Grammar-Based Lexicon Extension for Aligning German Radiology Text and Images

Claudia Bretschneider[1,2]
[1]Center for Information and Language Processing, University Munich
claudia.bretschneider.ext@siemens.com

Sonja Zillner[2]
[2]Corporate Research, Siemens AG
sonja.zillner@siemens.com

Matthias Hammon[3]
[3]Department of Radiology, University Hospital Erlangen
matthias.hammon@uk-erlangen.de

Abstract

For efficient diagnosis processes, the multitude of heterogeneous medical data requires seamless integration. In order to automatically align radiology reports and images based on the pathological anatomical entities they describe, a preceding sentence classification is necessary. However, the lexical resource used has to contain semantic information about the pathological classification of each entity. We introduce an approach to extend medical lexical resources with pathology classification information and, at the same time, with new classified vocabulary. Our algorithm is based on a semi-supervised learning algorithm and incorporates a semantic context-free grammar combined with a RadLex-based lexicon.

1 Introduction

In radiology, the health status of a patient is described using a multitude of formats. During the examination process, a radiologist creates machine readable descriptions such as radiology images, dictated reports about the image findings and written texts. Although, most of the radiology data are related via the anatomical entities shown or described, there is no link between them, since the information pieces are stored in distributed systems. This absence of links between the items is hindering the radiologist's workflow. Especially when reading reports, radiologists want to reference back from the described finding (in the text) to the correlating body location (in the images). Without automatically created links, this resolution is obviously time-consuming when dealing with images taken with modalities that deliver a mass of stacked images.

Today, radiologists add alignment information to the text that names the image that contains the described findings. But still, the resolution of these textual links requires manual interventions to find the correct image and detect the described finding in the image.

To simplify this workflow, we introduce a mechanism that automatically aligns pathological anatomical entities in radiology text and images based on semantic annotations. Figure 1 shows our concept of linking anatomical concepts from image and text: Both the images and the texts are annotated with the anatomical concepts that they describe. Combining annotations with the same RadLex ID (RID), the link from one format to the other can be established. As a result, the radiologist can easily navigate from the pathological *Leber* [liver] (RID58) described in the text to the correlating position in the images.

For the integration, the necessary semantic annotations of the images have been made available as a result of a previous project (Seifert et. al., 2009; Seifert, 2010). In order to align these RadLex-based annotations with anatomical entities described in radiology reports, our text analysis system has to annotate the texts with RadLex-based annotations, too. Our established mechanism operates in two steps: First, we identify the relevant sentences that describe pathological findings and, second, extract the anatomical annotations only from these sentence.

We include a preceding sentence classification step, because according to the radiologists we worked with, the extraction of *all* anatomical entities from the text to link them with the image annotations is inappropriate. A large portion of the findings is included in the reports in order to exclude differential diagnoses. These are normal or absent findings that do not describe pathologies. But radiologists are rather interested in automated alignment of images of anatomical entities described with pathological findings.

The sentence classification is conducted based

Proceedings of Recent Advances in Natural Language Processing, pages 105–112,
Hissar, Bulgaria, 7-13 September 2013.

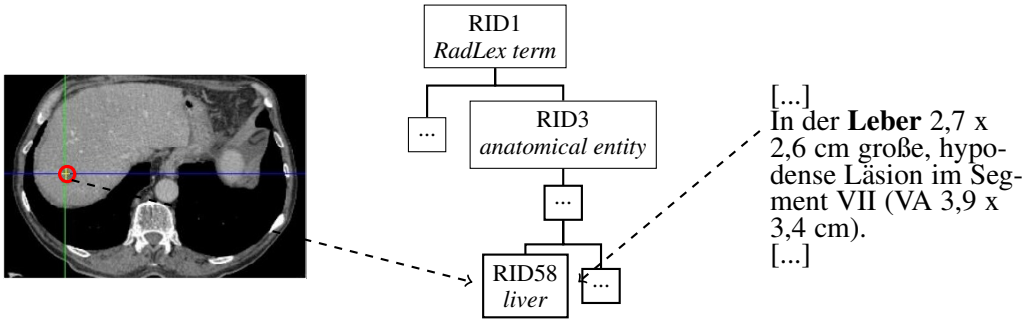

Figure 1: Aligning the anatomical concept *liver* from radiology text to image using RadLex-based annotations

on a lexicon and probabilistic semantic grammar rules (P-CFG). For parsing, we apply the standard probabilistic CKY algorithm (Kasami, 1965). During parsing, the most likely parse tree for the given sentence is determined. The topmost constituent in the resulting parse tree can be used to determine the pathology classification of the report sentences.

The chosen approach requires a full coverage lexicon including pathology classification of the entities. An initial linguistic resource based on the German RadLex taxonomy is provided. However, the German RadLex is lacking in terminology and pathology classification. The contribution of this paper is the description of a process to extend the German RadLex-based lexicon with vocabulary and pathology classification information in order to link heterogeneous medical data sources.

2 Related work

Medical grammar-based text analysis systems Theoretical work on the linguistic characteristics of the medical sublanguage has been conducted on the adaption of theories of Harris by (Friedman et. al., 2002). Early systems of (Sager et. al., 1994; Friedman et. al., 1994) are adaptations of the theories and implement own (context-free) medical language grammar for radiology reports. They show that parsing of medical texts based on a combined semantic-syntactic grammar can be successfully conducted. Even today, advances in grammar-based parsing of medical texts are reached (Fan et. al., 2011).

More recently, semantic text analysis systems have integrated the idea of parsing for medical text understanding for more sophisticated information extraction tasks (Savova et. al., 2010).

All those systems work with the advantage of elaborated lexicons that fully cover the vocabulary used in English report.

Terminology acquisition and semantic classification Semantic classifications beyond the hypernym information of taxonomies are still rare. Several approaches address this lack: Corpus-based approaches based on statistical analyses about the coverage and frequency of UMLS ontology concepts (Liu et al., 2012; Wu et al., 2012). (Johnson, 1999) derives semantic classes from ontology mapping and disambiguates multiple senses in contexts of discharge summaries. Limited to noun phrases, (Campbell et al., 1999) applies pattern-based rules and combines them with UMLS concepts to acquire new and semantically classified terminology. Finally, (Zweigenbaum et al., 2003) introduce a statistical approaches to automatically extending the UMLS ontology with French concepts.

Gap analysis While the grammar-based analysis of radiology reports has shown to be successful with complete lexical resources, we have to face the shortcomings of an incomplete lexicon. Furthermore, in other systems the grammar is used as mean for syntactic analysis of the content of the reports. Our approach to use it for pathology classification is novel and has not been applied so far.

Working with German clinical texts is another challenge in the field. English texts have been made available by a number of shared tasks and gained more and more interest in the last decade. Medical corpora in languages other than English are not available to that extend. At the same time, German language versions of medical ontologies are rare. Semantic classifications such as pathological information are particularly missing so far.

3 Corpus analysis

Our semi-supervised learning approach relies on a reference corpus, whose features are described shortly in the following section.

3.1 Reference corpus and development set

Since a publicly available corpus of German radiology reports is missing, we build our own annotated corpus. Our clinical partner, the University Hospital Erlangen, allocates the necessary texts: 2713 de-identified reports spanning the period from April 2002 until July 2007.

From this corpus, we selected 174 representative reports for a development set. Based on the findings described in the sentence, a radiologist classified each sentence. Sentences describing normal or absent findings are classified as 'non-pathological' and those containing descriptions of abnormalities are classified as 'pathological'.

3.2 Syntactic characteristics

One of the most apparent syntactic characteristics of the reports is their telegraphic style. The texts are rich in omission of verbs; the verbs are dispensable as they do not add semantics to the sentences. They are used to underline the absence or presence of symptoms - but are not necessary. Instead of noting

> *In der Lunge sind keine Ergüsse zu finden.* [In the lung, there are no effusions available.]

radiologists simply state

> *Lunge: Kein Erguss.* [Lung: No effusions.]

The average sentence length listed in Table 1 underline this finding.

3.3 Statistical characteristics

We annotated 4295 sentences in the development set of which less than half are classified as 'pathological'. This ratio is in line with the radiologists' experience. Table 1 shows further results of the statistical corpus analysis.

From comparing the numbers of word types, we conclude that the description of pathological findings requires a richer language than those of normal states and absent findings in non-pathological sentences. The linguistic resource has to cover this richness, which means that the multitude of entities should be classified as describing pathological findings.

Corpus characteristic	Sentence class *PATH*	*NOPATH*
Sentences	1,943	2,352
Tokens	16,437	11,572
Average sentence length	8.46	4.92
Word types	2,398	1,581

Table 1: Results of statistical analysis of the development set

4 Analysis of controlled vocabulary in RadLex

Furthermore, we use the vocabulary from the German RadLex taxonomy as initial linguistic input. What information is already available is analyzed in the following section.

4.1 RadLex taxonomy

RadLex (RSNA, 2012) is a taxonomy published by the Radiological Society of North America (RNSA) in order to deliver an uniform controlled vocabulary for indexing and retrieval of radiology information sources. The current English version 3.8 (n=39,542) contains terms organized in 13 main categories: anatomical entity as one among others such as treatment, image observation and imaging observation characteristics. A German version (Marwede et. al., 2009) has been worked-out in 2007. However, as the maintenance of this language version has been stopped, the latest version 2.0 contains only a subset of terms (n=10,003). Our approach covers this lack in terminology and extends the resource.

For a structured analysis of the controlled vocabulary, we filtered an initial lexicon containing 9,479 entries.

4.2 Vocabulary coverage

The 9,479 entries in the linguistic resource contain 23,588 tokens of which 6,326 are distinct. Comparing this number with the word types used in the development set (n=3,172), the first assumption is that the lexicon covers the vocabulary used in the reports without problems. However, we discovered that this is not the case. We identified the three major problems:

1. The lexicon contains quite rare terminology which is not used in the development set, e.g., *absorbierbarer Gelatineschwamm* (RID11213) [absorbable gelatin sponge].

2. Additionally, important terms that have both a high occurrence in the development set and relevance for the pathology classification are either not included in the lexicon (e.g. *Läsion* [lesion]) or are included but are not classified (e.g. *vergrößert* | RID 5791 [enlarged]).

3. As learned from the corpus analysis, the description of pathological findings requires a rich vocabulary. However, the lexicon entries classified initially as 'pathological' represent only 18.1% of the whole resource (Table 1; We deduce this number from an initial analysis and pathology classification of the topmost hypernyms and its substructures.). Our initial lexicon is obviously lacking a high amount of vocabulary to describe those pathologies.

Classification	#	
non-pathological	6,001	63.3%
pathological	1,714	18.1%
not to be determined	1,764	18.6%
	9,479	100%

Table 2: Pathology classification of RadLex entries

The analysis reveals that the initial lexicon does not fully cover the whole range of vocabulary used in the reports. Furthermore, not all words in the initial lexicon can be classified just by using the structural information of the taxonomy. That is why we introduce the following corpus-based learning approach to enhance the lexicon to enable a correct sentence classification and alignment.

5 Methods

5.1 Conclusions from the corpus and initial lexicon analysis

When comparing German and English reports, one can observe two characteristics in both languages: syntactic shortness and reduced semantic complexity. Based on this observation, (Friedman et. al., 1994; Friedman et. al., 2002; Sager et. al., 1994) successfully created semantic grammars for medical text parsing. We conclude, that this is also possible for German reports.

We use a semantic grammar for sentence classification, thus, we conduct that the learning of classified vocabulary from pre-annotated sentences is possible. The insights gained from the *statistical analysis* simplify the grammar creation: For deriving additional vocabulary from the reports, the short length of the sentences is of advantage. The short structure allows for derivation of knowledge with high certainty. Even if only little amount of seed vocabulary is available, the unknown vocabulary can be classified easily and with high reliability.

5.2 Derive grammar

The grammar rules are derived from the sentences in the development set. First, the semantic classes are defined and finally they are combined into valid grammar rules. The semantic classes are initially adapted from (Friedman et. al., 2002), but then reduced to 32 classes which either

1. are necessary for classification (distinguish between words containing pathological or non-pathological semantics),

2. carry special semantic properties (e.g. anatomical entities),

3. or carry linguistic features (negations, prepositions, enumerations, etc.).

The classes are combined into 238 grammar rules. The grammar follows the same intention as the grammars developed by (Friedman et. al., 1994; Sager et. al., 1994): to model the structure of the reports' sentences. But it pursues a different goal: The grammar is used to classify the sentences as either 'pathological' or 'non-pathological'.

The top-most non-terminals designate the classification: A sentence can be reduced to a PATH or NOPATH non-terminal. All subsequent grammar rules are hierarchially embedded into these non-terminals and form the semantic structure of sentences. Sample rules and sentences are listed below:

- PATH → DISEASE
  Tracheostoma$_{[DISEASE]}$.
- PATH → DISEASE MOD_PATH
  Nierenzyste$_{[DISEASE]}$ *rechts*$_{[MOD\_PATH]}$.
  [Kidney cyst right.]
- NOPATH → NEGATION DISEASE
  Kein$_{[NEGATION]}$ *Ödem*$_{[DISEASE]}$. [No edema.]

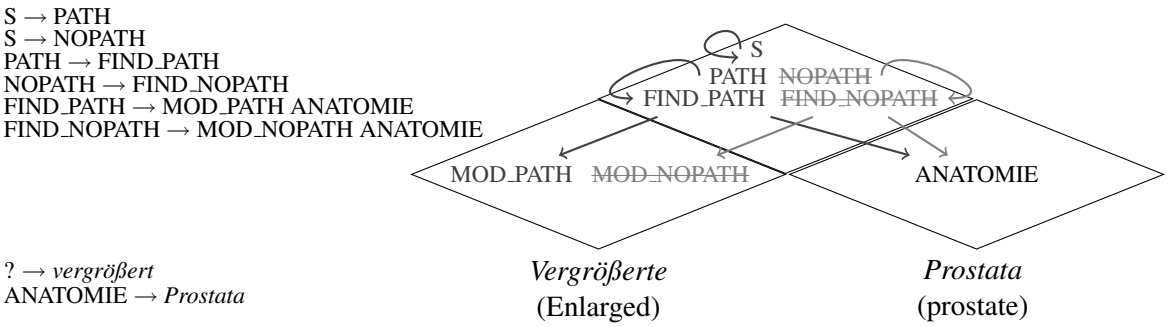

Figure 2: Learning lexical knowledge from sentence *Vergrößerte Prostata* (Enlarged prostate)

- NOPATH → ANATOMY MOD_NOPATH KOMMA NEGATION MOD_PATH
  Milz$_{[ANATOMY]}$ *homogen*$_{[MOD\_NOPATH]}$,$_{[KOMMA]}$ *nicht*$_{[NEGATION]}$ *vergrößert*$_{[MOD\_PATH]}$. [Spleen homogeneous, nor enlarged.]

As observed in the corpus analysis, the sentences describing pathological findings are longer, and thus, more complex in syntax compared to sentences describing non-pathological findings. This requires a higher amount of grammar rules for the description of the structure of pathological sentences. We manage this requirement by defining a set of rules of which the majority of 52% define the structures of sentences to be classified as pathological.

5.3 Learn from the development set

Rationale for learning method Our learning algorithm models the process medical students undergo when learning medical terms directly from texts. To align this model with our approach, we assume that the students know whether a sentence describes pathological or non-pathological findings. In addition, they have (basic) medical knowledge, which they can apply, e.g. about anatomical entities. When learning new vocabulary and its correlating pathology classification, they use this as seed knowledge. To validate their knowledge and derive new words with high certainty, they start with the shortest sentences. Proceeding with the sentences length-wise, they re-validate their knowledge and continue learning. The reliability of newly learned knowledge and classification decreases with the sentence length.

Learning process Our approach follows the same steps:

- We apply initial medical knowledge (in the form of pathology classification) from the lexicon.
- Knowledge about possible syntactic constructs is given with the grammar rules.
- Each sentence to learn from has information annotated about the correct pathology classification.
- We start with the shortest sentences to derive new vocabulary and pathology classification from. This is done, because learned knowledge from shorter sentences (with limited syntactic diversity) is correct with higher certainty.
- We apply the existing and learned knowledge in the following iterations to derive additional vocabulary and pathology classification.

Learning method We apply a semi-supervised learning algorithm: Each of the sentences to learn from is annotated with the target classification. But actually, we learn on the word level, where no annotations are available. Applying the rules of the semantic grammar, we derive the word-level semantic classification (which includes both the non-terminal assignment and the pathology classification) from the overall sentence classification.

Input for each parsing iteration is the sentence as an ordered list of words and the attached pathology classification. Starting with the shortest sentences, we can derive new vocabulary with high reliability, as those sentences are low in syntactic diversity. Additionally, the information about the target pathology classification reduces the rules that can be applied during the parsing process.

For learning, we adapt the standard probabilistic CKY parsing algorithm. How the algorithm operates in detail is illustrated in Figure 2. The goal

is to learn the pathology classification of the word *vergößert* [enlarged], which is currently not available.

The initial step of non-terminal assignment is mainly based on the lexical resource. If terms are contained in the lexicon, their non-terminal assignment can be derived from the semantic classification. (As the non-terminal for *Prostata* [prostate] is ANATOMIE.) If a term is not contained in the lexicon, we assign a number of possible non-terminals. Those non-terminals include one symbol that presumes that the terms describe a pathological state and one that presumes the opposite. (I.e., vergrößert is initially assigned the non-terminals MOD_PATH and MOD_NOPATH)

The disambiguation of the non-terminal assignment is resolved during the parsing process: On the one hand, the probabilistic nature of the grammar rules enable a disambiguation of the most probable constituent structures. On the other hand, the target pathology classification excludes invalid rules. (Which is in case of the example, the sentence is annotated as PATH, so any subsequent rule for this non-terminal is not considered; struck-through in the figure.) In the end, the non-terminals assigned to existing or unknown vocabulary is used to enhance the lexicon. (Finally, we can derive that *vergößert*, assigned MOD_PATH, describes a pathology.)

Results of the learning process After the learning step, the lexicon is extended to 10344 vocabulary terms (before 9479). But even more important, the overall amount of lexicon entries classified as 'pathological' increased by 18.8 % to now 2036 entries (before 1714). We consider this a key success of the learning, as our classification depends on this encoded knowledge.

6 Evaluation of the classification results

We evaluate the system using 40 randomly-chosen radiology reports containing 1294 sentences. We compare results of the sentence classification using the initial linguistic resource and the extended one. Table 3 shows the classification results for the two evaluated cases.

The learning resulted in an increase of vocabulary by 9.1%. At the same time, the pathology classification could be increased overproportionally by 18.8%. While the learning increased recall (from 45.4% with initial lexicon to 74.3% with additional, learned vocabulary), precision decreased.

Higher recall importance Before discussion these numbers, the higher importance of the recall value for our use case of aligning radiology text and images has to be underlined:

Only for sentences correctly classified as 'pathological', the contained anatomical entities are extracted and anatomical annotations are created. If sentences are misclassified as 'pathological' (although they describe non-pathological findings), this is a minor issue. As a result of this misclassification, anatomical entities in the sentence are extracted and links to the image annotations are created, although the images do not show any pathologies. We accept those additional, but not intended links.

In the workflow, links from textual findings to image positions for non-pathological findings are no problem compared to non-existing ones for pathological findings. In case links from text to images cannot be created because a sentence was misclassified as non-pathological, the radiologist still has to link the textual findings to the correlating image position manually. This should be avoided.

We conclude, that the true classification of pathological sentence is more important for the alignment, hence, the recall value indicating this case has higher weight for us.

Discussion But still the quality of the learning step can be improved: While the sentences correctly classified as 'pathological' increase using the learned vocabulary, the sentences correctly classified as 'non-pathological' decrease at the same time. The latter is indicated by the increasing 'false positive' (FP) value. This is the main reason for decreasing precision.

We see that the learned vocabulary contains several entries misclassified as 'pathological' **(Error type 1).** The consequence of this misclassification are more sentences classified as 'pathological' although they describe non-pathological findings.

Examples can be identified both from FP and FN cases in the test set: Terms that do not describe pathological properties such as *Voraufnahme* [previous examination] or *Lymphknoten* [lymph node] were classified as pathological. Even very obvious pathological findings such as *Läsion* [lesion] or *Infiltrat* [infiltrate] are not classified correctly. Because of their high usage frequency, these four terms are accountable for 169 of the misclassified sentences in the test set.

	vocabulary	PATH class	FP	TN	P	R
baseline	9,479	1,714	149	682	0.585	0.455
extended lexicon	10,344	2,036	288	543	0.544	0.743

Table 3: Classification results with initial lexicon

The application of a semi-supervised learning approach with sentence-level annotations for word-level vocabulary acquisition is obviously point for improvement. We will include a probabilistic feature in the learning process that takes into account all occurrences of a vocabulary term to be learn in order to increase the leaning certainty.

The second major issue for correct pathological classification is the lack of grammar rules for long sentence structures. Since those sentences are more likely describing pathological findings and they cannot be considered in the learning process, the contained pathology descriptions are missing in the lexicon **(Error type 2)**. A more sophisticated grammar engineering can help to bridge this gap.

Two further, but minor error types remain. **Error type 3** describe incorrectly resolved non-terminal matches because of not considered linguistic details:

- **Failed subtoken matching in composita** E.g. the term *Nasennebenhöhle* does not match the subtoken *Nase* as expected because the token itself was learnt before as new, non-anatomical lexicon entry.
- **Naming mismatch between lexicon and text** E.g. *Lebersegment II nach Couinaud* (RID62) is expected to match, but in the text it is only refered to as *Segment 2*. This can be resolved detecting synonyms.
- **Mismatch of (distant) multi-token matches** This is of special importance as 72 % of the lexicon entries are multi-token entries. Their individual components can be distributed within a sentence. E.g. The multi-token text *Lymphknoten im oberen Mediastinum* does not match the lexicon entry *Oberer mediastinaler Lymphknoten* (RID7739).

The failure of the type 3 errors can be solved by introducing more elaborated linguistic techniques.

And finally, **Error type 4** indicates the still missing amount of vocabulary not available for classification. Even though, we tried to extend the development corpus to a maximum, it is not possible to cover all possible description applied in radiology. For a higher learning rate, the development corpus has to be extended significantly.

The extension of the lexicon has a significant impact on the classification results. Comparing the results of the classification using the initial lexicon and using an extended lexicon, the impact of a complete controlled vocabulary becomes apparent. In particular, the completeness of the lexicon contributes to the correct classification of sentences describing pathological findings.

7 Conclusion

For implemented a system that aligns findings in radiology reports with findings in images based on semantic annotations, the incomplete linguistic resource has to be extended with vocabulary. We overcome this issue by introducing a semi-supervised learning approach that adapts the existing grammar rules to learn new and classified vocabulary. Incorporating this learned vocabulary, the grammar-based classification delivers a recall value of 74.3%.

The issue we are dealing with is relevant for further work on German clinical texts: Still, the coverage of controlled vocabularies and ontologies for medical texts written in languages other than English include a large gap. We believe that lexicons are the most crucial resources for language processing in the medical domain. That is why we will focus our future work on extending and enriching existing lexicons and establishing new resources for linguistic analysis.

Acknowledgments

This work was partially supported by EIT ICT Labs (in the Medical CPS activity).

This research has been supported in part by the THESEUS Program in the MEDICO Project, which is funded by the German Federal Ministry of Economics and Technology under grant number 01MQ07016. The responsibility for this publication lies with the authors.

References

D. A. Campbell, S. B. Johnson. 1999. A technique for semantic classification of unknown words using UMLS resources.. *Proc AMIA Symp*, 716–20.

J. W. Fan and C. Friedman. 2011. Deriving a probabilistic syntacto-semantic grammar for biomedicine based on domain-specific terminologies.. *J Biomed Inform.*, 44(5):805-14.

Carol Friedman, Philip O. Alderson, John H. M. Austin, James J. Cimino, and Stephen B. Johnson. 1994. A General Natural-Language Text Processor for Clinical Radiology. *Journal of the American Medical Informatics Association*, 1:161–174.

Carol Friedman, Pauline Kra, and Andrey Rzhetksy. 2002. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35:222–235.

S. B. Johnson. 1999. A semantic lexicon for medical language processing.. *J Am Med Inform Assoc*, 6(3):205-18.

T. Kasami. 1965. An efficient recognition and syntax-analysis algorithm for context-free languages. *Scientific Report AFCRL-65-758*, Air Force Cambridge Research Lab.

H. Liu, S. T. Wu, D. Li, S. Jonnalagadda, S. Sohn, K. Wagholikar, P. J. Haug, S. M. Huff, and C. G. Chute. 2012. Towards a semantic lexicon for clinical natural language processing.. *AMIA Annu Symp Proc*, 568-576.

D. Marwede, P. Daumke, K. Marko, D. Lobsien, S. Schulz, and T. Kahn. 2009. RadLex - German version: a radiological lexicon for indexing image and report information. *Fortschr Röntgenstr*, 181(1): 38–44.

Stephane M. Meystre, Guergana K. Savova, Karin C. Kipper-Schuler, and John F. Hurdle. 2008. Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research. *Yearbook of Medical Informatics*, 24(11):128–144.

Radiological Society of North America. 2012. RadLex. `http://rsna.org/RadLex.aspx`.

Naomi Sager, Margaret Lyman, Christine Bucknall, Ngo Nhan, and Leo J. Tick. 1994. Natural Language Processing and the Representation of Clinical Data. *Journal of the American Medical Informatics Association*, 1:142–160.

G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications.. *J Am Med Inform Assoc.*, 17(5):507-13.

Sascha Seifert. 2010. THESEUS-Anwendungsszenario MEDICO. `http://www.joint-research.org/das-theseus-forschungsprogramm/medico/`.

S. Seifert, A. Barbu, K. Zhou, D. Liu, J. Feulner, M. Huber, M. Suehling, A. Cavallaro, and D. Comaniciu. 2009. Hierarchical Parsing and Semantic Navigation of Full Body CT Data. *SPIE Medical Imaging*.

S. T. Wu, H. Liu, D. Li, C. Tao, M. A. Musen, C. G. Chute, N. H. Shah. 2012. Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis.. *J Am Med Inform Assoc*, 19(1):149–56.

P. Zweigenbaum, R. Baud, A. Burgun, F. Namer, E. Jarrousse N. Grabar, P. Ruch, F. Le Duff, B. Thirion, and S. Darmoni. 2003. UMLF: a Unified Medical Lexicon for French. *AMIA Annu Symp Proc*, 1062.