Chapter 4
Characteristics of Patient Records and Clinical Corpora

This chapter will describe the characteristics of patient records compared to other text types including: A comparison of the characteristics of patient records written in different languages, the number of spelling errors compared to other types of text, syntactic differences, word choices, abbreviations, acronyms, compounds and compound construction, negation expression and also speculative cues and factuality expressions in clinical text. Patient record text is different from standard text such as news text or novel text regarding style and grammatical correctness. Patient record text is also written by many different professionals with different writing skills. The varies writing style depending on whether it is a physician's note or a discharge letter, or nursing narratives. Also the style varies between different clinical units and specialities. Pathology reports in turn contain a very special type of writing, where the physician describes a sample taken from the patient without ever meeting the patient.

4.1 Patient Records

Patient records are primarily written for hospital internal use and for mnemonic reasons. Daily running notes might contain more spelling errors or noisiness than discharge letters that are read by a larger audience (Ehrentraut et al. 2012).

The linguistic term *corpus* meaning a document collection, the plural *corpora* means several document collections.

In a study comparing Finnish and Swedish nursing narratives the authors observed a large number of linguistic and grammatical errors in both languages. Complete sentences are rare; auxiliary verbs such as *be, is* and *are* are rarely used. *Patient* is not mentioned unless in abbreviated form, or sometimes the patient as a subject is not mentioned at all, the subject of the sentence is also missing, a sentence may contain only a number of adverbs such as *fever, sweating, trouble breathing* (Allvin et al. 2011).

H. Dalianis, *Clinical Text Mining*, https://doi.org/10.1007/978-3-319-78503-5_4

Septisk pat, oklart fokus, rundodlas före Zinacef.
(in Eng. Septic pat, unclear origin, roundcultured before Zinacef)

which means:

Patienten har sepsis med oklart ursprung, bakterieodling tas från samtliga möjliga infektionsfokus, inklusive blododling, innan behandling med Zinacef inleds.
(in Eng. The patient has sepsis of unclear origin, bacterial culture samples taken from all possible foci for infection, including blood culture samples, before commencing treatment with Zinacef.)

Fig. 4.1 An example of aggregated clinical text. It has been rephrased so as not to contain any redundant information, and it presumes background knowledge of the reader (© 2014 Springer International Publishing Switzerland—reprinted with permission. Published in Dalianis (2014))

According to Pakhomov et al. (2005) there are 30% non-word tokens, abbreviations, acronyms, misspellings, wrongly used grammar etc. in clinical text, which is a good indication of the noisiness of clinical text.

Generally patient records are written by highly skilled physicians and nurses using domain specific terms. For example, patient record text is very domain specific depending on which medical discipline it is written in. Each discipline or domain within medicine uses its own set of terms that can be incomprehensible by other disciplines.

Patient records are, as mentioned in Sect. 2.4, highly structured with headings such as *Subjective, Objective, Assessment* and *Plan*, but this is not always followed by individual physicians, or between different professions. The writing under the correct heading and also the names of the headings may differ in different clinical units or hospitals. The same for different electronic patient record systems.

The patient records are written under time pressure; the patient record systems do not contain any spelling correction (or grammar checking) system due to the difficulties of building such a function because of the complicated non-standard vocabulary used within healthcare.

Therefore, in clinical text non-standard abbreviations, domain specific acronyms, and incomplete sentences without a subject can be observed, meaning the patient is not mentioned, only his or her status. The text is short and efficient, and written in telegraphic style, see Fig. 4.1 on aggregated clinical text. Moreover, the text can be full of jargon and misspellings. The physicians reason and argue to find the diagnosis by excluding symptoms and mention them in negated form (Groopman 2007).

4.2 Pathology Reports

Pathology reports are written by pathologists, highly skilled physicians. Pathologists are experts in interpreting laboratory tests. They study samples and tissues from the human body, and describe the samples in free text in pathology reports. The structure and content of a pathology report is as follows: the name of the patient, how each

tissue sample was obtained, how it looks compared to normal tissue and normal cells, the diagnosis and a description of the diagnostic tests such as (possible) tumor such as size (typically in mm or cm), type and grade in the *TNM scale* (growing and spreading risk of the tumor), lymph node status, number of lymph nodes containing tumor metastasis, the position in body (Asamura et al. 2014). Whether the tumor is non-invasive or invasive and contains various hormone receptors (for example Her-2), proliferation rate (Ki-67, MIB1) or Gleason grade (also called Gleason score) depending on cancer type and finally the tumor margins (when performing a biopsy the tumor size is not known).[1,2]

Generally, pathology reports are more carefully written than patient records, often with correct spelling. The pathology reports have more semi-structure than the patient records. The pathologist is using the writing and the text as a tool for their profession, see Fig. 4.2 for an example of a pathology report for breast cancer.

The pathology report is sent to the physician who is treating the patient, so the physician can decide how to proceed with the treatment.

Samples are taken from the body at regular intervals to follow the progress, or hopefully the regression, of the cancer. The pathology reports are very often registered in regional or national cancer registries for statistics on cancer treatment and outcomes.

At the cancer registry well-trained coders read and interpret the content of the pathology report and then manually enter the information to the cancer registry. This work is time consuming and tiresome. The agreement between pathology report coder is not known. One hypothesis is that agreement between them may be around 0.8 in F-score as for other annotation tasks.

4.3 Spelling Errors in Clinical Text

The number of spelling errors in clinical text has been calculated in a few publications. Ruch et al. (2003) found around 10% spelling errors in French clinical text, while Patrick and Nguyen (2011) only found 2.3% spelling errors in Australian English clinical text, and finally Nizamuddin and Dalianis (2014) found 7.6% spelling errors in the Stockholm EPR PHI Corpus containing in total 174,000 tokens. In another smaller subset of the same Swedish clinical corpora 1.1% of the words were found to be misspelled (Grigonyte et al. 2014).

[1]Contents of a pathology report, http://ww5.komen.org/BreastCancer/ContentsofaPathologyReport.html. Accessed 2018-01-11.

[2]National Cancer Institute, Pathology Reports, http://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis/pathology-reports-fact-sheet#. Accessed 2018-01-11.

```
Mammaresektat (ve. side) med infiltrerende duktalt
karsinom, histologisk grad 3
Tumordiameter 15 mm
Lavgradig DCIS med utstrekning 4 mm i kranial
retning fra tumor
Frie reseksjonsrender for infiltrerende tumor (3 mm
kranialt)
Lavgradig DCIS under 2 mm fra kraniale
reseksjonsrand

ER: ca 65 % av cellene positive
PGR: negativ
Ki-67: Hot-spot 23% positive celler. Cold spot 8%.
Gjennomsnitt 15%
HER-2: negativ
Tidl. BU 13:

3 sentinelle lymfeknuter uten pÃěviste patologiske
forandringer
```

Translated to English:

```
Mamma specimen (le. side) with infiltrating ductal
carcinoma, histological grade 3
Tumor diameter 15 mm
Low-grade DCIS extending 4 mm in cranial direction
from the tumor
Free resection margins for infiltrating tumor (3 mm
cranially)
Low-grade DCIS less than 2 mm from the cranial
resection margin

ER: ca 65 % of the cells are positive
PGR: negative
Ki-67: Hot-spot 23% positive cells. Cold spot 8%.
Average 15%
HER-2: negative
Prev. BU 13:

3 sentinel lymph nodes without proven pathological
changes
```

Fig. 4.2 Extract from the free text part of an anonymised breast cancer pathology report in Norwegian (and its translation to English). The data in the figure is made up and can not be linked to any individual (© 2015, Association for Computational Linguistics (ACL). All rights reserved. Reprinted with the permission of ACL and the authors. Published in Weegar and Dalianis (2015))

In Fig. 4.3, shows a clinical text with a number of misspellings.

In Table 4.1 there are some examples of misspellings in Swedish patient record text and their correct spelling, together with the corresponding misspelled version in English and the correctly spelled English word.

Beklagar nissförstånd rek ayt provar mindre smaker som innehåller mindre Kolhydrater 8vilket pat benämner som smaken sött som diasip, komplett näring naturell samt provide x-tra tomat. Ut tar upp dessa till avd för utprovning . Vi ska se vad vi kna göra med de näringsdrycker som finns i hemmet då pat är åter hemma...
(in Eng; Sorry for the nisunderstanding rec tto try less flavours that contain less Carbohydrates 8which pat name as taste sweet like diasip, complete nutrition natural as well as provide x-tra tomato. Ut takes these to clin for try out . Let's see what we cna do with the nutrition drinks in the house when pat is back home. . .).

Fig. 4.3 Example of clinical text with spelling errors (© 2009 The authors—reprinted with permission from the authors. Published in Dalianis et al. (2009))

Table 4.1 Some examples on misspelled Swedish words in a patient record and their equivalents in English

Misspelled word Swedish	Correct word	Misspelled word Eng	Correct word Eng
Karnvatten	Kranvatten	Tpa water	Tap water
Deligation	Delegering	Deligation	Delegation
Rekomendation	Rekommendation	Recomendation	Recommendation
Inger förtroende	Inget förtroende	Gives confidence	No confidence

Misspelled parts are underlined. The last *confidence*-example is a real example where the misspelling *inger* (gives) is a real word while the author wanted to write *ingen* (no), which is not obvious since it created a real word and changed the whole meaning of the expression

Table 4.2 Spelling errors in various types of text

Type of texts	Misspellings
Text written in e.g. Word	0.2
Newspaper text	0.05–0.44
Web text	0.8
Hand written text	1.5–2.5
Typed textual conversations	5.0–6.0
Patient record text	10.0

© 2012 The authors—reprinted with permission from the authors. Published in Ehrentraut et al. (2012)

In Table 4.2, taken from Ehrentraut et al. (2012), the percentage of spelling errors in patient record text compared with other type of texts can be observed. One finding by Ehrentraut et al. (2012) was that patient records contained twice as many abbreviations as text messages, with 10.6% and 5.0% respectively.

4.4 Abbreviations

An abbreviation is a way to write a word not using the complete spelling, for example the word *patient* written as *pat*. Abbreviations are an efficient way to write text, but it makes the reading slower since the reader has to interpret the

abbreviations. In some cases the abbreviation can be ambiguous, for example, *pat* can be ambiguous since it also can mean *pathological*.

Here follow some studies on abbreviations in Swedish clinical text: in a study by Allvin et al. (2011), the authors found that 4.7% of the words in Swedish nursing narratives, a subset of the Stockholm EPR Corpus (Dalianis et al. 2009), were abbreviations. In a study by Isenius (2012), Isenius et al. (2012), 19,408 tokens from another subset of the Stockholm EPR Corpus were extracted and then manually annotated by a senior physician with previous experience in annotating clinical texts. In total 2050 tokens were identified as abbreviated tokens, of these 335 were unique. In total in this small subset 1% abbreviations were found. Skeppstedt et al. (2012) constructed a rule-based system to detect findings and disorders in a subset of the Stockholm EPR Corpus, 14% of the manually annotated disorders in the gold standard were written in an abbreviated format. Nizamuddin and Dalianis (2014) studied the Stockholm EPR PHI Corpus, which is a subset of the Stockholm EPR Corpus, and found 2.7% abbreviations. Finally, Olsson (2011) analysed patient records from a Swedish surgery department, and the author found 2.4% abbreviations.

Patrick and Nguyen (2011) found 0.7% abbreviations in Australian English clinical text. Wu et al. (2011) found in a small English clinical corpora containing 18,225 tokens a total of 1386 abbreviations, which is around 0.8% abbreviations.

Siklósi et al. (2014) studied a Hungarian clinical sub-corpus within ophthalmology consisting of 552,594 tokens and found 113,091 abbreviated tokens, which correspond to 20% of the total corpus.

Regarding ambiguity of the clinical abbreviations there are two studies: Liu et al. (2001) found that 33% of the abbreviations in English clinical text were highly ambiguous and Lövestam et al. (2014) analysed 40 different abbreviations in Swedish dietetics notes from the subset of the Stockholm EPR Corpus, written by three professions: dieticians, nurses and physicians. A contextual analysis showed that 33% of the abbreviations were ambiguous.

Many of the abbreviations, around 12%, in clinical text are compounds where one part is composed of an abbreviation and the other part of a complete word, see Table 4.3 for examples (Kvist and Velupillai 2014). For a dictionary of Swedish medical abbreviations see Cederblom (2005).

Table 4.3 Examples of Swedish clinical abbreviations from a Stockholm EPR Corpus, some of the them are ambiguous

Abbreviated word in Swedish	Resolved word in Swedish	Resolved word in English
ul	Ultraljud/underläkare	Ultrasound/assistant physician
rtg	Röntgen	X-ray
p5	Petrokantär femurfraktur	Hip fracture
Lungrtg	Lungröntgen	Lung X-ray

One abbreviation *lungrtg* (lung X-ray) is composed of a full word form *lung* and an abbreviation *rtg* meaning *röntgen* (X-ray)

Both the *Unified Medical Language Systems (UMLS)* and *Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)* contain abbreviations, and can be used for creating abbreviations lists. UMLS is only available for English, while SNOMED CT is available in several languages, but not all language versions contain abbreviations.

Many clinical words in Swedish are compounds, either full word form or combinations of full word forms and abbreviations as in the example *lungrtg* (lung X-ray).

4.5 Acronyms

Acronyms are a specialised form of abbreviations, usually using the first letters of each word in a phrase, or some combination of letters from words in a phrase forming an acronym with capital letters that also are easy to pronounce, while abbreviations consists of letters from one or more specific words, which are also easy pronounce.

An example of how an acronym has many different meanings is shown in Pakhomov et al. (2005). The UMLS concept *RA* has over 20 meanings, all of them different from each other, for example: *rheumatoid arthritis, renal artery, right atrium, right atrial, refractory anemia, radioactive, right aram, rheumatic arthriti, ragweed antigen, refractory ascites, renin activity, rheumatoid arthritis, renal artery, right atrium, right atrial, refractory anemia, radioactive, right aram* and *rheumatic arthritis*.

Patrick and Nguyen (2011) found 1.5% acronyms in Australian English clinical text. Kvist and Velupillai (2014) analysed both emergency unit records and radiology reports written in Swedish and found that 11% and 7.1% respectively contained abbreviations, and 33% and 55% respectively of the abbreviations where acronyms.

4.6 Assertions

Assertions are prepositions that have some sort of positive or negative polarity. They can range from completely negated to a speculative form through to a complete affirmed preposition.

4.6.1 Negations

Negations are very common in clinical text, since physicians use negations to exclude symptoms while reasoning about the cause of a patient's disease. The physician will write down the reasoning chain leading to the concluded disease; therefore, a lot of negated symptoms will be found in a patient record.

In a study by Chapman et al. (2001) more than half of the expressions in American radiology reports were found to contain negations. One explanation for this high amount is that these reports are mainly physicians' notes containing the physicians' reasoning (Groopman 2007). Physicians' notes contain more negations than nursing narratives that are about the daily healthcare of the patient. Another example is the English BioScope clinical corpus containing 13.6% negations (Vincze et al. 2008).

In Swedish, see Sect. 4.7.2 the texts from various clinical units under the heading *assessment* were studied and it was found that negated sentences or expressions encompassed 13.5% of the texts (910 negated sentences of a total of 6640 sentences) (Dalianis and Skeppstedt 2010).

4.6.2 Speculation and Factuality

Many of the expressions in clinical text are either negated or *uncertain* (sometimes also called *speculative*), or just asserted or assertions. Uncertain or speculative expressions in clinical text indicate that a statement is not affirmed or factual, and the uncertainty or level of speculation may range from slightly uncertain to strongly uncertain, see Fig. 4.4 for an example.

The English BioScope clinical corpus containing 6383 sentences was manually annotated for negation, speculation and scope. 13.4% of the sentences contained speculative keywords, so-called hedge sentences, and 13.6% contained negations, some of them overlapping (sentences containing both speculations and negations) (Vincze et al. 2008).

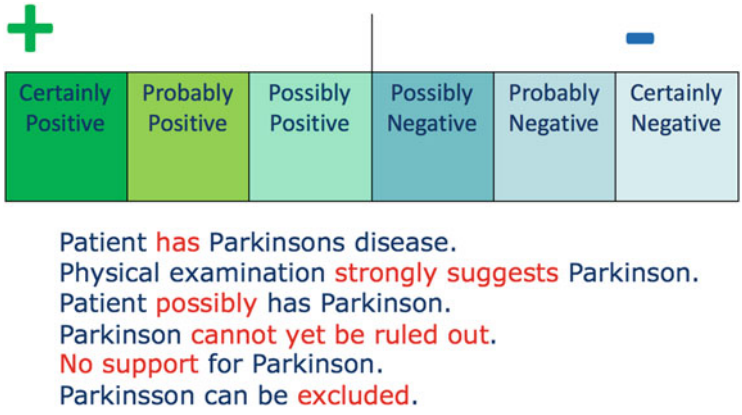

Fig. 4.4 Examples of different levels of certainty, ranging from completely affirmed to negated (© 2012 The author—reprinted with permission from the Author. Published in Velupillai (2012) p. 34)

A Swedish clinical corpus called the Stockholm EPR Diagnosis Factuality Corpus was developed to study uncertainty levels related to diagnostic expressions (Velupillai et al. 2011; Velupillai 2011). The corpus originates from a medical emergency ward, and specifically text under the heading *Assessment* or *Bedömning* was analysed. In total 3846 documents containing 26,232 sentences were manually annotated and 6483 diagnoses were found. The annotation was carried out by two senior physicians. These diagnoses were annotated for **certainty (positive or negative)**, and **uncertainty (possibly or probably)**, resulting in **six possible certainty levels**: *Certainly Positive, Probably Positive, Possibly Positive, Possibly Negative, Probably Negative* and *Certainly Negative*, see Fig. 4.4.

Levels of Certainty

According to Velupillai (2011) there were, on the negative polarity level, 11% *Certainly Negative* expressions, and 12.2% were in the middle of the scale (*Possibly Positive* and *Possibly Negative*) whilst 47.6% of the expressions were affirmed *Certainly Positive* in the final version of the corpus. The inter- and intra-annotator agreement on the subset that was annotated by both annotators is presented in a confusion matrix in Table 4.4. Velupillai was inspired to use the six different levels

Table 4.4 Confusion matrix, intra- and inter-annotator agreement

	CP	PrP	PoP	PoN	PrN	CN	ND	O	Σ
CP Intra	**990**	78	4	0	3	4	2	19	1100
Inter	***834***	*59*	*7*	*0*	*4*	*5*	*1*	*20*	*930*
PrP Intra	20	**236**	55	1	1	0	1	0	314
Inter	*66*	***134***	*10*	*1*	*0*	*0*	*2*	*1*	*214*
PoP Intra	4	38	**127**	25	9	0	0	2	205
Inter	*11*	*149*	***180***	*41*	*45*	*1*	*1*	*10*	*438*
PoN Intra	0	0	6	**14**	7	1	0	1	29
Inter	*0*	*0*	*0*	***1***	*5*	*1*	*0*	*0*	*7*
PrN Intra	1	1	1	10	**118**	25	0	5	161
Inter	*0*	*0*	*0*	*2*	***35***	*18*	*0*	*1*	*56*
CN Intra	2	0	4	0	51	**195**	0	1	253
Inter	*2*	*0*	*0*	*4*	*99*	***193***	*1*	*3*	*302*
ND Intra	0	0	0	0	0	0	**26**	0	26
Inter	*13*	*5*	*3*	*2*	*1*	*3*	***30***	*4*	*61*
O Intra	8	1	4	1	7	0	8	**65**	94
Inter	*1*	*1*	*1*	*1*	*5*	*3*	*1*	***49***	*62*
Σ Intra	1025	354	201	51	196	225	37	93	**2182**
Inter	*927*	*348*	*201*	*52*	*194*	*223*	*36*	*88*	***2070***

Table taken from Table 1 in Velupillai et al. (2011) © 2012 with permission from IOS Press. Published in Velupillai et al. (2011)

Columns: Al, first annotation iteration. Rows: Intra: Al, second annotation iteration (same set randomized), Inter: A2. CP = Certainly Positive, PrP = Probably Positive, PoP = Possibly Positive, PoN = Possibly Negative, PrN = Probably Negative, CN = Certainly Negative, ND = Not Diagnosis, O = Other, Σ = Total

by Saurí and Pustejovsky (2009). Saurí and Pustejovsky used a six level scale as well as a label for *Unknown* or *Uncommitted* when they annotated English newswire and broadcast news reports in the FactBank corpus. Their goal was to annotate the degree of factuality of the events.

In Velupillai et al. (2011) the inter-annotator and intra-annotator agreements are presented for the factuality annotations carried out by the two senior physicians. The main finding is that the overall agreement was fairly high (0.7/0.58 F-score and 0.73/0.6 Cohen's κ for intra-/inter- annotator agreement respectively).

Cohen's κ, or Cohen's kappa measures the probability of obtaining high agreement. A kappa value of 0.6 to 0.73 means the task is fairly difficult, indicating that the annotators have moderate to high agreement. A kappa score of 1.0 indicates full agreement (and an easy annotation task). An alternative way of measuring agreement is the F-score, an average result of 0.6–0.7 can be considered quite high. For the whole PhD study about factuality levels in Swedish clinical text see Velupillai (2012).

Negation and Speculations in Other Languages, Such as Chinese

One interesting approach in speculation detection for Chinese clinical notes can be found in Zhang et al. (2016). The authors obtained 0.922 in F-score with a CRF-machine learning approach using 5103 gold-standard speculation annotations as data. One critical point was to obtain a high quality word segmentation to obtain high performance in speculation detection in Chinese, since Chinese does not use spaces as word delimiters.

4.7 Clinical Corpora Available

Generally, it is difficult to get access to clinical corpora for research. This is mainly due to the sensitive information they may contain regarding individuals. We will describe the process of obtaining clinical corpora for research, with ethical permission and de-identification. Most of the few corpora available for research are in English, but some corpora are available in other languages.

4.7.1 English Clinical Corpora Available

The two most well-known clinical corpora are the Informatics for Integrating Biology & the Bedside (i2b2)[3] clinical corpus consisting of approximately 1000

[3]i2b2, http://www.i2b2.org. Accessed 2018-01-11.

notes in English and the Computational Medicine Center (CMC)[4] corpus containing 2216 patient records in American English (Pestian et al. 2007). Another well-used corpora is the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC II)[5] it is also called the De-id corpus and consists of 1934 discharge summaries and 412,509 nursing notes written in American English (Saeed et al. 2011). The BioScope Corpus and the Thyme corpus are two other well-known clinical corpora written in English. The BioScope corpus contains almost 6383 annotated sentences from the clinical domain and is described in Sect. 4.6.2. The Thyme corpus contains 1254 de-identified notes which are annotated for temporal relations. The Thyme corpus is described in Sect. 7.5.5.

The clinical corpora in English are de-identified with respect to sensitive identifiers such as personal names, phone numbers etc. These are identified and removed and some are also pseudonymised, meaning sensitive identifiers are replaced with pseudonyms or surrogates to keep the text natural. This is carried out by using automatic methods combined with manual review of the corpora for residues of sensitive information.

4.7.2 Swedish Clinical Corpora

The Swedish clinical corpora called the Stockholm EPR Corpus contained in the HEALTH BANK—Swedish Health Record Research Bank,[6] from Karolinska University Hospital in Stockholm. It contains over two million patients from over 500 clinical units and encompasses the years 2007–2014. The corpus is de-identified with regard to patient's personal names. Personal identity numbers are replaced with a serial number that makes it possible to follow the patient through the healthcare process, from admission to discharge from the clinical unit (Dalianis et al. 2015).

The Stockholm EPR Corpus contains both structured and unstructured information. The structured data includes gender and age of the patient, admission and discharge date and time, ICD-10 diagnosis codes, drugs both the name and the ATC-codes, blood values, laboratory values etc. The unstructured text consists of physicians' notes and nurses' narratives as well as other notes about the patient from other professionals in the healthcare process.

The Stockholm EPR Corpus is part of the HEALTH BANK—Swedish Health Research Bank. Many smaller subparts of it have been annotated for:

- De-identification, for privacy (Stockholm EPR PHI Corpus).
- and its corresponding Stockholm EPR PHI Pseudo Corpus containing pseudonymised PHI.
- Sentence uncertainty including negations.

[4]CMC, https://ncc.cchmc.org/prod/pestianlabdata/request.do. Accessed 2018-01-11.

[5]MIMIC II, http://www.physionet.org/physiotools/deid. Accessed 2018-01-11.

[6]Swedish Health Record Research Bank, http://dsv.su.se/healthbank/. Accessed 2018-01-11.

- Diagnosis factuality, six different levels ranging from affirmative expressions to negations, see Fig. 4.4 for details.
- Clinical (named) entities such as disorders, findings, diagnoses, and body structures and drugs.
- Abbreviations.
- Document classification in two classes, documents containing information on healthcare associated infection or not.
- Adverse drug events (ADE).
  - Attributes, such as negations, speculations, past and future.
  - Relations such as indication, adverse drug event, ADE outcome and ADE cause (these relations will be explained in Sect. 10.2.

All the manual annotations on clinical texts are stored in the HEALTH BANK and are described online.[7] The annotated data is clinical text written in Swedish. The descriptions are in Swedish, but can be understood since the annotation classes are in English and there are numerical values for the number of classes .

4.7.3 Clinical Corpora in Other Languages than Swedish

Here follows an enumeration of various clinical corpora in the following languages: Bulgarian, Danish, Dutch, English, Finnish, French, German, Italian, Norwegian, Polish, Spanish, Swedish and Japanese. Many of these corpora are closely connected to individual research groups, and special permission and good contacts are required for access.

- A British English clinical corpus[8] known as the general practice research database used for recognising symptoms automatically. The researchers annotated 6141 records in General Practice/ Primary care (Koeling et al. 2011).
- An Australian English clinical corpus from the Concord hospital's clinical progress summary containing 43,712 anonymised patient records from 2003 to 2008 (Patrick and Nguyen 2011).
- An Australian English pathology corpus from the state of Queensland in Australia containing 45.3 million pathology HL7 messages, including 119,581 histology and cytology reports (Nguyen et al. 2016).
- Another British English clinical corpus containing 20,000 cancer patient records used for semantic annotation and described in Roberts et al. (2009).
- Yet another British English corpus taken from the SLaM BRC Case Register (London area) on 31 December 2014, containing over 250,000 patient records within psychiatry (Perera et al. 2016).

[7] Annotated data in HEALTH BANK, http://dsv.su.se/healthbank/annotated-data/. Accessed 2018-01-11.

[8] British English clinical corpus, http://www.gprd.com. Accessed 2018-01-11.

- Another British English corpus is the Health Improvement Network (THIN)[9] database, containing 11 million British English patient records from general practices (Lewis et al. 2007).
- A Bulgarian clinical corpus containing several hundred thousand patient records from the specialties general practice, endocrinology, metabolic disorders cardiology, ophthalmology, gastroenterology, pneumology and physical therapy used for text mining and big data analytics (Boytcheva et al. 2015).
- Another Bulgarian clinical corpus containing 500,000 pseudonymised outpatient records on diabetes (Boytcheva et al. 2017b).
- A Danish clinical corpus containing 61,000 psychiatric hospital patient records from Center for Biological Sequence Analysis (CBS), University of Copenhagen and Technical University of Denmark (Eriksson et al. 2013).
- Another Danish clinical corpus containing 323,122 patient health records used for de-identification (Pantazos et al. 2016).
- A Dutch clinical corpus called the EMC Dutch clinical corpus[10] (Afzal et al. 2014).
- A Finnish clinical corpus[11] containing 2800 sentences from nursing notes from the University of Turku, Finland.
- A French clinical corpus containing 170,000 documents from 2000 patients with a stay of at least 20 days, covering five different hospitals within one geographical area, and several medical specialties (e.g., pneumology, obstetrics, infectious diseases) (Grouin and Névéol 2014).
- Another French clinical corpus containing 59,285 French patient records from three French hospitals. In total the article mentions 115,447 records from six hospitals, including Danish and Bulgarian patient records. The records were used to detect adverse drug events, Chazard et al. (2011).
- Yet another French clinical corpus containing 1500 discharge summaries, half of them containing hospital acquired infections, which are described in Proux et al. (2011).
- A German clinical corpus from Austria, containing 18,000 patient records from eight different clinical units (surgery, vascular surgery, casualty surgery, internal medicine, neurology, anesthesia and intensive care, radiology and physiotherapy) which has been used for document classification (Spat et al. 2008).
- Another German corpus of 12,743 clinical narratives describing laboratory results of leukaemia (Zubke 2017).
- Yet another German corpus of 6817 clinical notes and 118 discharge summaries in nephrology (Roller et al. 2016).
- An Italian clinical corpus containing 23,695 patient records used for entity extraction and determination of semantic relations (Attardi et al. 2015).

[9]THIN database, http://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/database. Accessed 2018-01-11.

[10]EMC Dutch clinical corpus, http://biosemantics.org/index.php/resources/emc-dutch-clinical-corpus. Accessed 2018-01-11.

[11]Finnish clinical corpus, http://bionlp.utu.fi/clinicalcorpus.html. Accessed 2018-01-11.

- A Norwegian clinical corpus containing 7741 patient records encompassing in total 1,133,223 unstructured EHR text documents used for identification of cancer patient trajectories (Jensen et al. 2017).
- A Polish clinical corpus containing 1200 children's hospital discharge records (Marciniak and Mykowiecka 2014).
- A Spanish clinical corpus, the IXAMed corpus from the Galdakao-Usansolo Hospital, collected during 2008–2012 containing 141,800 patient records (Pérez et al. 2017).
- An Argentian-Spanish clinical corpus, containing 512 annotated radiology reports (Cotik et al. 2017).
- A Japanese clinical corpus containing 3012 discharge summaries from the University of Tokyo Hospital annotated for adverse drug events (Aramaki et al. 2010).

For a nice overview of research carried out in clinical text mining in languages other than English see Névéol et al. (2018).

4.8 Summary

Patient record text is different from standard text. Patient records contain plenty of misspellings (up to 10%) and domain specific abbreviations (up to 10%) and acronyms (up to 5%). Patient records also contain incomplete sentences, often the subject or patient is missing in the sentence. In the assessment field there are many negations (up to 10%) since the physician tries to exclude symptoms while reasoning to find the disorder of the patient. In addition to the negations, the content contains vague or uncertain expressions (up to 12%) regarding the factuality of the findings and disorders. Patient record text is written by different professions and also varies between different medical specialties. Discharge summaries are often wellwritten and structured, since they are written for a broader audience than the personnel at the clinical unit. Most available clinical corpora for research are in English; however, there are some corpora in other languages available for research.