbreviation detection with machine-learning is facilitated with the new reference standard.

5 Conclusions

In this study, we have successfully characterised abbreviations in two subdomains of Swedish clinical text and used the results to improve detection of abbreviations and acronyms with SCAN, resulting in an overall F1-measure of 0.85 for emergency department assessment entries and 0.83 for radiology notes. Further, we have created lexicons with abbreviations and their expansions as found in ED and X-ray reports, and we have evaluated the coverage of correct expansions found in existing lexical resources. For abbreviation normalization in clinical text, it is essential to understand the many irregular ways of ad hoc creativity in abbreviation generation and work closely with domain experts.

Acknowledgments. The study was partly funded by the Vårdal Foundation and Swedish Research Council (350-2012-6658), and supported by Swedish Fulbright Commission and the Swedish Foundation for Strategic Research through the project High-Performance Data Mining for Drug Effect Detection (ref. no. IIS11-0053).

References

1. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proc. AMIA Symp., pp. 17–21 (2001)
2. Aronson, A., Lang, F.M.: An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association: JAMIA 17(3), 229–236 (2010)
3. Carlsson, E., Dalianis, H.: Influence of Module Order on Rule-Based De-identification of Personal Names in Electronic Patient Records Written in Swedish. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010, Valletta, Malta, May 19–21, pp. 3071–3075 (2010)
4. Cederblom, S.: Medicinska förkortningar och akronymer. Studentlitteratur (2005) (in Swedish)
5. Dalianis, H., Hassel, M., Henriksson, A., Skeppstedt, M.: Stockholm EPR Corpus: A Clinical Database Used to Improve Health Care. In: Nugues, P. (ed.) Proc. 4th SLTC, Lund, October 25–26, pp. 17–18 (2012), http://nlp.lacasahassel.net/publications/dalianisetal12sltc.pdf
6. Dannélls, D.: Automatic acronym recognition. In: Proceedings of the 11th Conference on European Chapter of the Association for Computational Linguistics, EACL (2006)
7. Friedman, C., Alderson, P.O., Austin, J.H., Cimino, J.J., Johnson, S.B.: A general natural-language text processor for clinical radiology. J. Am. Med. Inform. Assoc. 1(2), 161–174 (1994)
8. Friedman, C., Kra, P., Rzhetsky, A.: Two biomedical sublanguages: a description based on the theories of Zellig Harris. Journal of Biomedical Informatics 35(4), 222–235 (2002)
9. Henriksson, A., Moen, H., Skeppstedt, M., Daudaravicius, V., Duneld, M.: Synonym Extraction and Abbreviation Expansion with Ensembles of Semantic Spaces. Journal of Biomedical Semantics 5, 6 (2014)
10. Isenius, N., Velupillai, S., Kvist, M.: Initial Results in the Development of SCAN: a Swedish Clinical Abbreviation Normalizer. In: Proceedings of the CLEF 2012 Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis - CLEFeHealth2012. CLEF, Rome, Italy (September 2012)
11. Kvist, M., Velupillai, S.: Professional Language in Swedish Radiology Reports – Characterization for Patient-Adapted Text Simplification. In: Proceedings of the Scandinavian Conference on Health Informatics 2013, Linköping University Electronic Press, Linköpings Universitet, Copenhagen, Denmark (2013), http://www.ep.liu.se/ecp/091/012/ecp13091012.pdf
12. Larkey, L., Ogilvie, P., Price, M., Tamilio, B.: Acrophile: An Automated Acronym Extractor and Server. In: Proceedings of the Fifth ACM Conference on Digital Libraries, pp. 205–214 (2000)
13. Liu, H., Lussier, Y.A., Friedman, C.: Disambiguating Ambiguous Biomedical Terms in Biomedical Narrative Text: An Unsupervised Method. Journal of Biomedical Informatics 34, 249–261 (2001)
14. Lövestam, E., Velupillai, S., Kvist, M.: Abbreviations in Swedish Clinical Text – use by three professions. In: Proc. MIE 2014 (to be presented, 2014)
15. Mowery, D., South, B., Christensen, L., Murtola, L.M., Salanterä, S., Suominen, S., Martinez, D., Elhadad, N., Pradhan, S., Savova, G., Chapman, W.: Task 2: Share/clef ehealth evaluation lab 2013 (2013), http://www.clef-initiative.eu/documents/71612/599e4736-2667-4f59-9ccb-ab5178cae3c5

16. Mowery, D.L., South, B.R., Leng, J., Murtola, L., Danielsson-Ojala, R., Salanterä, S., Chapman, W.: Creating a Reference Standard of Acronym and Abbreviation Annotations for the ShARe/CLEF eHealth Challenge 2013. In: AMIA Annu. Symp. Proc. (2013)
17. Östling, R.: Stagger: an Open-Source Part of Speech Tagger for Swedish. Northern European Journal of Language Technology 3, 1–18 (2013)
18. Pakhomov, S., Pedersen, T., Chute, C.G.: Abbreviation and Acronym Disambiguation in Clinical Discourse. In: Proc. AMIA 2005, pp. 589–593 (2005)
19. Park, Y., Byrd, R.: Hybrid text mining for finding abbreviations and their definitions. In: Proceedings of Empirical Methods in Natural Language Processing, pp. 126–133 (2001)
20. Savova, G., Masanz, J., Ogren, P., Zheng, J., Sohn, S., Kipper-Schuler, K., Chute, C.: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association 17(5), 507–513 (2010)
21. Skeppstedt, M., Kvist, M., Dalianis, H.: Rule-based Entity Recognition and Coverage of SNOMED CT in Swedish Clinical Text. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23–25, pp. 1250–1257 (2012)
22. Smith, K.: Treating a case of the mumbo jumbos: What linguistic features characterize Swedish electronic health records? Master's thesis, Dept. of Linguistics and Philology, Uppsala University, Sweden (2014)
23. Tengstrand, L., Megyesi, B., Henriksson, A., Duneld, M., Kvist, M.: Eacl - expansion of abbreviations in clinical text. In: Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR), pp. 94–103. Association for Computational Linguistics, Gothenburg (2014), http://www.aclweb.org/anthology/W14-1211
24. Wu, Y., Denny, J.C., Rosenbloom, S.T., Miller, R.A., Giuse, D.A., Xu, H.: A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries. In: AMIA Annu. Symp. Proc., pp. 997–1003 (2012)
25. Wu, Y., Rosenbloom, S.T., Denny, J.C., Miller, R.A., Mani, S., Giuse, D.A., Xu, H.: Detecting Abbreviations in Discharge Summaries using Machine Learning Methods. In: AMIA Annu. Symp. Proc., pp. 1541–1549 (2011)
26. Xu, H., Stetson, P.D., Friedman, C.: A Study of Abbreviations in Clinical Notes. In: Proc. AMIA 2007, pp. 821–825 (2007)
27. Yeates, S.: Automatic Extraction of Acronyms from Text. In: Proc. Third New Zealand Computer Science Research Students' Conference, pp. 117–124 (1999)