Semantic Mining – Work Package 27 – Deliverable 1

Literature Review on Patient-Friendly Documentation Systems

Hans Åhlfeldt[4], Lars Borin[2], Philipp Daumke[1], Natalia Grabar[3], Catalina Hallett[5], David Hardcastle[5], Dimitrios Kokkinakis[2], Clara Mancini[5], Kornél Markó[1], Magnus Merkel[4], Christian Pietsch[5], Richard Power[5], Donia Scott[5], Annika Silvervarg[4], Maria Toporowska Gronostaj[2], Sandra Williams[5], and Alistair Willis[5]

[1]Universität Freiburg, Germany
[2]Göteborgs universitet, Sweden
[3]INSERM, France
[4]Linköpings universitet, Sweden
[5]The Open University, UK

19th May 2006

Contents

1 Overview **7**

2 Official guidelines on legal aspects **9**
2.1 Legal aspects in France . . . 9
2.1.1 Regulatory documents . . . 9
2.1.2 Rights under the law n° 2002-303 . . . 9
2.1.3 Procedures for gaining access to medical records . . . 10
2.1.4 Exemptions . . . 11
2.1.5 Records of dead patients . . . 12
2.2 Legal aspects in Sweden . . . 12
2.3 Legal aspects in the UK . . . 13
2.3.1 Regulatory documents . . . 13
2.3.2 Rights under the Data Protection Act 1998 . . . 13
2.3.3 Procedures for gaining access to medical records . . . 14
2.3.4 Exemptions . . . 15
2.3.5 Records of dead patients . . . 16

3 Terminology issues **18**
3.1 Differences between lay and expert terminologies and related problems . . . 18
3.1.1 What to call them: *patient*, *client*, *customer*, etc? . . . 18
3.1.2 Communicating clinical documents to patients . . . 18
3.1.3 Psychological aspects of communicating with patients . . . 20
3.1.4 Specific needs . . . 21
3.1.5 Patient understanding of medical terms . . . 23
3.1.6 Which words to use? . . . 26
3.1.7 Analysis of few concepts . . . 27
3.1.8 Matching patient vocabularies and problem lists with standard terminologies . . . 27
3.1.9 Conclusion . . . 30
3.1.10 Technical Term Translation . . . 30
3.2 Terminological Resources . . . 31

4 Computational Methodologies **33**
4.1 Existing Resources . . . 33
4.1.1 UMLS . . . 33
4.1.2 GALEN . . . 34
4.1.3 SNOMED . . . 34
4.1.4 MeSH . . . 34
4.2 Divergence of Patient and Practitioner Language . . . 34

4.3 Computational Methods in Terminology Management . . . 35
4.4 Construction and Maintenance of Ontologies . . . 36
4.5 Knowledge Representation . . . 36
4.6 Medical WordNet . . . 37
4.7 Analysis of records . . . 37
4.8 Text Simplification and Dialogue . . . 38
4.8.1 Text simplification . . . 38
4.8.2 Dialogue systems . . . 38
4.9 Automated acquisition of lexicon . . . 39
4.9.1 Acquisition of monolingual lexicon . . . 39
4.9.2 Acquisition of multilingual lexicon . . . 42

5 Survey of Natural Language Generation systems aimed at Patients or Doctors 43
5.1 Alphabetical list of systems with brief descriptions . . . 43
5.1.1 BabyTalk . . . 43
5.1.2 CLEF . . . 43
5.1.3 GRASSIC . . . 43
5.1.4 HealthDoc . . . 44
5.1.5 Linguistic String Parser . . . 44
5.1.6 MAGIC . . . 45
5.1.7 MDA (Multilingual Document Authoring) . . . 45
5.1.8 MeDView . . . 45
5.1.9 MIGRAINE . . . 45
5.1.10 PERSIVAL . . . 46
5.1.11 PIGLIT . . . 46
5.1.12 PILLS . . . 46
5.1.13 STOP . . . 47
5.1.14 SumTime-Neonate . . . 48
5.1.15 SUREGEN-2 . . . 48
5.1.16 TAS . . . 48
5.1.17 TraumAID/TraumGEN . . . 48
5.2 Comparison Tables . . . 48
5.3 Use of Multimedia . . . 59

6 Survey of laymen/expert (e.g. patient/doctor) ontology translation systems 60

7 Empirical studies with patients on information provision 62
7.1 Providing computer-based information . . . 62
7.2 Tailoring patient education materials . . . 62
7.2.1 Asthma Management Advice . . . 62
7.2.2 Cancer Information . . . 63
7.2.3 Dietary Advice . . . 63
7.2.4 Smoking Cessation Advice . . . 63
7.2.5 Mammography Advice . . . 64
7.2.6 Discussion . . . 64
7.3 Patients' understanding and recognition of medical terminology . . . 65
7.4 Use of pictures in patient information materials . . . 65

7.5 Clinicians' communications with patients . . . 65
7.6 Summary . . . 66

8 Corpus Annotation Tools 67
8.1 Requirements . . . 67
8.2 Forces . . . 67
8.3 Annotation Types . . . 68
8.4 Schema . . . 69
8.5 Tool Overview . . . 69
8.5.1 Orthographic Annotations . . . 70
8.5.2 Linguistic Annotations . . . 71
8.5.3 Semantic Annotation . . . 78
8.5.4 Discourse-Level Annotation . . . 78
8.6 Other tools of potential interest . . . 80
8.6.1 MontyLingua . . . 80
8.6.2 Unitex . . . 81
8.7 IDEs and Workbenches . . . 81
8.7.1 IMS Corpus Workbench (CWB) . . . 81
8.7.2 Annotate . . . 81
8.7.3 Alembic Workbench . . . 81
8.7.4 CLaRK . . . 81
8.7.5 GATE . . . 81
8.7.6 RASP . . . 82
8.7.7 Stanford Parser . . . 82
8.8 Survey of existing corpus annotation tools (Sweden) . . . 82

9 Survey of Corpora of Patient Information 83
9.1 Previously used and/or annotated corpora . . . 83
9.1.1 Identified corpora . . . 83
9.1.2 Unidentified corpora . . . 86
9.2 Non-annotated corpora . . . 87
9.2.1 From experts to laymen . . . 87
9.2.2 From expert to expert . . . 88
9.2.3 From novice laymen to novice laymen . . . 89
9.2.4 From expert laymen to novice laymen . . . 89
9.3 Survey of web portals addressing patient information needs (Sweden) . . . 90

10 Internet as Corpus 92
10.1 World Wide Web . . . 92
10.2 Usenet . . . 94
10.3 Email . . . 95
10.4 Internet Relay Chat . . . 96
10.5 Conclusion . . . 98

Index of Acronyms 99

Index of Authors 102

Bibliography **109**

1 Overview

This literature review forms a deliverable in the European Network of Excellence on Semantic Interoperability and Data Mining in Biomedicine. More specifically, it is part of a work package (WP27) which aims to develop and evaluate generic methods and tools for assisting patients to understand their health and healthcare by generating patient-friendly readable texts that paraphrase the content of their electronic health records. We have reviewed the literature in topics that we consider to be relevant to this work package. When appropriate, we cover variations in conditions in the four countries of the collaborating research groups (France, Germany, Sweden and the UK) and we cover corpora, tools and language technologies for the European languages of interest to these groups.

First, we consider legal issues involved in patients gaining access to their medical records. Who can view the records? What data do they have the right to access? Are there any data that patients cannot access? Who can access records of dead patients? What about security and data protection? See chapter 2 for brief surveys of the current state of affairs in France, Sweden and the UK.

Patient records are packed with jargon, acronyms and medical terms that clinical staff know and understand. Often there is a learning curve before patients become familiar with medical terms associated with their own particular illnesses and they may require more familiar words and phrases to describe medical concepts in an accessible form. The development of large-scale medical term banks, thesauri and lexicons (e.g. UMS SPECIALIST and Metathesaurus) enable Language Technology developers to generate reports for medical staff, but how can we communicate the same concepts to patients? We review the current literature on communicating technical medical terms in everyday language for patients and related issues. See chapter 3.

Our survey on computational methodologies for generating patient-friendly texts included the following topics: extraction of terminologies from corpora, comparison of terms from different sources, automatic analysis of patient records, use of ontologies, logics to model terminological use and changes, text simplification and dialogue systems. See chapter 4.

There have been many past NLG systems that generated output aimed at patients or doctors. We present an overview of these systems and compare them in 5 dimensions:

1. application area,
2. knowledge used (domain knowledge, generic medical knowledge and linguistic knowledge),
3. user models and personalisation,
4. system evaluation,
5. use of hypertext

See chapter 5.

We have reviewed work, particularly in medical informatics, on automatic translation of technical medical language into language aimed at patients. Chapter 6 presents a survey of such systems.

A number of empirical studies with patients have focused on the question of whether the provision of personalised information for patients is superior in various ways to general information. Will personalising information help patients be better-informed? Will it help them manage their illnesses better and comply with medical guidelines? Will it help them to take their medication in the correct manner? Will such information ultimately reduce hospital admissions? See chapter 7 for a survey of this literature.

Our survey of existing corpus annotation tools, see chapter 8, describes existing tools and what they do. It also includes their availability, the languages they cover, their formats, platforms and locations. The tools are classified into:

1. Tools for orthographic annotations (document information and document structure),
2. Tools for linguistic annotations (e.g. tokenization, stemming, morphology, POS, syntax),
3. Tools for semantic annotations (discourse-level, semantic tags and UMLS tags),
4. Workbench and IDE tools.

Our survey of existing corpora of patient information addressing patient information needs includes corpora that language engineers have used in the past in building systems as well as general linguistic corpora, medical corpora and others. See chapter 9.

Chapter 10 surveys the Internet as a corpus, including access to and potential use of the web, Usenet, email and Internet relay chat. See chapter 10.

2 Official guidelines on legal aspects

2.1 Legal aspects in France

2.1.1 Regulatory documents

Throughout France, the rights of access by living people to their health records are as set out in the Law on public health n° 2002-303 adopted on 4 March 2002 and its modifications. This information is set out precisely in the Titre II, Chapitre II, Article 11, Chapitre 1 "Informations of users of the health system and their will". Légifrance[1] publishes the corresponding law and its modifications.

AP-HP[2] (Assistance Publique - Hôpitaux de Paris) publishes documents in relation to:

- patients' rights for access to their health records,
- procedure that must be followed for obtaining health record documents.

From this site it is possible to download a booklet which describes this information as well as a questionnaire to be filled in when asking a hospital for health documents.

2.1.2 Rights under the law n° 2002-303

According to the law n° 2002-303, any patient can access all the information relevant to his health, as held by medical professionals or hospitals. This information can be any document which is formalised and has been used for the definition and follow-up of the patient's diagnosis and treatment, or prevention, or any written communication among medical professionals. This information concerns:

- proposed investigations, treatments or prevention acts,
  - their usefulness
  - their urgency
  - their consequences
  - their frequent and severe known risks can normally be foreseen
- other possible solutions in the case of refusal of previously proposed treatments
- if new risks of applied treatments are identified, the patient concerned has to be informed unless it is impossible to find him.

When leaving the hospital, a patient can ask to obtain documents from his health record:

1. In any case, the patient receives the *leaving order* at the end of his hospitalisation. This order is completed with information useful for his further care. This information

[1] `http://www.legifrance.gouv.fr/`
[2] `http://www.aphp.fr/`

can also contain some practical information and indicate the date of the next medical consultation.

2. The *patient discharge* document is sent to a medical doctor as indicated by the patient, together with information useful for medical follow-up. The maximum deadline is 8 days. A patient can ask for a copy of these documents.

3. The *patient's health record* can contain numerous documents which can be communicated to the patient. If desired by the patient, documents of this record can be sent to his medical doctor. Main documents of the health record are the following:

   · a medical document which indicates reason(s) for hospitalisation

   · a patient discharge which indicates the patient diagnosis at the time of leaving the hospital. The discharge can contain as well:

     – results of main clinical examinations realized during the hospitalisation

     – report(s) on different para-clinical examinations and of main complementary examinations

     – indications and precautions to take

     – surgical or labour discharge(s)

     – documents related to anaesthesia

     – therapeutic prescriptions and orders, including the leaving order

     – documents resuming the transfusion acts

     – file resuming the nurse care

     – depending on the case, other significant documents (scans, etc.)

If a patient doesn't desire to be informed about his diagnostic or his prognosis, this will must be respected, except if a third party runs a risk of contagion.

2.1.3 Procedures for gaining access to medical records

The patient can access his medical documents by:

1. *Consulting them in the hospital.*
   The patient must schedule an appointment with the medical department concerned. This appointment can be planned directly with the department or through the hospital administration. If the patient desires it, a physician can answer his questions about the topics and terms mentioned in the records. During this appointment, the patient can obtain copies of document(s) he wants.

2. *Receiving them by post.*
   The patient must first send the written form to the hospital department concerned and, if possible, dates of hospitalisation and patient ID number. When filling in the form, following fields have to be completed:

- personal patient information (names, postal address, birth date)
- a third party has to be identified (father, mother, legal representative or eligible party)
- necessary documents
- postal address (patient, third party, medical doctor)
- reason of the request in cases the patient is deceased
- date and signature

When identifying a patient or third party, identity papers have to be sent. And the third party has to legally certify himself as legal representative or eligible party.

Phone requests are not possible, because of the confidential nature of documents delivered.

Consulting health records in the hospital is free. Copying them and sending them by post mail will be billed (current reference fee is set to 0.18 euro per page).

The patient's family can be informed about his health at dates and hours specific for each department, except if the patient disagrees with this. Phone communication of such information is forbidden.

The transmission to patient of his health information often involves expert explanations. Medical doctors can ask for the transmission of such documents to be sent with specific precautions. But this cannot prevent the department from delivering these documents.

If the medical doctor who prescribed the patient's hospitalisation asks to see the health record of this patient, this can be done only when the patient, or his legal representative, agrees with this.

Patient health records are stored in the hospital for at least 20 years (current regulation). Patients can obtain only copies of their original documents. The time necessary for the communication of requested documents is:

- at least 48 hours (time for reflexion on behalf of the medical staff),
- at most 8 days when documents have been produced less than 5 years ago,
- no more than 2 months when documents have been produced more than 5 years ago.

The patient can access these documents directly or through the agency of a medical doctor he indicates.

At any time, clinical professionals are available to answer all the patient questions related to this topics.

2.1.4 Exemptions

As health record documents are protected by confidentiality, some rules can prevent them being accessed.

Documents from health records cannot be communicated to the patient when they contain:

- information containing the name of third party (*ie.* relative to some patient acquaintances):

- communicated by a third party, who is not involved in the therapeutic treatment
- concerning a third party
- confidential information which cannot be communicated to a third party.

It seems that psychological observations about patient are not delivered to him.

2.1.5 Records of dead patients

When asking for communication of health documents on a deceased patient, the reason of the request has to be specified.

Documents from the health record of a deceased patient can be communicated to his eligible party only when using them in order to:

- know the reason for the death
- protect the memory (recollection) of the patient
- assert rights of an eligible party

Nevertheless, the patient should not have been opposed to this while he was alive.

If a hospital department refuses to deliver these documents to an eligible party, this refusal has to have a reason. Anyway, it cannot prevent the delivery of the medical certificate, especially when this certificate does not contain information recognised as being a medical secret.

2.2 Legal aspects in Sweden

On The Swedish National Board of Health and Welfare (*Socialstyrelsen*) web site under "Publicerat" / "Published" there are about 30 documents intended for the general public. Some of these documents are written in "easy Swedish" and are particularly intended for immigrant learners, people with low literacy skills and the elderly. A number of these documents are also available in an English version. *The Social Services Act* (*Socialtjänstlagen*) is available in Arabic, English, Finnish, Persian, Serbo-Croatian, Spanish and easy Swedish. *The Social Services Act* in Swedish: `http://www.sos.se/fulltext/114/2002-114-3/2002-114-3.pdf`, and in English: `http://www.sos.se/sose/sos/general.htm`.

On the official gateway to Sweden `http://Sweden.se/` there is a section intended to inform the general public (primarily outside Sweden) about the health care system in Sweden while under "Society & Welfare" there are several fact sheets about a number of policies e.g. disability, public health objectives etc. Some of the documents are available in multiple languages.

The Swedish government's national IT strategy in healthcare published in 2006 on patient safety and quality of care: *Nationell IT-strategi för vård och omsorg*
`http://www.sweden.gov.se/content/1/c6/05/96/62/abac6cb0.pdf`.

2.3 Legal aspects in the UK

2.3.1 Regulatory documents

From 1 March 2000 throughout the UK the rights of access by living people to their health records whether computerised or manual are as set out in the Data Protection Act 1998 and its regulations. It applies equally to all records regardless of when they were made. Its provisions supersede the previous rights of access under legislation specific to health records, such as the Access to Health Records Act (1990) and Access to Health Records (Northern Ireland) Order (1993). The access to medical records of deceased patients is not covered by the Data Protection Act 1998, instead former regulations imposed by the aforementioned Access to Health Records acts apply. Specific rights in respect of medical reports written for insurance or employment purposes are covered by separate legislation which applies to reports written by doctors who are or have been involved in the subject of the report's clinical care and treatment.

The British Medical Association (BMA) publishes guidelines for medical practitioners regarding the disclosure of medical records in view of the Data Protection Act 1998. Most of the information below comes from BMA official documents, accessible from http://www.bma.org.uk.

2.3.2 Rights under the Data Protection Act 1998

According to the Data Protection Act, art. 68, sec. (2):

> *"health record" means any record which —*
>
> *(a) consists of information relating to the physical or mental health or condition of an individual, and*
>
> *(b) has been made by or on behalf of a health professional in connection with the care of that individual.*

This definition applies to all manual and computerised medical records, regardless of when they were produced. The BMA believes that this includes reports written by doctors who examine patients for the sole purpose of writing a report and who have no other clinical relationship with the patient. This interpretation rests on the definition of 'care', which is said to include examination, investigation, diagnosis and treatment. Thus a doctor who writes a report following an examination does so in connection with that patient's 'care' and as such makes what the Act defines as a 'health record'.

Under the Data Protection Act 1998, sec. (7) patients have the following rights:

> *(a) to be informed by any data controller whether personal data of which that individual is the data subject are being processed by or on behalf of that data controller,*
>
> *(b) if that is the case, to be given by the data controller a description of —*
>
> *(i) the personal data of which that individual is the data subject,*
>
> *(ii) the purposes for which they are being or are to be processed, and*
>
> *(iii) the recipients or classes of recipients to whom they are or may be disclosed,*
>
> *(c) to have communicated to him in an intelligible form —*

> *(i) the information constituting any personal data of which that individual is the data subject, and*
>
> *(ii) any information available to the data controller as to the source of those data, and*
>
> *(d) where the processing by automatic means of personal data of which that individual is the data subject for the purpose of evaluating matters relating to him such as, for example, his performance at work, his creditworthiness, his reliability or his conduct, has constituted or is likely to constitute the sole basis for any decision significantly affecting him, to be informed by the data controller of the logic involved in that decision-taking.*

Although the BMA believes that under these regulations patients have the right to request access to independent medical records, this has been challenged and is a matter of legal dispute.

2.3.3 Procedures for gaining access to medical records

Competent patients may apply for access to their own records, or may authorise a third party, such as their lawyer, to do so on their behalf. Parents may have access to their child's (under 18, or, in Scotland, under 16) records if this is in the child's best interests and not contrary to a competent child's wishes. If a person is not capable of giving his or her permission, a court-appointed person with 'welfare power of attorney' may have access to information necessary to fulfil their function.

Exemptions to this rule are outlined in section 2.3.4.

Normally, a medical record will be released to a patient following an official written request from the patient or their legal representative. However, nothing in the law prevents doctors from informally showing patients their records or, bearing in mind duties of confidentiality, discussing relevant health issues with carers. This follows a long standing practice of disclosing to patients information about their own health history.

Requests for access are made to the person in charge of keeping the records; the data controller. This is usually the health professional responsible for the patient's care, but may in some circumstances be another health professional or, for example, a member of records management staff.

Irrespective of who is the data controller, decisions about disclosure must be made by the 'appropriate health professional'. This is usually the health professional currently or most recently responsible for the clinical care of the patient in respect of the matters which are the subject of the request. If there is more than one, it should be the person most suitable to advise. If there is none, advice should be sought from another health professional who has suitable qualifications and experience.

The courts have the power to order disclosure or non-disclosure. Patients or other people likely to be affected by disclosure (for example a person likely to suffer serious harm if information is disclosed) can apply to the courts.

The health provider is entitled to charge a fee for the release of the medical record. Fees vary with the type of record the patient requires and whether they want a hard copy or just to see them.

Access is available to all records whenever they were made. Unlike previous legislation there is no date restriction. Health records and any information as to the source of information in

them (for example the identity of a health professional who has contributed to the record) must be communicated to patients in an intelligible form.

Patients are also entitled to a permanent copy of the information, for example a print out of computerised records or a photocopy of manual records. The copy must be accompanied by an explanation of any terms which are unintelligible. The Act does not require a permanent copy to be provided if this is impossible or involves disproportionate effort. Even if providing a permanent copy is impossible, the law still requires the patient to be shown the records and the relevant explanations of terms given.

When records are requested, those supplied must be those in existence at the time of the request. There can be amendments or deletions between the request and the supply of the records provided these would have been made regardless of the request.

Patients are entitled to be informed of the logic involved in any automatic decisions made about them, for example decisions made by a computer system.

2.3.4 Exemptions

Certain information, described below, must not be released, and there is no obligation to inform patients if information is withheld on any of these grounds. There is still an obligation to disclose the remainder of the records. The main exemptions are that information must not be disclosed if it:

- is likely to cause serious physical or mental harm to the patient or another person; or
- relates to a third party who has not given consent for disclosure (where that third party is not a health professional who has cared for the patient).

The BMA guidelines defines in more detail the specific cases where the disclosure of information is illegal or not advisable.

Third parties Where records contain information which relates to an identifiable third party, that information may not be released unless:

- the third party is a health professional who has compiled or contributed to the health records or who has been involved in the care of the patient (thus there is no requirement to contact other health professionals who have contributed to records, or whose correspondence is part of the records, although this may be helpful in some cases); that other individual gives consent to the disclosure of that information; or
- it is reasonable to dispense with that third party's consent (taking into account duty of confidentiality owed to the other individual, any steps to seek his or her consent, whether he or she is capable of giving consent and whether consent has been expressly refused).

Where health records include information from other identifiable sources, it is advisable to distinguish this information in the records to avoid inadvertent disclosure. Doctors must still disclose as much of the information in the records as is possible without revealing the identity of the third party. The Act suggests that this might be done by omitting names and identifying particulars from the records before disclosure, and care should be taken to ensure that the information is genuinely anonymous. Doctors are not required under the Act to approach a third party for consent to disclosure, although may wish to in some circumstances.

Harm Access must not be given to any information which, in the opinion of the appropriate health professional, would be likely to cause serious harm to the patient or another person The decision about likely harm must be taken by the appropriate health professional, usually the treating doctor. Circumstances in which information may be withheld on these grounds of harm are extremely rare. This exemption does not justify withholding comments in the records because patients may find them upsetting. The BMA advises that if harm could arise from providing access, advice from others involved in providing care may be helpful in assessing the nature and extent of the risk. For example it is particularly recommended that psychiatrists and GPs liaise before psychiatric records are released although there is generally no duty to inform or seek advice from any other health professional.

Confidentiality When a third party applies for access on behalf of a patient no information can be disclosed which the patient had provided on the understanding that it would be kept confidential or about which the patient had requested non-disclosure. Similarly, no results of examinations or investigations can be disclosed if the patient had expected the results to be kept confidential.

Legal privilege Access may not be given to records which are subject to legal professional privilege or, in Scotland, to confidentiality as between client and professional legal advisor. This may arise in the case of an independent medical report written for the purpose of litigation.

Court proceedings The courts have the power to restrict access to information as to the physical or mental health or condition of the patient supplied to the court in a report or other evidence from a local authority, Health and Social Services Board, Health and Social Services Trust, probation officer or other person in the course of certain family and children court proceedings.

Fertility treatment When disclosing information under the Data Protection Act 1998, no information may be disclosed about the keeping or use of gametes or embryos. Similarly no information may be disclosed which reveals whether an identifiable individual was, or may have been, born as a result of fertility treatment (in vitro fertilisation or the use of donated ova, sperm or embryos).

Children The Act does not allow disclosure of information whose disclosure is already prohibited in legislation concerning adoption records and reports, statements of a child's special educational needs and parental order records and reports and (for Scotland only) information provided by the principal reporter for children's hearings. Doctors who believe their records may contain such information should seek legal advice.

None of these exemptions apply where the disclosure is required by law, or is necessary for the purposes of establishing, exercising or defending legal rights.

2.3.5 Records of dead patients

The Data Protection Act 1998 only covers the records of living patients. If a person has a claim arising from the death of an individual, he or she has a right of access to information

in the deceased's records necessary to fulfil that claim. These rights are set out in the Access to Health Records Act 1990 or Access to Health Records (Northern Ireland) Order 1993.

Any person with a claim arising from the death of a patient has a right of access to information covered by the Act and directly relevant to that claim. No information which is not directly relevant to the claim may be released. Thus a personal representative or executor can access information to benefit the deceased's estate, as can an individual who was a dependant of the deceased and who has a claim relating to that dependency which has arisen from the death.

The Access to Health Records Act 1990 covers manual health records made since 1 November 1991. In Northern Ireland the corresponding legislation, the Access to Health Records (Northern Ireland) Order 1993, covers manual records since 30 May 1994. Access must also be given to information recorded before these dates if this is necessary to make any later part of the records intelligible. There is no statutory right of access to records of deceased patients which fall outwith the time period covered by the legislation.

There are certain exemptions to this right, and information may be withheld if:

- it identifies a third party without that person's consent unless that person is a health professional who has cared for the patient;
- in the opinion of the relevant health professional, it is likely to cause serious harm to somebody's physical or mental health; or
- the patient gave it in the past on the understanding that it would be kept confidential. Similarly no results of examinations or investigations which the patient thought would be confidential at the time they were carried out can be disclosed. No information at all can be revealed if the patient requested non-disclosure.

After a patient's death the health records may be held by the Primary Care Trust or Central Services Agency. These bodies are required to take advice before making a decision about disclosure.

Access can be given by allowing the applicant to inspect the records or extract or by supplying a copy if this is requested.

The courts may enforce compliance with the legislation if access is not given within the required time limits. The court may also require that the records be made available for its own inspection in order to come to a decision.

3 Terminology issues

3.1 Differences between lay and expert terminologies and related problems

The communication between care givers and patients has been a research topic for a few decades, as early Medline citations (since the seventies) attest. The success of communication between these two social groups is, first of all, the main condition of the successful heath care of every patient. Actually, this success depends on his understanding of his diagnosis, usefulness of the provided medical care, clinical trials, medical research, medication prescribed, etc., and on the confidence the patient has in his doctor.

When reviewing literature on these questions, several topics can be distinguished, and we briefly present them here. Notice that the research on this topic is very widespread as scientists from many countries are involved (Sweden, Hungary, France, USA, UK, Brasil, Germany, India, Israel, Scotland, Norway, etc.)

In general, a lot of work has shown that there is an important difference between patient's and medical doctor's knowledge and languages. Sometimes, nurses and medical students seem to represent the middle position between patients and doctors.

3.1.1 What to call them: patient, client, customer, etc?

There are very few references on the question of what to call patients: *patient*, *client*, *customer*, *consumer*, *user*, *recipient* (Herxheimer & Goodare, 1999; Neuberger & Williams, 2001; Naseem, Balon, & Khan, 2001; Ramdass et al., 2001; Mulhall, Ahmed, & E, 2002; Wittich, Junnila, & Buller, 2003; Kehler & Rice, 2004; Deber, Kraetschmer, Urowitz, & Sharpe, 2005). Although some people strongly dislike the label *patient*, most people across different countries still prefer to be called *patient*. This indicates the desire to maintain of doctor-patient relationship rather than commercial relationship through words like *client*, *customer*, etc. This means also that a doctor-patient relationship involves a more complex interaction than simply a market transaction (Mulhall et al., 2002). Physicians prefer even to call their patient by their last or first name (Naseem et al., 2001). On the other side, patients prefer to call their physicians *doctors*. But these questions are still challenging as the discussion will go on for many years.

3.1.2 Communicating clinical documents to patients

There is common agreement on the point that clinical documents should be communicated to patients, because informed patients are better equipped to participate in their own health care decisions (Bouhaddou & Warner, 1995). But notice that the terminology of these documents should be adapted to patient knowledge.

To discover the views of patients, patient representatives and doctors on copying referral letters to patients, White, Singleton, and Jones (2004) realized a three-part study:

- an analysis of 50 GP referral letters against a standard template;
- 35 patient interviews using a semi-structured questionnaire in outpatient waiting rooms;
- 3 focus groups of patients, patient representatives and doctors.

There was general agreement that copying referral letters to patients could improve information and decision sharing with patients. Copying referral letters could provide an opportunity for patients to correct mistakes, prepare for their appointments and have a personal record that they could keep and show to others. However, there were concerns about letter content, particularly medical terminology, character judgements and "sensitive" patient information. It was also recognised that providing more information to patients could increase patient anxiety. The style and content of some referral letters may need to change. This is particularly relevant where certain types of information included in referral letters could cause distress for patients or influence the time that patients have to wait for their outpatient appointments.

Scanlan, Siddiqui, Perry, and Hutnik (2003) are interested in determining patients' understanding and opinions about the usefulness of the informed consent (IC) document for cataract surgery and evaluate the deterioration in the effectiveness of verbal and written IC over time. The study showed that patients about to consent to cataract surgery had a reasonable grasp of basic terminology. A standardised IC discussion was effective in educating patients. Patients considered IC to be important and expected all pertinent information to be communicated. Patient recall of outcome probabilities was poorer than that of nonnumeric facts; however, memory decay may be slowed by providing supporting take-home literature.

The work presented by Waisman et al. (2003) is led by the hypothesis that understanding discharge instructions is crucial to optimal healing but may be compromised in the hectic environment of the emergency department. It shows that full understanding was found in 72 % and 78 % of the parents at the respective centres for the diagnosis, and in 82 % and 87 % for the treatment instructions. There was no statistical correlation between level of understanding and parental age, gender, education, level of anxiety before or after the ED visit, or time of day. The most contributory factor to lack of understanding was staff use of medical terminology. Parents suggested further explanations by a special discharge nurse and written information as auxiliary methods. There remains a considerable number (about 20 %) who fail to fully comprehend the diagnosis or treatment directives. This subset might benefit from the use of lay terminology by the staff, institution of a special discharge nurse, or use of diagnosis-specific information sheets.

Miyawaki, Takada, Furukawa, and Adachi (1995) show that much content which may be hard to understand without having expert knowledge is related to the temporal change in dentofacial structures known as the growth, development and physiological aspects of masticatory apparatus. This work shows as well that consultation of multimedia software can provide a comfortable environment to the patients and their families to learn where the orthodontic problems lie and how they could be solved.

An inquiry made by Habeck, Engel, and Munstermann (1977) by means of a questionnaire among 1043 persons (566 in-patients, 269 employees of industrial firms, 208 teachers) revealed: 32.5 % were "often not" and 26.5 % were "nearly always" satisfied with the medical information provided by their doctors. They were mainly interested in the causes of complaints (77.7 %), in the prognosis (66.4 %), in the effect of the medicament (44.5 %) and in the diagnosis, using technical terminology (37.7 %). Nearly all of them wanted to know about the contents of the specialist's report: 35.5 % would like to read the report for themselves and 61.4 %

wanted an intelligible explanation by the family doctor. In incurable malignant disease, 49.5 % preferred the patient to be fully informed, whereas 24.4 % wanted information about a serious illness which might lead to death. Other questions concerned the information of relatives and between doctors.

The effect of giving hospitalised medical patients access to their problem-oriented hospital records was investigated by Stevens, Stagg, and Mackay (1977). There was no measurable effect of seeing the record on the subjects' ability to list their diagnoses or medications, their self-assessment of depression, anxiety or contentment, or their attitudes toward selected components of the health care system. On the other hand, in individual instances access to the hospital record seemed to facilitate communication and provide an opportunity for hospital in-patients to monitor objectively their hospital course.

Golodetz, Ruess, and Milhous (1976) show that patients were generally comfortable about reading the record, found it educational and appreciated the trust implied. No substantial difficulties arose. Few records were expurgated. The staff has accepted this style as crucial to an appropriate sharing of responsibility between themselves and the patients. Authors conclude that giving the patient his record is a safe and inexpensive aid to the rehabilitation process, and is probably mandated by the changing relationships between professionals and their clients, and by the patient's need to negotiate his own health care in an increasingly complex and mobile society.

3.1.3 Psychological aspects of communicating with patients

The communication between these two social groups can become more difficult because of its psychological impact. Areas in which such problems arise, are mostly related to psychic and reproduction cares.

In early pregnancy loss, Cameron and Penney (2005) recommend to say *miscarriage* instead of *abortion*, *blighted ovum*, *incompetent cervix* or *pregnancy failure*. In other studies, the term *infertility* receives a very negative connotation and is hardly accepted by patients (Madsen, 2005; Davies, Delacey, & Norman, 2005). In the same area, Mullin, Mills, and Kirkman (2004) report that patients preferred to use familiar terms for contraceptive names, e.g. *pill*, *mini-pill*, *coil* and *morning-after pill*. Although precise terms were not widely known or understood, when used they were associated with more information than were the familiar terms.

As words can inadvertently signal negative messages, Hitt (1998) considers that medical caregivers should be careful when dealing with HIV-positive individuals: use phrases that are neutral and non-judgemental.

In connexion with psychic disorders, Karla (2003) recommends not to use the term *depression*. Changing this term would reduce its negativity. More generally, Shackle (1985) argues that the harmful effects of psychiatric diagnosis, such as social stigmatisation and possible loss of freedom, obligate clinicians to minimise the harm done to patients even when mental illness is correctly diagnosed. One way of achieving this goal is to discuss fully with patients the details of their diagnosis in language that they can understand. In so doing, psychiatrists may enable patients to acquire a conceptual framework with which to make sense of their illness, satisfy their cognitive needs, reduce their sense of isolation, and provide them with a route to restoring their mental health.

A specific case has been reported by Steensma (2006). It is related with the *myelodysplastic syndromes* (MDS), and physicians vary in whether they refer to MDS as a *cancer* when

discussing the diagnosis with patients. Patients who carry one of the dubious cancer-specific health insurance policies are usually not eligible for financial benefits when they receive a diagnosis of MDS. Likewise, patients with MDS who have been led to believe they do not have a form of cancer by their primary physician may become upset when seeing another health care provider who does refer to MDS in this way.

3.1.4 Specific needs

In this section, we present previous work where the differences between patient and expert terminologies and knowledge can present difficulties in different specific areas: information retrieval, formulation of problems during consultations, building of problem list, clinical trials and judgement about hospital quality.

Information retrieval In the context of information retrieval (IR) and in order to facilitate consumer health information seeking, retrieval and understanding, Zeng and Tse (2006) undertake the comparative study of this vocabulary. The precise content of this study can be found in the full paper. During the same project, the researcher's group decide to build a tool to assist people in health-related queries formation (Zeng et al., 2006). Thus, 213 subjects have participated in the study. Researchers observed the improvement of successful queries, but found no statistically significant impact on user satisfaction. The same tool has been evaluated through the matching to the UMLS (Plovnick & Zeng, 2004). Such reformulation showed that 42 % of queries gave better results, 19 % worse results, and 39 % showed no difference. Otherwise, this study allowed to identify that ambiguous lay terms, expansion of acronyms, arcane professional terms have an impact on performances of the reformulation system.

In another IR application, developed on the basis of the French medical gateway CISMeF, the reformulation of queries is done through already registered synonyms of MeSH terms (Soualmia, Darmoni, Douyère, & Thirion, 2003).

The PERSIVAL project (McKeown et al., 2001) also proposes access for patients to online information that can help them to understand their medical situation.

When patients formulate their problems When patients are unable to express all their major concerns, they are less likely to follow the physician's prescribed treatment plan and they are less satisfied (Larsen, Risor, & Putnam, 1997). We report here few studies on this point.

Dyche and Swidenski (2005) analyse different situations during medical visits. As physicians often interrupt the patient's opening statement, researchers decided to observe the impact of such interruption on physicians understanding of patients.

1. in 26 % patients have not been interrupted,
2. in 37 % patients have been interrupted,
3. in 37 % no question has been asked during the first 5 min.

Results show that there is no significant difference between 1 and 2, but when no question is asked during the first 5 min (3) the understanding between physicians and patients is worse.

The work presented by S. L. Smith and Hamm (1998) observes the agreement between patient and physicians. Twelve patients identified an average of 4.58 problems, the physician

identified an average of 7.75 problems, and the other professionals identified an average of 6.54 problems. This lack of agreement has implications for patient education and certification as full collaborators in primary care.

On the basis of 439 return primary care visits, Freidin, Goldman, and Cecil (1980) also studied the agreement between the patients and physician as for the patient problems detected. The concordance was defined as complete when both cited a problem in the same organ-system (208 visits, 47 %), as partial when the patient cited a problem that was anywhere but first on the physician's problem list but both parties agreed on the biological or psychosocial nature of the principal problem (114 visits, 26 %), or as absent (117 visits, 27 %). Concordance scores were significantly lower when physicians identified a principal psychosocial problem or when patients identified a principal problem related to psychological issues, preventive medicine, the musculoskeletal system, or accidents.

Study of side-effects perceptions was lead by Hyland and Stahl (2004). It showed some differences between clinicians' and patients'/parents' perceptions of treatment. For patients, side effects meant long-term effects (10–20 years), for clinicians, it meant occasional local problems.

Thus, during medical visits, Assal, Lacroix, and Aufseesser-Stein (1992) recommend to physicians: listen more attentively, avoid interruptions, excuses, simplifications, interpretations and premature information and to learn to repeat the flow of thoughts of the patient; inform more adequately without usage of medical terminology, abstain from confused explications, frequent change of subject, imprecision and attitudes forcing the patient into an inferior role; improve teaching of the patient by proposing distinct goals and by inviting him to find solutions to his problems by learning from errors, all this without being patronising and being considerate towards the attitude of the patient towards his disease.

Patient problem list The problem lists comprise terms suitable to describe patient's concerns. The collection of these terms is often bootstrapped manually, on the basis of medical consultation notes made by physicians. But for different medical and economic purposes, these terms must be matched to existing terminologies (sec. 3.1.8).

Bui, Taira, El-Saden, Dordoni, and Aberle (2004) propose to automatically generate medical problem lists on the basis of ICD-9: given a set of ICD-9 codes associated with a patient records, the system maps the codes (and related data) to an anatomy-centric hierarchy.

With regard to a maternal record system, described in J. Miller, Driscoll, Kilpatrick, and Quillen (2003), patient problems were (manually) assigned a unique term that corrected for spelling, spacing, synonyms, and abbreviation variations. Analysis of these terms suggested design changes that would increase the number of automated problem entry functions in the hospital's record system.

Lauteslager, Brouwer, Mohrs, Bindels, and Groundmeijer (2002) address the question of to what extent the problem list can be improved when general practitioners ask their patient about his or her own medical problems. This study showed that patients were in agreement with 88 % of all listed problems. The completeness of the problem list could be increased by 28 %, while 4 % ultimately were removed.

Shared decision making between patients and their health care providers and the inclusion of patient preferences in patient care have been, in theory, embraced as models for good clinical practice. Ruland and Bakken (2001) show that this integration requires that patient preference-related concepts be represented non-ambiguously and in a manner that renders

them suitable for computer rather than human processing. As results, the use of the LOINC semantic structure as a terminology model to create fully specified names for a sample of 15 preference elicitations from 8 published research articles.

Clinical trials Clinical trials are another context where the lack of understanding has negative impact (Helgesson, Ludvigson, & Stolt, 2005), and where researchers try to improve the nature of guidelines on written information for trial participants (Stead, Eadie, Gordon, & Angus, 2005).

Hospital quality Cheng, Ho, and Chung (2002) investigate Taiwanese patients' ability to judge hospital quality and to examine their knowledge of commonly used quality indicators. A total of 31–50 % of the participants claimed that they could judge a hospital's quality on the basis of medical equipment, technical competence, or medication. The most frequently mentioned reasons on which their judgements were based were related to their own experiences and to the hospital's reputation. The percentage of participants reporting that they understood the quality indicators was 6.7–42.1 %. In general, patients lack the ability to judge hospital quality and are unfamiliar with the commonly used quality indicators. Public education should be enhanced, or more understandable indicators should be developed in the future.

3.1.5 Patient understanding of medical terms

A lot of work address the ability of patient to understand given medical terms in different medical areas. The ability of patients to understand medical slang is usually recognised as useful. The extent of this ability may vary. Thus the Nordby (2003) analysis implies that it is sufficient that patients have a minimal understanding. But according to Egerod (2002), when the terminology is unclear, the indications, interventions and outcomes become unclear too. This last opinion is usually shared by researchers. In some studies, there is a correlation between understanding the medical slang by patients and their social position.

Lerner, Jehle, Janicke, and Mosati (2000) carried out studies in an emergency department, where patient were asked to explain common medical terms used by health care providers. Several observations were found:

- The mean number of correct responses was 2.8.
- The percentage of patients that did not recognise analogous terms was:
  - 79 % for *bleeding* vs. *hemorrhage*
  - 78 % for *broken* vs. *fractured* bone
  - 74 % for *heart attack* vs. *myocardial infarction*
  - 38 % for *stitches* vs. *sutures*
- The percentage that did not recognise non analogous terms was:
  - 37 % for *diarrhoea* vs. *loose stools*
  - 10 % for *cast* vs. *splint*

As conclusion, the authors consider that medical terminology is often poorly understood, especially by young, urban, poorly educated patients. Emergency care providers should remember that even commonly used medical terminology should be carefully explained to their patients.

The work described in McCormack, Evoy, Mulcahy, and Walsh (1997) shows that many patients in orthopaedic departments willingly consent to procedures that they do not fully understand. This implies that there is an element of trust involved in the process of giving consent.

Gittelman, Mahabee-Gittens, and Rey (2004) studied the understanding of common medical terms by parents. Caregivers agreed on the definitions of *diarrhoea*, *constipation*, *dehydration*, *fever*, and *seizure*. However, *diarrhoea* and *constipation* were mainly defined by either *stool consistency* or *frequency*, not both. *Dehydration* was appropriately defined as *lack of body fluids* (92 %), but many parents had difficulty identifying more than one sign of dehydration. *Fever* was thought to be an *elevated body temperature* (76 %), yet 69 % felt that a temperature less than 100.5 degrees F was considered a fever. Most respondents did not know the definitions of *meningitis* (70 %), *lethargy* (64 %), and *virus* (40 %). Although commonly used in everyday conversation, there seems to be a large disparity between a caregiver's perception and the actual definition of medical terms. More precise communication may help both parties to understand the true situation.

Blake, Weber, and Fletcher (2004) report that three (2.7 %) of the 111 adolescent participants provided an accurate definition of the term *Pap smear*. Sixty-eight percent mistakenly believed that a *Pap smear* was the same as a *pelvic examination*. Age, history of sexual intercourse, and having had a *Pap smear* correlated with a better *Pap smear* definition rating.

The study presented by Chapman, Abraham, Jenkins, and Fallowfield (2003) assessed lay understanding of terms used by doctors during cancer consultations. The questionnaire included scenarios containing potentially ambiguous diagnostic/prognostic terms, multiple-choice, comprehension questions and figures on which to locate body organs that could be affected by cancer. Respondents also rated how confident they were about their answers. About half the sample understood euphemisms for the *metastatic spread of cancer* e.g. *seedlings* and *spots in the liver* (44 and 55 % respectively). Sixty-three per cent were aware that the term *metastasis* meant that the cancer was spreading but only 52 % understood that the phrase *the tumour is progressing* was not good news. Knowledge of organ location varied. For example, 94 % correctly identified the lungs but only 46 % located the liver. The findings suggest that a substantial proportion of the lay public do not understand phrases often used in cancer consultations and that knowledge of basic anatomy cannot be assumed. Awareness of the unfamiliarity of the lay population with cancer-related terms could prompt further explanation in cancer-related consultations.

Ogden et al. (2003) studied patients' views about the relative impact and function of lay and medical diagnoses for stomach and throat problems. The results showed consistent differences between the lay and medical labels for both stomach and throat problems in terms of their impact upon the patient and their function for the doctor. In particular, the medical labels were rated as beneficial for the patient in terms of validating the sick role and improving their confidence in the doctor. In contrast, the lay labels resulted in a greater sense of ownership of the problem which could be associated with unwanted responsibility and blame. In addition, the medical labels were seen to provide the doctor with a greater sense of professionalism, as giving them a clearer role in the consultation and to imply less blame on the part of the patient. *Stomach upset* was also seen as a more pragmatic label than *gastroenteritis*.

Although much current prescriptive literature in general practice advocates the use of lay language in the consultation as a means to promote better doctor-patient partnerships, the issue of diagnosis is more complex than this. Patients attribute greater benefits to the use of medical labels for themselves and state that such medical labels are of greater benefit to the doctor.

The study proposed by Aufseesser, Lacroix, Binyet, and Assal (1995) evaluates how diabetic patients understand the meaning of 8 medical terms related to retinopathy. The results show only one third of correct answers. Results illustrate a big diversity in understanding according to the terms which, nevertheless, were currently used by the doctors during their ophthalmological consultation. The same team (Binyet, Aufseesser, Lacroix, & Assal, 1994) asked sixty diabetic patients to give their definitions of 12 terms concerning the *diabetic foot*. On average, only half of these terms were understood by these patients. The level of correct replies is associated neither with the patient's socio-cultural level nor with other variables such as the socio-demographic determinants (sex, age, life-style).

According to Peckham (1994), many patients apparently believe there is a difference between a *fracture* and a *break*. In a survey of 100 patients, 81 thought there was a difference. Of these, 71 thought a fracture was better than a break, and 65 believed that bone was undisplaced in a fracture and displaced in a break.

To determine mothers' level of comprehension of terminology used by health care providers when discussing the care of a newborn baby, DiFlorio (1991) asked 60 patients on a postpartum unit in a general hospital who gave birth to healthy newborns to define 56 terms related to the care of newborn babies that were commonly used by health care providers. Analyses of variance results revealed significant differences among mothers for age and level of education. Mothers who were 30 years or older and who had more than a high school education demonstrated the more overall knowledge of terms than younger mothers with less than a high school education.

Hadlow and Pitts (1991) realized a survey by multiple choice questionnaire of 40 doctors, 60 health support staff and 120 patients investigated the understanding of common medical and psychological terms. Significant differences in levels of understanding were found between these groups, and level of medical education predicted the level of correct understanding. The widest gap in doctor-patient understanding was shown for common psychological terms.

Spees (1991) observed the knowledge of 50 common medical terms by 25 family members. Only nine terms were correctly understood by all respondents. Older persons with higher education and moderate length of illness had higher scores.

Henley and Hill (1990) presents a cross-sectional survey conducted among 60 families with a child with cystic fibrosis to assess their medical knowledge of the illness. A 63-item, multiple-choice test with acceptable psychometric properties was administered to 60 mothers, 54 fathers, 29 siblings (aged 10 to 23 years), and 18 patients (aged 9 to 22 years). Parents and patients correctly answered approximately three quarters and siblings two thirds of all items. Family members were most knowledgeable about general cystic fibrosis facts, physiotherapy, gastrointestinal symptomatology and treatment, and anatomy. They were less well-informed about respiratory symptomatology and treatment and nutrition. Parental knowledge of genetics and reproductive risks was mediocre, and that of patients and siblings was poor. Knowledge of terminology was uniformly low. Social class was a significant predictor of parental knowledge. If left uncorrected the misconceptions, gaps, and errors in family members' knowledge of cystic fibrosis identified in this study could result in inadvertent noncompliance in treatment of the patient.

Spiro and Heidrich (1983) led a study with 1606 adult patients of a community family practice program which were questioned about their understanding of the terms *hypertension*, *virus*, *strep throat*, *herpes*, *tumor*, *Pap smear*, and *uterus*. Significant misconceptions were common among patients of all ages and educational backgrounds, although a positive association of education and knowledge was demonstrated. In using these and similar terms, clinicians must be cautious to ensure that the patient is receiving the intended message.

3.1.6 Which words to use?

Knapp, Raynor, and Berry (2004) study aims to determine whether the use of verbal descriptors suggested by the European Union (EU) such as *common* (1–10 % frequency) and *rare* (0.01–0.1 %) effectively conveys the level of risk of side effects to people taking a medicine. The verbal descriptors were associated with more negative perceptions of the medicine than their equivalent numerical descriptors. Patients want and need understandable information about medicines and their risks and benefits. This is essential if they are to become partners in medicine taking. The use of verbal descriptors to improve the level of information about side effect risk leads to overestimation of the level of harm and may lead patients to make inappropriate decisions about whether or not they take the medicine.

Zeng, Kogan, Ash, and Greenes (2001) observe that lack of familiarity with medical vocabulary is a major problem for patients in accessing the available information. As a first step to providing better vocabulary support for patients, authors collected and analysed patient and clinician terms to confirm and quantitatively assess their differences. The results showed that patient terminology does differ from clinician terminology in many respects including misspelling rate, mapping rate and semantic type distribution, and patient terms lead to poorer results in information retrieval. Thus, Avenarius (1994) considers that it is necessary for doctors to use different codes depending on whether they communicate with each other, with their patients, or outsiders. Such codes are every day language, technical language, scientific language and language to the knowledge of different groups of non-specialists.

According to Rodning (1992), since the the medical slang can lead to incompleteness, incorrectness, misinterpretation, the ambiguity, uncertainty, anxiety, and animosity among the individuals (patients and physicians) involved in dialogue can arise.

Pediatric care becomes more complex and technical, and there is a continuing tendency among health care professionals from all fields to use labels, jargon, and abbreviations when talking about and talking to young patients. McCue (1991) offers possible explanations for the use of labels and jargon, and provides a summary of the dangers inherent in this sort of communication. Simple suggestions for alternative forms of communication are provided.

Lindsley (1991) gives examples of translating technical jargon into plain English application words, acronyms, letter codes, and simple tests were necessary as we developed Precision Teaching. The author shows that accurate plain English translations do not come easily. They cannot be made from scratch at the desk.

Linguistic differences From the linguistic point of view, notice that the medical slang has the specificity to use abbreviations, acronyms and Latin terms (Surjan & Heja, 2003). The Latin abbreviations especially does not consistently promote patient safety (Dunn & Wolfe, 2001). The use of plain English is suggested as the prescribing practice most consistent with professional values. While the specificity of patient language is to use nouns instead of verbs and adjectives instead of nouns (Richardson, 1996).

Legal problems Two court cases in British Columbia in the nineties reveal the importance of using plain, simple language to communicate with patients (Gordon, 1996). This is particularly important because almost half of Canadians have low literacy levels. The CMA, which promotes the use of plain language in professional practice, is participating in the Canadian Public Health Association's National Literacy and Health Program. Resources are available to help physicians better serve patients.

3.1.7 Analysis of few concepts

Fever H. J. Thompson (2005) analysed the current state of the science literature in order to develop an accurate conception of *fever*. Literature for this concept analysis was obtained by computerised searches of PubMed, CINAHL and BIOSYS for the years 1980–2004. Additional sources were obtained after reviewing the bibliographies of the literature identified by the initial search. Fever has characteristically been recognised as a cardinal sign of illness and has traditionally had negative connotations for patient well-being. Substantive advances over the past 20 years in immunology and neurophysiology have expanded understanding of the process of fever. This new knowledge has shifted the perception of fever as part of the acute-phase response to one of an adaptive nature. This knowledge has yet to be fully translated into changes in the fever management practices of nurses. Consistent usage of terminology in relation to fever should lead to improved and evidence-based care for patients, and to fever management practices consistent with current research. It is important to use clear language about fever and hyperthermia in discussions and documentation between nurses and among disciplines.

Rheumatism The word *rheumatism*, introduced in ancient times, is still used directly or indirectly, in parallel with the terms of the modern nosography. The reasons for this persistence can be sought in the history of the concept, which can be approached via quotations from texts written either by authors who describe popular beliefs or their own sufferings; or by physicians known to have played a prominent role in the individualisation of rheumatology. The word *rheumatism* was first used mainly to designate a painful fluxion of the tissues located between the skin and the internal organs. It gradually lost ground to more descriptive terms suggestive of joints. Thus, the concept of rheumatism still bears the hallmark of its "popular" roots and is on a level parallel to but distinct from that of modern nosography. Awareness of its origins may improve communication between physicians and patients and also raises questions about the foundations of the concept of *rheumatic disease*.

3.1.8 Matching patient vocabularies and problem lists with standard terminologies

Once the need of matching these two vocabularies is defined, how this can be done? Following references, resume already proposed approaches. Most of them use the UMLS Metathesaurus and tools. Few others try specific international medical terminologies (ICD-9) or local ones.

Matchings with UMLS The linguistic analysis of e-mails to a cancer information service allowed to detect 504 unique terms (C. A. Smith, Stavri, & Chapman, 2002a). These terms have been matched to the UMLS Metathesaurus and following results have been found:

- 185 (36 %) terms present exact match

- 179 (35 %) terms present partial match
- 119 (24 %) are known synonyms
- 2 (< 1 %) are lexical variants
- 19 (4 %) have no matchings

As a result, up to 96 % of terms detected in patient e-mails could be matched to the UMLS Metathesaurus terms.

In the context of information retrieval (IR), Plovnick and Zeng (2004) designed a system for the reformulation of patient queries through their matching to the UMLS Metathesaurus. Such reformulation showed that 42 % of queries gave better results, 19 % worse results, and 39 % showed no difference. Otherwise, this study allowed to identify that ambiguous lay terms, expansion of acronyms, arcane professional terms have an impact on performances of the reformulation system.

The objective of the work described in Brenna and Aronson (2003) is to search electronic knowledge resources and bring health information resources into the hand of patients. The way to realize this is to detect relevant concepts from UMLS within the free text of lay people's e-mail. The work shows that the UMLS nursing vocabularies provide an excellent starting point for this exercise because their domain encompasses patient's responses to health challenges. The best performance was obtained when the nursing vocabularies were complemented with selected clinical terminologies.

Out of 1262 unique terms found in medical records, 999 terms (79 %) have matches in UMLS (H. Goldberg et al., 1998):

- 986 of them map to the UMLS concept of the corresponding lexical match
- 952 of them have semantic types that comply with the operational definition of clinical problems.

A more detailed evaluation has been done in further work on 2810 disease-related labels with UMLS-based semantic parser (H. S. Goldberg, Hsu, Law, & Safran, 1998):

- parser correctly recognised and validated 1398 of terms (49.8 %)
- correctly excluded 1220 of 1312 non-disease-related labels (93 %)

Among 1181 failures of the parser:

- 812 match failures (68.8 %) were caused by terms either absent from UMLS or modifiers not accepted by the parser
- 369 match failures (31.2 %) were caused by labels having patterns not recognised by the parser.

By enriching the UMLS lexicon with terms commonly found in provider-entered labels, it appears that performance of the parser can be significantly enhanced over a few subsequent iterations.

One of objectives when matching problem list terms with the UMLS terms is to incorporate information about relationships between UMLS concepts into the problem list. Such experience is described in Hales, Schoeffler, and Kessler (1998). 67 % (1627/2436) of terms could be matched with my normalised string matches to the UMLS KS. Of these matches, 91 % participated in at least one UMLS-identified parent relationship but only 28 of the matched

concepts participated in parent relationships that already matched to a patient problem list terms. Not surprisingly, patient list problem's terms is less rich in terms than the UMLS, which comprises about hundred of terminologies.

As healthcare consumers often have difficulty expressing and understanding medical concepts, Tse and Soergel (2003) propose to identify and characterise medical expressions or terms used by consumers and health mediators. In particular, these terms were characterised according to the degree to which they mapped to professional medical vocabularies. Lay participants identified approximately 100,000 term tokens from online discussion forum postings and print media articles. Of the over 81,000 extracted term tokens reviewed, more than 75 % were mapped as synonyms or quasi-synonyms to the Unified Medical Language System (UMLS) Metathesaurus. While 80 % conceptual overlap was found between closely mapped lay (consumer and mediator) and technical (professional) medical terms, about half of these overlapping concepts contained lay forms different from technical forms. This study raises questions about the nature of consumer health vocabularies that authors believe have theoretical and practical implications for bridging the medical vocabulary gap between consumers and professionals.

Matchings with other terminologies An other IR application, developed on the basis of the French medical gateway CISMeF, the reformulation of queries is done through already registered synonyms of MeSH terms (Soualmia et al., 2003).

Like in Brenna and Aronson (2003), Travers and Haas (2003) propose to match patient and expert terminologies through the the construction of concept-oriented nursing terminologies.

Bui et al. (2004) propose to automatically generate medical problem list on the basis of ICD-9: given a set of ICD-9 codes associated with a patient records, the system maps the codes (and related data) to an anatomy-centric hierarchy.

In the work reported by Yarnall, Michener, Broadhead, Hammond, and Tse (1995), a computer system is designed to translate patient diagnoses noted by a physician into appropriate ICD-9-CM codes and maintain a patient-specific up-to-date problem list. Since an additional locally built dictionary is used, following results are obtained:

- Visits in which all diagnoses matched increased from 58 % to 76 % with use of the computer system
- Visits in which no computer diagnoses matched the chart decreased from 22 % to 8 %
- Errors of omission declined from 38 % to 18 %
- Errors of commission decreased from 19 % to 11 %
- Overall accuracy increased from 62 % to 82 %

Scherpbier, Abrams, Roth, and Hail (1994) designed a system which allows a quick and easy method to enter and maintain a patient's problem list. Physicians can use their own terminology. ICD-9 codes are included where possible, but free text is allowed.

J. R. Campbell and Payne (1994) present a comparison of matchings between patient problem list terms and terms from four international standard terminologies: UMLS, SNOMED, Read and ICD-9-CM. Matching with UMLS and SNOMED performed substantially better in capturing the clinical content of the problem lists than Read or ICD-9-CM.

In order to automate processes based on problem lists, Fabry, Baud, Ruch, Despont-Gros, and Lovis (2005) use controlled vocabularies. Terms extracted from physicians' notes have

been matched to such terminologies. 88,6 % of 1546 terms could be related to a relevant problem statement.

Ruland and Bakken (2001) use of the LOINC semantic structure as a terminology model to create fully specified names of patient preference-related concepts to be represented non-ambiguously and in a manner that renders them suitable for computer rather than human processing.

Henry and Holzemer (1994) examined the ability of SNOMED International to represent patients' perceptions of health-related problems. The majority of concepts used by patients to describe health-related problems could be matched with existing SNOMED terms. The addition of the social context module as an adjunct to existing terminologies of medical diagnoses, NANDA diagnoses, and signs/symptoms provided additional matching terms. Patient goals did not match existing SNOMED terms. The findings of this study suggest that SNOMED International has the potential to adequately represent patients' perceptions of health-related problems for the computer-based patient record.

What the matching could mean Research in general practice emphasises the importance of matched models, beliefs and vocabulary in the consultation. The study presented in N. Williams and Ogden (2004) aimed to explore the impact of matched and unmatched vocabulary on patient satisfaction with consultations for problems relating to sexual or bodily function or anatomy. Matched consultations required the doctor to use the same vocabulary as the patient. Unmatched consultations required the doctor to use medical vocabulary. Completed questionnaires were received from 60 patients. Patients in the matched consultation group had significantly higher total satisfaction scores and higher ratings of rapport, communication comfort, distress relief and compliance intent than those in the unmatched group. The results indicate that a doctor's choice of vocabulary affects patient satisfaction immediately after a general practice consultation and that using the same vocabulary as the patient can improve patient outcomes.

3.1.9 Conclusion

Communication between patients and medical staff has been researched for few decades and showed significant differences between these two vocabularies. The importance of their matching has been discussed.

The improvement of the communication between patients and their physicians can be supported by automated tools. But these tools would need suitable terminological resources. Cost and efforts needed for their building should not be underestimated. Their content can be inspired from Zielstorff (2003): *Structured vocabularies comprised of lay terms, with definitions, variant spelling and regional dialects, along with mappings to equivalent or related professional terms ...* Moreover, a gradual and progressive shift from exclusive to shared knowledge and responsibility can be achieved by patient/physician collaboration (Sadan, 2002).

3.1.10 Technical Term Translation

Term translation approaches recur – with different focus – in several different research contexts including Cross Language Information Retrieval (Levow, Oard, & Resnik, 2005), Corpora Alignment (Resnik, 1999), Word Sense Disambiguation (Markó, Schulz, & Hahn, 2005) or Automatic Lexicon Acquisition (Markó, Schulz, Medelyan, & Hahn, 2005). In all these

contexts the translation of unknown, so called out-of-vocabulary terms are a major challenge. Usually existing bilingual word lists are used as seed lexicons, or parallel, related or even unrelated corpora are exploited.

Baud, Lovis, Rassinoux, Michel, and Scherrer (1998) applied a multilevel method to automatically create a bilingual English-French dictionary of nearly 10,000 word pairs exploiting co-occurrences of words in the ICD-10 classification. Similar to our subword based approach they transform compounds or derivational words to underlying concepts using a dictionary with 8,000 entries. The resulting word pairs proved correct in 98 % of the cases.

Schulz, Markó, Sbrissia, Nohama, and Hahn (2004) introduced a method of directly translating terms from Portuguese to Spanish using simple string transformation rules. These translations are then validated in the local context of language-specific corpora resulting in a list of biomedical cognate pairs.

Claveau and Zweigenbaum (2005) propose an algorithm that infers transducers from examples of bilingual word pairs. They achieve up to 85 % of correct translations for translations between French and English. This approach, again, counts for biomedical simple terms (composed of one word) only and may be less effective in languages in which word compounding is used extensively (such as German, Dutch or Swedish).

In a previous work Y. C. Chiao and Zweigenbaum (2002) identified translational equivalents of out-of-dictionary words from French to English in the medical domain relying on non-parallel, comparable corpora and an initial bilingual medical lexicon. They achieved about 60 % of correct translations in the top ten candidates.

For more recent work, see Markó et al. (forthcoming) and Daumke, Schulz, and Markó (forthcoming).

3.2 Terminological Resources

The National Board of Health and Welfare in Sweden works with classification and terminology issues. Its terminology bank `http://app.socialstyrelsen.se/termbank` covers about 600 search terms recommended for use in communication within health care services and in communication with patients. This term bank does not include however professional medical terms dealing with diseases or anatomy.

In the same site as above, the report entitled "Begrepp och termer inom vård och omsorg – Rapport från InfoVU-projektets kunskapsnätverk för begrepp och termer", "Concepts and terms within healthcare", covers 150 general terms (including definitions and comments).[1]

The Swedish Council on Technology Assessment in Health Care (SBU) `http://www.sbu.se/` works with the promotion of health technology assessment. SBU's aim is to identify interventions that offer the greatest benefits for patients while utilising resources in the most efficient way. Its terminological bank `http://www.sbu.se/ordlista/list.asp` covers about 200 terms used in e.g. clinical drug trials and non-interventional trials.

The "Methods and principles in terminology work" (Spri, 1999) report provides an introduction to terminology work for health care professionals. It is available from the Swedish Centre for Terminology which provides terminological services and support to authorities, organisations, enterprises in Sweden who pursue terminological work of their own within various subject fields, and also to individuals.

[1] `http://www.socialstyrelsen.se/NR/rdonlyres/465DC500-CB71-40A6-8FDA-E38763DB6308/4539/200513121.pdf`

The need for the identification of a "common language" between different professions in health care has been emphasised in an article by a group of health care professionals (Broberg et al., 2006). Their article was a response to a critical view in the hospital doctor's web site `http://www.sjukhuslakaren.org` regarding the efforts to replace the doctor's "reliable vocabulary" with "unfamiliar search words" (Zur-Mühlen, 2005/05).

The book *Medicinens språk* (Nyman, 1996) provides a general introduction to the medical language used in Sweden. The book is aimed at those who either are interested in medical language or work within health care services. It has a linguistic profile with a normative touch. Explicit formulated recommendations guide the reader into the realm of orthography, pronunciation, derivation of medical terms and some basics of Latin grammar necessary for interpretation of multiword medical phrases. The contrastive approach reveals many structural differences between Latin or Greek loan terms and their Swedish equivalents. The linguistic data in the book not only provides a thorough introduction to medical language but also can be useful for some applications within medical informatics, e.g. for writing algorithms to handle orthographic variation or to segment compounds.

In the Semantic Mining network the Department of Biomedical Engineering at Linköping university, has collected a sample of English-Swedish terminology from a number of terminology systems. The collection is called TermColl, and contains a total of 39,500 parallel rubrics from ICD-10, ICF, MeSH, NCSP, KSH97-P. A rubric is a short informative term accompanying each code in the terminology system. The TermColl material has also been extended by using word alignment techniques to extract more fine-grained terminology.

There is a US initiative on formulating and compiling a so-called "Consumer Health Vocabulary" (CHV) as a freely available electronic resource (see `http://www.consumerhealthvocab.org/`). In connection with this initiative, and also more generally, there has been some research to elucidate which linguistic parameters influence the accessibility of medical texts to laymen (Zeng & Tse, 2005; Zeng, Kim, Crowell, & Tse, 2005; Zeng, Tse, et al., 2005, see further the bibliography on the CHV website); the results of this research indicate that the main factor here is vocabulary complexity, although syntax also plays a role (Ownsby, 2005). The CHV initiative is for English. At the moment, a similar initiative does not exist for Swedish specifically in the health domain, although t here is a small number of studies related to medical terminology which deal with "readability issues", e.g. Borg (2005) and Grehn (2004). If we widen our scope somewhat, however, there is a Swedish government agency, *Centrum för lättläst* ("the Centre for Easy-to-Read"; see `http://www.lattlast.se/`), with a remit to adapt written texts (fiction, news, official publications), mainly for the estimated 25 % of the population who have reading difficulties of various sorts, but more recently they have also started offering gisting of written materials for the information-overloaded segment of the population, normally highly educated and mostly with no reading difficulties. Relevant for this literature survey, the centre publishes guidelines for writing (or adapting) texts for accessibility and is also actively involved in research on reading efficiency and reading difficulties.

In the AAC (augmentative and alternative communication) community, there is ongoing European work on a common (general) conceptual coding system for translating among AAC symbol languages (Bliss, Picto, etc.) and between these and various written languages (among which are Swedish and English). The target groups of this work are by and large a sub set of those of WP27. See `http://www.dart-gbg.org/`; `http://www.wwaac.org/`; `http://www.conceptcoding.org/`.

4 Computational Methodologies

This chapter briefly looks at existing computer systems or techniques that have been used to deal in some ways with the text found in medical records, either as patient records or in text aimed at patients (i.e. a non-specialist audience). Some common themes in the existing work are:

- extraction of terminology from corpora
- comparison of terms from different sources
- the analysis of patient records with reference either to controlled terminologies or models of reports' narrative structure
- the automated or computationally assisted construction of ontologies, and
- the use of description logics to model terminology use and change.

Several of the papers discussed in this chapter do not address the questions of doctor/patient language directly, but address similar tasks within the domain of biomedicine. Such papers have been included if they address a similar task required by WP27, and use similar resources (for example, SNOMED or UMLS).

4.1 Existing Resources

Existing work uses the the following resources:

- UMLS
- GALEN
- SNOMED
- MeSH

These are existing resources for managing biomedical terminology. Work on NLP in this area generally uses one or more of these.

4.1.1 UMLS

UMLS, the Unified Medical Language System[1] (Bodenreider, 2004) contains three parts:

1. Metathesaurus. Approximately 1 million concepts and 5 million concept names. The data is obtained from over 100 controlled vocabularies and is intended to systematise the relationships between vocabularies.
2. Semantic Network. Arcs between concepts encoding the relationships between the concepts in the Metathesaurus.

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

3. SPECIALIST lexicon. Terms from both common English vocabulary and biomedicine. An example[2] for anaesthetic is:

```
{base=anaesthetic
spelling_variant=anesthetic
entry=E0008769
cat=noun
variants=reg}
{base=anaesthetic
spelling_variant=anesthetic
entry=E0008770
cat=adj
variants=inv
position=attrib(3)}
```

4.1.2 GALEN

The GALEN system[3] (Rector, Gangemi, Galeazzi, Glowinski, & Rossi-Mori, 1994) is intended to provide reusable terminology resources for clinical systems. Their are two main components:

1. The GALEN Representation And Implementation Language (GRAIL)(Rector et al., 1997). A description logic for encoding medical knowledge.

2. The Common Reference Model (CRM). An model of medical knowledge written in GRAIL.

A recent progress report (Rector & Rogers, 2006) describes the successes of GALEN in representing clinical information and some of the relationships between terms.

4.1.3 SNOMED

The Systemized Nomenclature Of MEDicine, SNOMED[4] is a terminology data set in the form of SNOMED Clinical Terms (SNOMED CT).

4.1.4 MeSH

Medical Subject Headings (MeSH)[5] is a database based upon a controlled vocabulary that is used to classify or index books and journal articles.

4.2 Divergence of Patient and Practitioner Language

Soergel, Tse, and Slaughter (2004) consider how UMLS can provide a translation (or "interpretive") layer between doctor and patient language. They also discuss several of the ways that a patient typically misunderstands, or misrepresents, the clinical data. C. A. Smith,

[2] from the UMLS entry in Wikipedia
[3] http://www.opengalen.org
[4] http://www.snomed.org
[5] http://www.nlm.nih.gov/mesh

Stavri, and Chapman (2002b) and Hsieh, Hardardottir, and Brennan (2004) have both investigated the language that patients use in email communication with their nurses. Hsieh et al. (2004) investigated how many key terms in patients' emails could be extracted just by using the standard UMLS vocabulary; they found that their MetaMap tool, most of the medical information in their mails could be extracted using UMLS. C. A. Smith et al. (2002b) have also investigated the language used by patients, and found that a large majority ($> 90\,\%$) of terms used by patients were exact matches to terms used in the UMLS Metathesaurus. They conclude from this that the notion of a healthcare consumer with his or her own particular healthcare language may be ill-founded (although this is not a widely held view).

To investigate the difference in language used by different writers, Bodenreider and Pakhomov (2003) have explored the behaviour of adjectival modifiers across the two written genres using texts from Medline[6] and the Mayo clinic[7]. They found that a much greater range of adjectives was used for the wider audience, and suggest that sensitivity to the broader range of adjectival modification is necessary in systems dealing with medical texts.

An application to multilingual medical documents is that of Widdows et al. (2003), who have attempted to use parallel corpora for disambiguation. It is possible that the techniques they have developed could be applied to documents for different audiences using the contextual awareness techniques discussed by Spasic and Ananiadou (2005) and D. A. Campbell and Johnson (1999). Finally, a similar technique has been tried by Yeh, Wu, Chen, and Yu (2004) using alignment of multilingual *ontologies*, rather than languages.

4.3 Computational Methods in Terminology Management

An important task will be to recognise from existing documents where different terminology is used to refer to common concepts. Although some examples (Spasic and Ananiadou, for example) are targeted more at recognising gene names, the techniques are still relevant to WP27, as a method of recognising common terms used across patient and practitioner documents. In particular, as these systems are used within the biomedical domain, the use of resources such as UMLS raises issues that are relevant to the analysis of medical documents.

The use of controlled medical terminologies has been investigated by Liu and Friedman (2000) and Oliver, Shahar, Shortliffe, and Musen (1999), who have been looking at the automatic management of controlled vocabularies. Liu and Friedman have used description logics to represent the diverging use of terminology over time according to a set of core concepts. Oliver et al. (1999) present the MedLee system, which identifies word phrases, represents them in XML, and then matches the XML trees. Kornai and Stone (2004) have given a useful overview of the general problems in translating between medical terminologies.

Spasic and Ananiadou (2005) have investigated how a measure of the similarity of the context in which terms appear can be used to recognise similarity. Their work is particularly aimed at the recognition of terms that do not currently exist in the vocabulary (of UMLS in this case). They use a two-stage process; the first stage requires POS tagging and shallow parsing to recognise the similarity of terms in different documents, then reference to UMLS is used at the second stage to hypothesise term similarity. The validity of this method for medical records is illustrated by D. A. Campbell and Johnson (1999), who earlier applied very similar methods to one year's discharge summaries from a New York hospital. Campbell

[6] `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed`
[7] `http://www.mayoclinic.com/`

and Johnson have been particularly concerned with detecting previously unknown words from document collections; this has particular relevance to the cases highlighted by Bodenreider and Pakhomov (2003), who recognised that patients are likely to use a wider linguistic range than practitioners (and therefore, what is likely to appear in controlled vocabularies).

Computational analysis of morphological change, and how that affects what can be obtained from medical documents is discussed by Lovis, Baud, Rassinoux, Michel, and Scherrer (1998) and Zweigenbaum and Grabar (2000). In particular, Zweigenbaum and Grabar show how morphological knowledge can improve information extraction of systems based upon SNOMED. A system that both uses morphological disambiguation and contextual knowledge to disambiguate is described by Mukherjea et al. (2004), who attempt similar work based upon UMLS rather than SNOMED.

4.4 Construction and Maintenance of Ontologies

Throughout the literature, ontologies are used to guide techniques such as terminology acquisition and document understanding. However, the construction of such ontologies is expensive and time consuming. A valuable use for NLP is applied to the automated (or assistance with automation of) ontology construction and maintenance. Good et al. (2006) have experimented in automated text mining using the TermExtractor tool with volunteers to improve term recognition from multiple document sources. Their results suggest that human aided similarity recognition should provide a basis for ontology building from text.

Gangemi, Pisanelli, and Steve (1998) have recognised that a major problem is the alignment and integration of separate ontologies (for WP27, the issue will be the comparison of ontologies reflecting the knowledge of a lay person, against that of a specialist). The ONIONS system attempts illustrates how the terminologies in UMLS and SNOMED can be integrated by by using a general encoding in first order logic. Their aims have been is to ensure that term classification and definitions are now available in a common, expressive formal language and that for upgrading of terminology systems, term classification and definitions are translated so that they can be included in an ontology library which has a subset constituted of motivated generic ontologies.

4.5 Knowledge Representation

While most of the applications discussed in the literature use the standard existing ontologies to represent medical knowledge, Hahn, Romacker, and Schulz (1999) have used a model formulated in description logics using a KL-ONE based formalism. This model allows the knowledge representation to be more intimately connected to their task of understanding medical language. Their system MedSyndikate (the design issues for the system are discussed in Hahn, Romacker, & Schulz, 2000) is evaluated on German pathology texts, and in ongoing research projects has been applied to concept formation and text mining (Hahn, Romacker, & Schulz, 2002) and Natural Language Understanding (Romacker, Schnattinger, Hahn, Schulz, & Klar, 2004).

4.6 Medical WordNet

Medical Wordnet (B. Smith & Fellbaum, 2004; Fellbaum, Hahn, & Smith, forthcoming) is an attempt to develop a dictionary of medical vocabulary along the same lines as WordNet (G. A. Miller, 1995). It is a stated aim of Medical WordNet to address problems with doctor/patient communication issues; Fellbaum et al. (forthcoming) point out that although WordNet contains many medical terms, it was not developed by practitioners and so does not contain the degree of structure that may be required for medical applications.

Medical WordNet consists of three parts: medically relevant word forms, structured in the same way as WordNet, Medical FactNet, which constitutes a set of medical facts, and Medical BeliefNet, which is a collection of propositions reflecting laypersons' medical beliefs. The development of Medical WordNet is expected to be a lengthy project, requiring major input from both medical specialists (validating the medical facts embodied in Medical FactNet) and non-specialists (affirming the propositions in Medical BeliefNet). The two subcorpora, Medical FactNet and Medical BeliefNet are derived from online healthcare information from factsheets on Airborne allergens and from the UK Netdoctor's diseases encyclopedia.

Medical WordNet currently exists only as pilot projects, and there does not yet exist a large-scale implementation.

4.7 Analysis of records

There are several examples of attempts to extract information from a set of medical records using NLP techniques; this is held to be a useful step in forming representational structures of those records.

Oliver and Altman (1994) have identified the interpretation of given patient records as a possible first step in comparing differing terminologies for describing common conditions. They use shallow parsing techniques and matching against SNOMED to identify which terms in patient records are controlled terms from the SNOMED vocabulary. The ideas are extended in Oliver et al. (1999), in which a description logic (similar, though not identical to, that of GRAIL) is used to encode a model of how controlled terminologies change over time.

SNOMED has also been used by Elkin et al. (2005) to assist with automated understanding of electronic health records; in particular, they have shown that using SNOMED improves the accuracy with which negation can be recognised.

There are several instances where a model of the patient narrative has been used as the basis for understanding a patient record document. The earliest example is given by Sager, Lyman, Bucknall, Nhan, and Tick (1994), who used a custom-built model of the clinical narrative to build an understanding of the patient record. Similar techniques have since been developed using pre-existing, rather than custom-built models of the narrative. Wu, Yu, and Jang (2005) show that the symptoms of depression can be identified from patient records by assuming an underlying narrative based upon the Hamilton Depression Rating Scale (Hamilton, 1960). Rector and Rogers (2006) describe similar results, using the medical models in GALEN, rather than the specific case of Wu et al. (2005).

A similar tack is taken by Niu, Hirst, McArthur, and Rodriguez-Gianolli (2003) in their work on Medical QA systems; text understanding (and thereby, question answering) is improved by identifying the roles (for example, patient, treatment, outcome) of different named entities in the documents.

Some Medical knowledge acquisition systems have been built directly into parsers. Steimann (1998) and Taira and Soderland (1999) have shown how the relationships between medical concepts can be built directly into a dependency grammar. As an implementation issue, note that Ruch, Baud, and Geissbühler (2002) have shown several methods of improving understanding of patient records by using spelling correction techniques.

4.8 Text Simplification and Dialogue

Considering the computational issues involved in automatically adapting health record contents and other professional medical information to the needs of laymen – who are increasingly interested to understand more about their personal health status, and about past or planned health care interventions including the purpose and safety of their medications – we note that there are several strands of existing work in computational linguistics and in other fields which could be brought to bear on this problem. Thus, there is work of interest going on in (at least) the areas of text simplification, dialogue systems and natural language generation.

4.8.1 Text simplification

There has been some work on text simplification for various purposes, e.g. accessibility for aphasics (Carroll, Minnen, Canning, Devlin, & Tait, 1998; Carroll et al., 1999; Canning, Tait, Archibald, & Crawley, 2000)[8] or deaf people (Inui, Fujita, Takahashi, Iida, & Iwakura, 2003) and semi-automatic production of controlled-language texts (Chandrasekar, Dorian, & Srinivas, 1997; Chandrasekar & Srinivas, 1997). Other potential target groups for text simplification services would be second/foreign language learners, users of small-screen, limited-bandwidth devices (Siddhartan, 2002) and, of course, laymen (patients) wishing to access text aimed at professionals (physicians), e.g. popular science texts (Senda, Sinohara, & Okumura, 2004, for Japanese).

In order to automatically produce a simplified version of some text, a system must be able to: (1) (possibly) select the most relevant portions of the text, an objective shared with systems for *automatic summarization* (Mani, 2001) and *(topical) text segmentation;* (2) paraphrase the portions of the text the content of which is to be (hopefully faithfully) conveyed in the simplified version of the text, making text simplification in some ways similar to *example-based machine translation* (EBMT[9]) and of course *automatic paraphrasing* (including ontology paraphrasing in natural language, e.g., Hewlett, Kalyanpur, Kovlovski, & Halaschek-Wiener, 2005).

There is also some work on generation of documents for people with poor literacy skills (S. H. Williams, 2004). The system used discourse constraints and optimisation rules for selecting solutions with short sentences and short, common discourse connectives.

4.8.2 Dialogue systems

Dialogue systems are systems allowing users to formulate their information needs in natural language and also engage in a dialogue with the system. The use of dialogue means for instance that it is possible to formulate information needs in a fragmented fashion. The dialogue system can collect further information from the user and ask for clarifications before

[8] see also `http://osiris.sunderland.ac.uk/~pset/`

[9] see, e.g., `http://iai.uni-sb.de/~carl/ebmt-workshop/papers.html`

searching the information base. Once information is presented the user can ask follow-up questions; ask for more information, ask for clarification, precision etc.

Today dialogue system technology is used more and more in interactive guides for various domains and mainly with an informative-seeking purpose (simple question-answering). In 1966 ELIZA (Weizenbaum, 1966) was developed and since then people have tried to create conversational systems that emulate interactions between health providers and patients. More recent systems like interactive guides have been built attempting to provide low-cost and widely accessible health care in restricted treatment areas, and many of these systems have been proven effective in large-scale clinical trials (Bickmore & Giorgino, 2004). Interactive guides are front-ends to information systems where users interact with an animated figure. In Sweden there are very few interactive guides for medical information. The company Humany has developed one for *Jämtlands Läns Landsting* (*Jämtland County Health Care*) where citizens can ask for information about health care in the their region. The web assistant is intended for citizens and the information is primarily focused on administrative and practical procedures in the county, such as "Where can I find a cardiologist?" or "What is the budget for health care in the county?"

The web assistant from Humany can also consult the external web site from *Sjukvårdsrådsgivningen* for specific medical questions, and has then the same problems as their question-answering system, mainly in the processing of FAQs. Humany's web assistant is currently undergoing further development and customisation to the medical domain in order to add automatic dialogue capabilites to *Sjukvårdsrådgivningen* (see section on *Health Care Counselling / Sjukvårdsrådgivningen*).

4.9 Automated acquisition of lexicon

4.9.1 Acquisition of monolingual lexicon

Acquisition of monolingual lexical resources, especially those which contain morphologically related words, is used to help different NLP tasks (ie. term variants detection, POS tagging, text typology, semantic tagging). The main problem consists in finding the semantic context in which automated acquisition tools can provide with morphologically and lexically relevant relations. Different approaches are proposed.

Using electronic dictionaries Dictionaries can provide rather constrained semantic context for the acquisition of lexicon. Results depend on the richness and exhaustivity of used dictionaries.

Krovetz (1993) exploits the electronic version of *Longman Dictionary of Contemporary English* in two ways:

- the dictionary structure is used for the detection of words related formally (through their form) and semantically: if the entry lexeme is formally close to one of words from its definition, both words are considered as well as semantically close. For instance, the definition of *cylindrical* contains the word *cylinder*;
- the Porter stemming algorithm (Porter, 1980) is iteratively applied to entries, but before each iteration the system checks is the stemmed word corresponds or not to an existing dictionary entries. If so, the stemming process is interrupted. With such processing

interpretation of obtained stems is eased as resulting forms correspond to existing words: for instance *general* instead of *gener* further to the stemming of *generalize*.

The same dictionary is used by Pentheroudakis and Vanderwende (1993). First, authors build the inflexional lexicon according to the grammatical information indicated in the dictionary. They build then the derivational lexicon:

- as in Krovetz (1993), the system checks if the definition of an entry contains words formally close to the entry: as the definition of *journalism* contains the word *journal*, both words are supposed to be related semantically as well;
- if an entry indicates suffixes, the system is able to construct suffixed words on the basis of the entry word. For instance, *journalism* indicates suffixes -istic/Adj and -istically/Adv and the system is then able to construct *journalistic/Adj* and *journalistically/Adv*;
- entries are considered semantically close if their definitions contain formally close words. Like in the case of *geographer* / *geography*, and *cartographer* / *cartography*.

Hathout (2001) uses a dictionary of synonyms. Simples words of this dictionary are compared and recorded as morphologically related if they present a suffix alternation and if this alternation is verified on at least one more pair of words. For instance, the alternation *er|ation* is recorded because it is verified on three verbal bases: *adorer*, *vénérer* and *permuter*.

Using corpora Corpora as well provide with context useful for the acquisition of lexical resources.

The hypothesis in Xu and Croft (1998) establishes the semantic proximity between words if they occur in a text within a window of given size (50 to 100 tokens). Words are considered to be related morphologically if a stemmer detects that they share at leat three first characters. The association score between these words is further computed and output pairs thus filtered. Word pairs can be grouped into morphological families such as: *uniformity*, *uniformly*, *uniformed*, *uniforms*, *uniform*. This approach provides with resources which depend on the reliability and completeness of used corpora. A similar experiment has been realized on medical French corpora (Zweigenbaum, Hadouche, & Grabar, 2003). The common string length is set to 4 characters, and the window size to 50 to 200 tokens.

Using structured terminologies Semantic relations between terms, such as synonymy, hierarchy, etc. can also be used for the generation of lexical resources (Grabar & Zweigenbaum, 2000).

Using corpora and lists of terms Jacquemin (1997) uses list of controlled terms and their variants from corpora to induce new lexicon of simple words. For instance, term pairs:

{*gene expression*, *genic expression*}
{*gene expression*, *genes expression*}

allow to induce word pairs {*gene*, *genic*} and {*gene*, *genes*}.

Using pairs of suffixes and their frequencies In Gaussier (1999), the comparison of words is supported by the frequency of the suffixes, which is supposed to indicate the reliability of the semantic context. If the frequency is high enough (at least 2), words are considered as

related lexically and semantically. For instance, words *deplorable* and *deploringly* present the alternation of suffixes *able|ingly*. As the same pair of suffixes occurs in other word pairs, this pair of words is considered as candidate to the semantic relation.

Application of distributional approaches To discover the most frequent affixes of a given language, Déjean (1998) applies the distributional approach. If, after a sequence of characters an important number of different characters can occur, the system induces the morpheme border:

- direc- can be followed only by *t*, and there is no morpheme border ;
- direct- can be followed by *i*, *l*, *o* and *e* (as in direct*ion*, direct*ly*, direct*or* or direct*ed*), and there is morpheme border.

In order to enrich the list of affixes, the system verifies is already registered affixes can alternate with other affixes. In this way, on the basis of words already known by the system (light, light*s*, light*ed*, light*ing*, light*ly*, light*er*), it is possible to induce new related words (light*ness*, light*est*, light*en*).

Schone and Jurafsky (2001) bootstrap the acquisition of suffixes from the corpora with a method similar to Déjean (1998), but prefixes are identifies in the same way as suffixes. The system is able to process not only prefixation and suffixation, but also the *circumfixation*, simultaneous adding of prefix and suffix to a basis. The first filtering of word pairs is done with Latent Semantic Analysis (LSA). The word pairs are then confirmed with similarity measure and weighted by the frequency of affixes and syntactic tag of words.

Application of learning algorithms Learning algorithms are applied in work such as Bosch, Daelemans, and Weijters (1996); Pirrelli and Yvon (1999); Theron and Cloete (1997). These algorithms are trained on a set of positive examples which are generalised in order to provide new word pairs. The performance of these algorithms depends on lexical regularities and analogies of a language, and on the completeness of the acquisition set.

Bosch et al. (1996) use as the acquisition set a Dutch lexicon Celex (Burnage, 1990). Pirrelli and Yvon (1999) propose a lazy algorithm, which in 4-tuple of words *eat:eater* = *cheat:X*, is able to compute the form of missing word *X* (*cheater*) through the exploitation of analogy of word formation. Theron and Cloete (1997) use a set of word pairs to deduce rules of their formation and induce new word pairs. These rules are expressed in a model of two-level morphology (Koskenniemi, 1983).

Using of lexeme formation rules Lexeme formation rules as proposed by linguistic studies can also provide the reliable basis for the acquisition of lexical resources. These rules propose all the "possible", but not necessarily existing, lexemes. As they over-generate results, they must be coupled with filtering process.

Rules proposed by constructional morphology (Corbin, 1987; Corbin, 1991) are applied in the project MorTAL (Dal, Namer, & Hathout, 1999; Hathout, Namer, & Dal, 2001). Automatically generated forms are validated through the reference list compiled from TLFi (Trésor de la Langue Française) and through the Internet (Namer, 2002b).

In the project Verbaction, Berche, Mougin, Hathout, and Lecomte (1997) use a list of verbs extracted from TLF. The Unix tool findaffix is used to detect rules for the formation of deverbal nouns. This is followed by manual filtering. In the further extension of this

project (Tanguy & Hathout, 2002), searching and filtering of new affixed lexemes is realized on Internet.

Viegas, Gonzales, and Longwell (1996) validate generated lexemes through a dictionary and corpora.

4.9.2 Acquisition of multilingual lexicon

Alignment and acquisition of multilingual terms and lexicon is usually bootstrapped from parallel (translated documents) or comparable (non translated documents, but which have the same topic) corpora. Among various work, one can distinguish lexical (simple lexical units) and terminological (multilexical units) alignment. Word alignment can be done on the basis of comparison of the length of words in two languages (Gale & Church, 1991), on the basis of symetric translations (Hiemstra, 1998), on cognates (words from different languages which present the formal similarity) (Simard, Foster, & Isabelle, 1992), on syntactic patterns of terms (Vintar, 2001), on cooccurrence of words and on bilingual dictionaries extracted from parallel (Kaji & Aizono, 2001) or comparable (Fung & McKeown, 1997; Y.-C. Chiao & Zweigenbaum, 2002) corpora. Among techniques used for the alignment of complex expressions, one can identify candidates for translation, when recording word sequences between the first and last words of a source term (Dagan & Church, 1994), when calculating the frequency and the expected position of target term (Eijk, 1993; Conley, 2002), when combining simple words of the target language in order to deciphe its translation (Smadja, McKeown, & Hatzivassiloglou, 1996), when looking for the alignments through the semi-automated system (Bourigault, Chodkiewicz, & Humbley, 1999), when exploring HTML tags in parallel documents (Grabar & Haag, 2003). Complex expressions can be found during the acquisition process or already known.

5 Survey of Natural Language Generation systems aimed at Patients or Doctors

5.1 Alphabetical list of systems with brief descriptions

Medicine is a popular domain for NLG developers. There have been numerous systems from the 1980s on. See Cawsey, Webber, and Jones (1997) for an overview of systems up to 1998. We have chosen a selection of seventeen of the best-known to include in this review. We briefly describe each in the list below.

5.1.1 BabyTalk

Date(s): 2006

Description: A project just beginning at the time of this survey which proposes to generate reports based on neonatal data from premature babies in a neonatal unit for hospital workers and for parents. Input data is time series health monitor data collected continually from probes attached to each baby (e.g. heart rate, blood pressure, etc.). Presentation is normally via graphics on computer monitors in the neonatal unit. Alternative forms of presentation (e.g. generated textual summaries) will be explored. Nurses normally write daily baby "diaries" for parents, an associated PhD project will attempt to generated these automatically.

Research Focus: Summarising time-series data for clinicians and patients and the interface between textual and graph-based information presentation.

References: Hunter, Reiter, & Sripada, 2006

5.1.2 CLEF

Date(s): 2005

Description: Generates reports based on simulated cancer patient Electronic Health Records for clinical staff. Part of a larger project on information retrieval and generation in biomedicine.

Research Focus: Building semantically linked knowledge structures based on data in medical records. Paraphrasing data from different viewpoints (e.g. problems, interventions and investigations).

References: Hallett & Scott, 2005

5.1.3 GRASSIC

Date(s): 1994

Description: Tailored booklets were produced for asthma patients. These were personalised using mail-merge according to the patients' health records, e.g. instructions for medication taken and anti-smoking information if the patient is a smoker. Education in the booklets was focussed on how to manage symptoms rather than on general knowledge of asthma. A large clinical trial with 800 patients and longitutinal over one year was used to evaluate GRASSIC. The personalisation group received personalised booklets (4 booklets during the course of the year) while a control group received no booklets, but normal oral education was given during consultations. Results showed that the group who received the personalised booklets reduced their hospital admissions by 51 % over the year of the trial.

Research Focus: Patient education and social benefits, especially reduction of hospital admissions.

References: Osman et al., 1994

5.1.4 HealthDoc

Date(s): 1995–1998

Description: An NLG system that produced health education documents aimed at patients about illness management (e.g. living with diabetes), how to follow medical guidelines (e.g. how to prepare for bowel surgery) and general health education (e.g. about smoking). The inputs were electronic medical records and other patient information related to personality (e.g. does the patient feel "in charge" of her health) and technical medical literacy. Many possibilities were discussed in the 1997 paper, but the exact nature and extent of user tailoring that was actually implemented is not clear. The proposed system was ambitious but only part was implemented (authoring tool and sentence repair mechanism). HealthDoc uses a "generation by selection and repair" NLG technique. This has a master document that represents content both in English and in a knowledge representation language. Content is selected and "repaired" by fixing anaphoric references and rhetorical relations. Final output was produced by the PENMAN realiser.

Research Focus: The research focus was on three issues: user tailoring (to health record, and aspects of personality); the design and construction of a tailorable "master document" and on linguistic problems.

References:
DiMarco, Hirst, Wanner, & Wilkinson, 1995;
Hirst, DiMarco, Hovy, & Parsons, 1997;
DiMarco et al., 2005

5.1.5 Linguistic String Parser

Date(s): 1986

Description: Early "generation" from either a stroke database, or from output of a decision support system. It generated isolated sentences by reverse-parsing, the sentences were not linked into a structured text.

Research Focus: Data-to-text by reverse-parsing.

References: Li, Evens, & Hier, 1986

5.1.6 MAGIC

Date(s): 1996–1997

Description: English medical, intensive care data
From `http://i3p-class.itc.it/projects_scheda.asp?id=14`: "MAGIC is an intelligent multimedia presentation system for the medical domain. After a patient has heart surgery, the physicians in the operating room (OR) must inform the caregivers in the intensive care unit (ICU) what happened during the surgery in order to prepare for the patient when he/she arrives in the ICU. MAGIC replaces the OR physicians in this scenario by presenting similar information using coordinated text, speech and graphics."

Research Focus: Integration and coordination of graphics, speech and text. Evaluation.

References:
McKeown, Pan, Shaw, Jordan, & Allen, 1997;
McKeown et al., 2002

5.1.7 MDA (Multilingual Document Authoring)

Date(s): 2000

Description: Interactive multilingual document editor tools were built using extensions to DTDs (document type definitions) for semantic drop-down menu choices to be rendered in English, French or German. The system produces information about pharmaceutical products.

Research Focus: Semantic grammars and multilingual issues.

References: Brun, Dymetman, & Lux, 2000

5.1.8 MeDView

Date(s): 2002

Description: Simple template-based NLG in an applied oral medicine application. Summarization of electronic patient records using text and graphics.

Research Focus: Building the underlying corpus of oral medical examinations.. Knowledge acquisition from exerts to develop protocols for data entry which facilitate a formalised logical interpretation of database fields that is suitable for computer-based reasoning.

References: Torgersson & Falkman, 2002

5.1.9 MIGRAINE

Date(s): 1994

Description: Tailored, interactive information for migraine patients.

Research Focus: User tailoring to groups of migraine patients, individual migraine patients and to the previous dialogue.

References: Mittal, Carenini, & Moore, 1994

5.1.10 PERSIVAL

Date(s): 2001

Description: Quote from `http://www1.cs.columbia.edu/nlp/projects.html`: "PERSIVAL (Personalized Retrieval and Summarization of Image, Video And Language resources) aims to provide personalized access to a distributed patient care digital library. PERSIVAL is a joint research initiative between the fields of NLP, human-computer interaction, medical informatics, video processing, library and cognitive science. Key features of PERSIVAL include personalized access to distributed, multimedia resources available both locally and over the Internet, fusion of repetitive information and identification of conflicting information from multiple relevant sources, and presentation of information in concise multimedia summaries that cross-link images, video, and text. When the latest medical information is provided at the point of patient care, it can help practicing clinicians to avoid missed diagnoses and minimize impending complications. When expressed in understandable terms, it can empower patients to take charge of their healthcare."

Research Focus: Summarising and merging results from searching a database of medical journal articles (independent research on textual and video summaries); automatic identification of medical terms; query interface; automatic layout.

References: Elhadad & McKeown, 2001

5.1.11 PIGLIT

Date(s): 1993–1998

Description: Planned small explanatory hypertexts based on electronic health records. It linked explanation plans ordered by relevance to patient and number of preconditions.

Research Focus: Personalisation, user modelling and selection of relevant content.

References:
Binsted, Cawsey, & Jones, 1995;
Cawsey, Binsted, & Jones, 1995a;
Cawsey, Binsted, & Jones, 1995b;
Cawsey, Jones, & Pearson, 2000

5.1.12 PILLS

Date(s): 1998

Description: PILLS explored the possibility that a computational tool, based on language generation, might allow a company to produce technical documentation more cheaply and more quickly; the production of medical information documents by the pharmaceutical industry served as a test domain. The main problem in applying NLG commercially

was to find a convenient way of creating and maintaining the content model. PILLS used a method known as 'WYSIWYM editing': instead of presenting the knowledge by means of a diagram, the system generated a feedback text through which editing operations was performed by opening pop-up menus on mouse-sensitive phrases.

An innovation in PILLS, compared with earlier WYSIWYM systems, was that the program could generate documents of several different types, the overlapping content being defined only once in a master model. A second innovation was that the ontology employed during knowledge editing was derived, in part, through an automatic extraction from a large medical database, the Unified Medical Language System or UMLS. As a result, the program began to address the problems of scale that would arise in a commercial application, with thousands of lexical entries for some common medical categories like diseases and ingredients.

Research Focus: The feasibility of WYSIWYM for producing multilingual multi-user documentation.

References:
Power & Scott, 1998;
Bouayad-Agha, Scott, & Power, 2001;
Bouayad-Agha, Power, Scott, & Belz, 2002

5.1.13 STOP

Date(s): 2000

Description: The STOP NLG system (Reiter, Robertson, & Osman, 2003; Lennox et al., 2001) produced personalised smoking cessation advice based on a questionnaire about health, smoking habits and smoking pros and cons. Over 2500 smokers took part in a clinical trial to compare cessation rates six months after receiving either a personalised letter, a generic letter, or a thank you letter with no smoking advice at all. There was no significant difference between cessation rates for people who received tailored letters and people who received non-tailored letters. Both groups had higher cessation rates (3.5 % or 30 out of 857 in the tailored group and 4.4 % or 37 out of 846 in the generic group) than the group who received just a thank you letter (2.6 %, or 22 out of 850). These results showed that perhaps elicitation of user details and user tailoring was not extensive enough, or perhaps a generic letter is sufficient. The study included a biochemical saliva test to verify the claims smokers who said they had quit (154 smokers, or 6 %, claimed to have quit, but only 89, or 3.5 %, were validated by the test as having quit smoking).

Research Focus: Personalised, persuasive generation, education, knowledge acquisition from experts, large-scale evaluation.

References:
Reiter et al., 2003;
Lennox et al., 2001

5.1.14 SumTime-Neonate

Date(s): 2003

Description: Generated summaries of time-series neonatal data from monitors attached to the patient (e.g. blood pressure, heart rate, etc.).

Research Focus: Proof-of-concept, comparison of text and graphics presentations.

References: Sripada, Reiter, Hunter, & Yu, 2003

5.1.15 SUREGEN-2

Date(s): 2003

Description: Applied NLG system (deployed?) – a clinical information system that generates German medical findings, procedure reports and referral letters from a medical ontology.

Research Focus: Simple, workable, application-driven solutions.

References: Hüske-Kraus, 2003

5.1.16 TAS

Date(s): 2005

Description: Generates summaries of references to journal articles that are relevant to patients' medical records.

Research Focus: Personalisation of medical records, evaluation.

References: Elhadad, McKeown, Kaufman, & Jordan, 2005

5.1.17 TraumAID/TraumGEN

Date(s): 1999

Description: Part of a larger decision support system which compares clinicians' decisions about patient management to those it produces automatically. NLG is used to generate "critiques" of physicians' decisions.

Research Focus: Microplanning and coherency.

References:
Carberry & Harvey, 1997;
Cawsey et al., 1997

5.2 Comparison Tables

We now compare the systems along four dimensions in the following tables: applications (Table 1), knowledge used (Table 2), user models and personalisation (Table 3) and evaluation (Table 4).

Date	Name	DOMAIN (branch of medicine)	USERS (intended audience)	INPUT (data source(s) input to generator)	OUTPUT (type of document(s) generated)	Technology
1986	Linguistic String Parser	Stroke	Medical Staff	Strings parsed from documents in a Stroke Database (a database of human-authored stroke case reports)	A list of unrelated sentences summarising stroke cases.	Reverse string parsing
1994	MIGRAINE	Migraine	Patients	Patient questionnaires, data entered by medical staff, stored user models.	Interactive documents.	Rule-based(?)
1994	GRASSIC	Asthma	Patients	Patient Health Records (computer-based record system on spreadsheets, e.g. asthma drugs taken, smoker, information requests)	Printed information booklets on asthma management, how to use medication, relaxation techniques, lifestyle, etc.	Mail-merge
1995	PIGLIT	Diabetes	Patients	Patient medical records	Hypertext explanations of patients' medical records.	Template-based plans
1997	HealthDoc	Health education (many diseases)	Patients	Electronic medical records and personal information about patients (e.g. psychological factors)	Documents (no examples in paper).	Generation by "selection and repair" from a master document.
1997	TraumAID/ TraumGEN	Medical decision support	Medical Staff	Output from an system that compares patient management plans (physician's plan vs. automatic plan).	Document criticising the physician's plan.	Rule-based.

Date	Name	DOMAIN (branch of medicine)	USERS (intended audience)	INPUT (data source(s) input to generator)	OUTPUT (type of document(s) generated)	Technology
2000	MAGIC	Transfer from operating theatre to intensive care	Medical Staff	Data from operating theatre monitors and data entered by operating theatre staff and online patient records.	Multimedia (text, graphics and speech) reports to inform clinicians of a patient's status on handover from one clinical team to another.	Multimedia planner, content planner and lexical chooser. Realisation by FUF/SURGE grammar and text-to-speech.
2000	MDA	Pharmaceutical products	Patients	Author input (MDA is an interactive authoring system)	Documents about pharmaceutical products.	XML and extensions to DTD
2000	STOP	Smoking cessation and related diseases	Patients	Patient Questionnaire (e.g. smoking habits, medical conditions, etc.)	Patient information leaflets	Rule-based
2001	PERSIVAL	Cardiology	Medical staff	Electronic health records and a database of journal articles.	Textual summaries of journal findings.	Rule-based
2002	PILLS	Pharmaceutical products	Patients, Medical Staff, Pharmacists	Product data.	Information leaflets about pharmaceutical products.	Rule-based
2002	MedView	Oral medicine	Medical Staff, Students (Patients in future)	Electronic case records (patient examinations and photos)	Multimedia (text and graphics) medical histories, discharge summaries and educational materials	Template-based
2003	SUREGEN-2	Cardiology	Medical Staff	Interactive input.	Medical "findings", procedure reports, referral letters.	Template-based

Date	Name	DOMAIN (branch of medicine)	USERS (intended audience)	INPUT (data source(s) input to generator)	OUTPUT (type of document(s) generated)	Technology
2003	SumTime-Neonate	Neonatal care and related diseases	Medical Staff	Time-series data from monitors attached to the patient (e.g. blood pressure, heart rate, etc.).	Summary reports for medical staff.	Signal processing and rules.
2005	TAS	General	Medical Staff and trainee medical staff.	Patient records and journal articles returned by a search.	Summary report of journal articles for medical staff tailored to patient.	Rule-based?
2005	CLEF	Cancer	Medical Staff	Simulated patient records.	Summary reports for medical staff.	Rule-based
2006	BabyTalk	Neonatal care and related diseases	Medical Staff and parents of Patients	Time-series data from monitors attached to the patient (e.g. blood pressure, heart rate, etc.)	Summary reports for medical staff and parents	Not yet known

Table 1. Comparison of Applications (systems in date order)

Table 1 shows the seventeen system in data order. It compares application types (or domain), intended users, system inputs and outputs and system technologies.

The systems cover a wide variety of medical topics. Roughly half generate documents for patients and some generate from health records. Of these, GRASSIC, PIGLIT and HealthDoc, generate from EHRs. HealthDoc and PIGLIT summarise and explain the EHRs themselves, while GRASSIC uses EHR data to generate personalised education materials. STOP generates from patient questionnaires. Using knowledge from EHRs is obviously preferable to forcing patients or medical staff fill in long questionnaires. Another valuable source of medical data is time-series data from monitors attached to patients and from these can provide up-to-the-minute status reports. The remaining systems generate summaries for medical staff. SumTime-Neonate, MAGIC and BabyTalk generate from data from monitors attached to the patient (MAGIC refers to EHRs in addition). The Linguistic String Parser, MedView, PERSIVAL and CLEF also generates from EHRs. PILLS generates from product data.

The majority of these systems are rule-based, SumTime-Neonate additionally employs signal processing algorithms to detect features in the input time-series data.

Date	Name	Expert domain knowledge (for planning document content e.g. for different kinds of users)	Generic medical knowledge source (used for content or lexical/phrasal alternatives)	Linguistic knowledge (used for microplanning linguistic expression e.g. domain-specific lexical/phrasal alternatives used for, for example, user style preferences)
1986	Linguistic String Parser	Analysis of human expert-authored case notes.	none	none
1994	MIGRAINE	Sample information sheets written by medical experts used to derive content and document structure.	Knowledge from ethnographic studies of migraine. Knowledge about therapies.	??
1994	GRASSIC	MailMerge document designed and written by medical experts	None	None
1995	PIGLIT	Expert criticisms & suggestions received on prototype system.	Typology from UK medical Read codes	None
1997	HealthDoc	Mock up of an expert authoring tool. Some informal interaction with experts.	None	Co-reference and rhetorical structure knowledge (from linguistic theories?).
1997	TraumAID/ TraumGEN	From expert system (presumably developed using KA from experts).	From expert system (presumably developed using generic medical knowledge).	None
2000	MAGIC	KA from expert-authored admission notes used for content. Audio recordings of expert briefings on patient status.	??	Lexicalisations derived from experts' notes and briefings. FUF/SURGE grammar.
2000	MDA	None	Corpus of drug notices from a medical textbook.	Derived from a corpus of domain texts.
2000	STOP	KA from medical staff	None	Derived from a corpus of domain texts
2001	PERSIVAL	no	Knowledge of medical terms and from on-line journal articles.	Linguistic knowledge used to identify repetitions and contradictions in search results and consequently to express these in the summary.
2002	PILLS	None	Medical database UMLS used to derive ontology.	None

Date	Name	Expert domain knowledge (for planning document content e.g. for different kinds of users)	Generic medical knowledge source (used for content or lexical/phrasal alternatives)	Linguistic knowledge (used for microplanning linguistic expression e.g. domain-specific lexical/phrasal alternatives used for, for example, user style preferences)
2002	MedView	Expert oral medical knowledge elicited to derive protocols for database entry.	None	None
2003	SUREGEN-2	No	Medical ontology	Lexical items derived from UMLS.
2003	SumTime-Neonate	KA from medical staff, expert-authored corpus of domain texts.	None	Linguistic analysis of corpus of domain texts.
2005	TAS	None	None	Linguistic knowledge on combining phrases from articles to form coherent summaries.
2005	CLEF	No	Generic cancer knowledge, medical ontology.	None
2006	BabyTalk	Planned	Not known	Planned

Table 2. Comparison of knowledge used both in building the system and during generation (systems in date order)

Table 2 shows the seventeen system in data order. It compares knowledge and data used to build the systems: knowledge acquired from domain experts, knowledge from generic medical sources (such as textbooks, medical ontologies and term banks), and linguistic knowledge from grammars or derived from linguistic analysis of domain documents.

Systems tend either to acquire knowledge from experts, or from generic medical sources, such as UMLS. When developers have access to relevant experts as, for example, the STOP and MAGIC developers did, they tend to use them to derive both domain knowledge (through interviews or audio recordings made while they are working) and linguistic knowledge (by asking them to write notes or example output texts). This involves time-consuming and expensive hand-crafting, but leads to very high quality texts. On the other hand, systems such as CLEF, PILLS and MDA rely exclusively on generic medical knowledge which has the advantage that there is no hand-crafting involved.

Date	Name	Takes account of users' Domain knowledge	Takes account of relevance to users (e.g. to select relevant content)
1986	Linguistic String Parser	No	Uses stroke record to select content.
1994	MIGRAINE	No	Uses individual patient data, including discourse history and group data about migraine patients to select content.

Date	Name	Takes account of users' Domain knowledge	Takes account of relevance to users (e.g. to select relevant content)
1994	GRASSIC	No	GRASSIC takes account of the asthma drug taken and other user information from an electronic health record & varies content of the document accordingly.
1997	HealthDoc	Yes (but there are no examples of this in the paper)	HealthDoc assumes relevance from the medical record (which is is the user model). Types of information are selected according to mock up of user characteristics (e.g. likely to read technical information).
1995	PIGLIT	Yes, infer users' knowledge from length of illness & vary use of tech. terms accordingly	infer relevance from medical record and vary content accordingly
1997	TraumAID/ TraumGEN	No	No
2000	MAGIC	No	No
2000	MDA	No	User makes authoring choices.
2000	STOP	No	Yes, uses user information from questionnaires to vary content.
2001	PERSIVAL	No	Yes, uses EHR to determine relevance.
2002	PILLS	No	No
2002	MedView	No	No
2003	SUREGEN-2	No	No
2003	SumTime-Neonate	No	Yes, identify episodes in input signal and vary content.
2002	TAS	No	Yes, selects information from journal articles relevant to patient's medical record.
2005	CLEF	No	Yes, generates from user queries.
2006	BabyTalk	Not known	Not known

Table 3. Personalisation in the system (systems in date order)

Table 3 compares personalisation (or user tailoring) in the systems: both personalisation in terms of domain knowledge and personalisation of content by taking into account the relevance to users. Overall, personalisation efforts have concentrated very much on choosing content rather than on vary linguistic expression of that content for different individuals. None of the systems take account of domain knowledge except PIGLIT which does use a crude measure of length of illness to estimate familiarity with relevant medical terms and provides additional explanations if necessary and HealthDoc which has a questionnaire asking if users are likely to read technical information and it selects content accordingly. Otherwise, it is presumably

assumed that all medical staff have the same level of medical knowledge and that all patients also have a uniform level of knowledge. This is clearly unrealistic.

When the users are medical staff, personal skills, for instance literacy or numeracy skills, or first language, are never taken into account and presumably it is assumed that they will also have equally high levels of these skills, although this may not necessarily be the case. Obviously this is more important when users are patients, but again, none of the systems take account of this and it is especially important when numerical data is being communicated.

Finally, none of the systems take account of users' preferences, such as their preferences for style and layout of the output document.

Date	Name	User Comprehension of generated documents	User Preferences	Usability Study	Longitudinal Study with Users (e.g. 3-year clinical trial)	Other Evaluation Technique
1986	Linguistic String Parser	No	No	No	No	No
1994	MIGRAINE	No	Interviews with small numbers of patients after using the system asked if they liked it, if they learnt anything and if it was helpful.	Small numbers of patients were observed using the system.	No	No
1994	GRASSIC	No	No	"usefulness" questionnaire	A one-year clinical trial with 800 patients. One group received personalised booklets (4 booklets during the course of the year) while a control group received no booklets and the normal oral education given during consultations. Results showed that the group who received the personalised booklets reduced their hospital admissions by 51 % over the year of the trial.	User evaluation questionnaire on each section of each booklet asking to rate newness, readability and usefulness.

Date	Name	User Comprehension of generated documents	User Preferences	Usability Study	Longitudinal Study with Users (e.g. 3-year clinical trial)	Other Evaluation Technique
1997	HealthDoc	No	No	No	No	No evaluation in 1997 or 2005 papers.
1995	PIGLIT	Yes, self-assessed at interview	Yes, interviews	User interviews	No	No
1997	TraumAID/ TraumGEN	No	No	No	No	Comparison of two styles of automatically generated texts and evaluation by one human subject of coherence and quality.
2000	MAGIC	No	No	No	No	Compared content of NLG output with content of clinicians' verbal briefings and written notes. Found 78 % recall in NLG compared to notes and 55 % compared to verbal briefs (because NLG in information is finer-grained).
2000	MDA	No	No	No	No	No

Date	Name	User Comprehension of generated documents	User Preferences	Usability Study	Longitudinal Study with Users (e.g. 3-year clinical trial)	Other Evaluation Technique
2000	STOP	No	No	None	6-month clinical trial with over 2500 smokers that compared cessation rates six months after receiving either a personalised letter, a generic letter, or a thank you letter with no smoking advice at all. The results showed there was no significant difference between cessation rates of people who received tailored letters and people who received non-tailored letters.	No
2001	PERSIVAL	No	No	No	No	Pilot evaluation of precision and recall in generated summaries of 27 journal articles. Precision was high (92 %) but recall was very low (50 %).
2002	PILLS	Study on layout and comprehension.	No	No	No	No
2002	MedView	No	No	No	No	No formal evaluation, but the system has been deployed in several clinics for two years.

Date	Name	User Comprehension of generated documents	User Preferences	Usability Study	Longitudinal Study with Users (e.g. 3-year clinical trial)	Other Evaluation Technique
2002	SUREGEN-2	No	No	No	No	No formal evaluation, but the system was integrated into a hospital information system and texts appear to be acceptable.
2003	SumTime-Neonate	No	No	Task-based (can medical staff use system summaries to make correct treatment decision?)	No	No
2005	TAS	No	Task-based study with medical staff. Compared standard search results, generic summary and personalised summary all with gold-standard. Subjects preferred personalised summary.	Usability – videos of medical staff using system. Task was completed more successfully with personalised summary.	No	No
2005	CLEF	No	User queries	Study on user queries	No	No
2006	BabyTalk	Not known	Not known	Not known	Not known	Not known

Table 4. System Evaluation (systems in date order)

Evaluation of the seventeen NLG systems in Table 4 ranges from large-scale longitudinal clinical trials with users designed to demonstrate positive medical outcomes, to evaluations of part of the NLG process, such as content selection (as in MAGIC). Large-scale studies were successful in GRASSIC, but unsuccessful in the STOP project. Clinical trials are very expensive and time consuming. It is also very hard to get access to patients and medical staff for such trials.

Overall, it is surprising how little evaluation has been done of most of the systems in our survey and this leaves us with very little data on how well the systems would work in practice. There are two exceptions, MedView and SUREGEN-2, both of which are deployed in clinics.

5.3 Use of Multimedia

Use of multimedia in NLG medical systems has been disappointing to date. MAGIC should have had full multimedia coordination of graphics, text and speech, in an ambitious integrated application. In practice, however, the graphics and speech coordinators were prototypes only and were not included in the final system evaluation. MedView generated both speech and graphics (photographs of patients' mouths), but there was no attempt to link these in any formal way. PERSIVAL investigated summarisation of both videos and texts, but these were not integrated activities.

6 Survey of laymen/expert (e.g. patient/doctor) ontology translation systems

Laymen and experts, e.g. patients and doctors, often use different terminologies, which reflect the underlying differences in ontology, i.e. conceptualisation of the domain. Typical differences between expert and laymen ontologies and vocabularies are that laymen vocabularies are simpler, they are less elaborate in upper and lower levels of the ontology and they have shallower taxonomies with fewer intermediate categorical distinctions. Laymen cover only a small part of professional vocabulary, mainly the taxonomic orders. Laymen vocabularies are also fuzzier; the terms cover a large range of referent types or vice versa (B. Smith & Fellbaum, 2004). Ontologies can differ in meta-model, i.e. the syntax, logical representation, semantics of primitives, and expressivity of ontology language, or they can differ in ontology model level, i.e. conceptualisation mismatches or explication mismatches (Klein, 2001). Typically lay expert ontologies differ in conceptualisation, they do not have the same scope nor model coverage and granularity. They also differ in explication, for example, the style of modelling and the terminology.

Information systems often utilise information sources, e.g. documents and thesauri, which reflect the expert view. Thus means for mapping of concepts and properties in different ontologies are needed for an information system to be able to understand both expert and laymen vocabulary. Although many information systems in the medical domain utilise ontological knowledge sources, the problem of handling differences in laymen and expert ontologies in information systems is little explored.

A recent project within the medical domain that addresses the problem is Medical WordNet (B. Smith & Fellbaum, 2004). It aims at capturing and describing laymen vocabulary for the medical context. The result will be an ontology and a database of example sentences that illustrate the use of concepts in the ontology. The intended use of the ontology is for NLP tasks like machine translation, text summarisation and question answering. The example database will hold both statements considered to be facts (true) and beliefs (both true end false) as held by laymen.

The lay expert problem has also been studied in other domains. In the BEST project (Laarschot, 2005) the goal is to bridge the gap between laypersons' terminology and documents written by experts in the domain of tort law. A user ontology captures laypersons' view of cases and hold common sense concepts used to describe these. A legal thesaurus holds the expert view and is used to annotate legal documents. The approach in this project to mediate between these two different ontologies was to use a third neutral knowledge source, in this case the structure of the law, to which the two ontologies were linked. Hovy (2003) exemplifies the lay expert problem in access of government databases. In this domain it is important that the domain terminology is detailed in order to make all relevant technical distinctions and allow experts to use the system. At the same time it must also include lay

terminology to allow laymen to use the system. The approach taken is to use one integrated ontology with more general terms at the top that the laypersons can use, and more specialised at the bottom of the ontology that the experts use. Through links between the different types of terms the laypersons can navigate the data starting with common sense terms, e.g. price, which is refined step by step to the right expert concept, e.g. cost, charge, amount, fee, payment. A similar approach is used in an information-providing dialogue system in the domain of encyclopaedic information on birds (Jönsson et al., 2004). A bird encyclopaedia was used to extract information and populate a database. To do this an ontology for the encyclopaedia was developed, which reflected the experts view of the domain. A question corpus was used to extract the laypersons view of the domain. The two ontologies were then integrated. Different categorisations were allowed by use of multiple inheritance, and vague user properties were introduced and linked to the properties in the expert ontology (Flycht-Eriksson, 2004).

Although little work has been done on how to relate and use lay expert ontologies, there has been much general work on ontology alignment and merging. Mapping of ontologies can be done as a one-to-one mapping between pairs of ontologies or through the use of a global ontology. The first approach was used by the dialogue system above, and the latter in the tort law and e-gov domains discussed above. Mapping can also be done as an alignment where the original ontologies are preserved but at least one of them is modified to match the overlapping parts of the other, as in the tort law example, or as merging where a new ontology is formed either as a union or intersection of the original ontologies, as in the dialogue system. There have also been many tools developed to support the mapping process. For a comprehensive survey of ontology alignment and merging, and the available tools, see Predoiu et al. (2004).

7 Empirical studies with patients on information provision

This chapter concerns empirical studies on computer-based provision of information; user tailoring of information, i.e. effects of tailoring information to individuals; how well patients understand technical terms; do icons help; and a pointer to a survey of studies on physicians' communication with patients.

7.1 Providing computer-based information

A 2002 European Union survey (Spadaro, 2003) on where Europeans look for health information revealed that the vast majority (45 %) ask a health professional (pharmacist, doctor or chemist); this is more than twelve times the numbers who search the Internet for information (3.5 %). Over the intervening four years, the numbers using the Internet have increased, but not much. Jones et al. (2006) found in their study of 384 cancer patients that only 14 % of patients used the Internet to look up medical information. In the future, the numbers using the Internet should increase still further. However, we do need to consider the implications of this for applications we develop if we are serious about developing technology for computer-based information provision that will be deployable in the near future. For instance, we should take accessibility issues seriously.

7.2 Tailoring patient education materials

Many of the following studies found significant positive effects on patient behaviour resulted from tailoring patient education materials. The medical community carried out many of these studies using mail merge technology, or similar. Fewer empirical studies, particularly large-scale ones, have been conducted by the NLG community. Where NLG technology was used, the results tended to be disappointing and there is conflicting evidence on whether patient tailoring is effective, or not. A good summary of issues in this area is Reiter and Osman (1997) and see the discussion at the end of this section.

7.2.1 Asthma Management Advice

Osman et al. (1994) used mail merge to automatically produce personalised information booklets for asthma patients. Sections of the letter were chosen based on information in a spreadsheet medical record. A large longitudinal study with 800 asthma patients over six months found that hospital admissions were significantly reduced (by 51 %) in the group receiving tailored information compared to a group receiving generic information.

7.2.2 Cancer Information

Jones et al. (1999) carried out a longitudinal study with 525 cancer patients over three months to compare their preferences on three kinds of cancer information: computer-based information tailored to patients' medical records, generic computer-based information and cancer booklets. They found that patients preferred computer systems. However, very few in the study did actually use the computer (only 20 out of 169 in the personalised information group and only 4 out of 155 in the general information group). They also found that patients preferred information based on their medical records, but this was countered by the fact that 49 % of these people thought that the personalised information was too limited!

More recently, Jones et al. (2006) tested the hypothesis that methods of selecting and printing information for cancer patients could affect their interactions with others, improve their level of support, reduce their anxiety levels and increase their feelings of wellbeing. There were eight subject groups (each of around 50 patients). Information was selected for patients either automatically or in an interactive session, the information selected was either personalised or general and patients were either given information containing advice on anxiety, or not. Again, this study failed to show that personalisation of information had a significant effect.

7.2.3 Dietary Advice

M. K. Campbell et al. (1994) produced dietary messages (using mail merge?) for patients to persuade them to decrease their fat intake and increase their fruit and vegetable intake. The study found that tailored information significantly decreased fat intake four months after receiving the message. No group increased fruit and vegetable intake! Significantly more people remembered receiving a tailored message than than a generic message.

7.2.4 Smoking Cessation Advice

Tailored smoking cessation letters were produced (not using an NLG system) by Strecher et al. (1994). Two studies were carried out. The first study (N=51) found significantly more moderate-to-light smokers (30 %, roughly 7–8 out of 25) gave up smoking within six months of receiving a tailored letter, whereas only 7 %, or 1–2 out of 25)gave up in the control group (each of which received a generic letter). A second study (N=197) found significantly more moderate-to-light smokers (19 %) gave up smoking within four months of receiving a tailored letter, whereas only 7 % gave up in the control group (which received no letters). Letters were tailored according to consumption of cigarettes, interest in giving up, and "perceived benefits and barriers to quitting".

The STOP NLG system (Reiter et al., 2003; Lennox et al., 2001) produced personalised smoking cessation advice based on a questionnaire about health, smoking habits and smoking pros and cons. Over 2500 smokers took part in a clinical trial to compare cessation rates six month after receiving either a personalised letter, a generic letter, or a thank you letter with no smoking advice at all. The results did not reproduce those of the Strecher study above; instead, there was no significant difference between cessation rates of people who received tailored letters and people who received non-tailored letters. Both groups had higher cessation rates (3.5 % or 30 out of 857 in the tailored group and 4.4 % or 37 out of 846 in the generic group) than the group who received just a thank you letter (2.6 %, or 22 out of 850), showing that perhaps elicitation of user details and user tailoring was not extensive

enough, or perhaps a generic letter is sufficient. The low proportions of subjects who gave up smoking in the larger STOP study makes the high proportions who quit in the small Strecher study seem unrealistic. Indeed, closer examination of the two studies revealed that the STOP study included a bio-chemical saliva test to verify the claims smokers who said they had quit (154 smokers, or 6 %, claimed to have quit, but only 89, or 3.5 %, were validated by the test as having quit smoking), whereas the Stretcher study merely trusted smokers' claims about quitting.

7.2.5 Mammography Advice

Skinner, Strecher, and Hospers (1994) found tailored letters about mammograms are more effective (i.e. the recipient remembers it better and is more likely to have a mammography) than generic letters, especially for women of low socioeconomic status. Tailoring was done by gathering information from telephone calls (not automatic) and this is most likely where the advantage was gained, since the amount, type and quality of personal data that can be gathered by a human in a 10-minute phone call would vastly outperform any existing methods for automatic data elicitation.

7.2.6 Discussion

It is remarkable how conflicting the evidence is in this area! The medical informatics studies of Skinner et al. (1994), Strecher et al. (1994) and M. K. Campbell et al. (1994) all found positive effects of tailoring, as did the study of Osman et al. (1994). On the other hand, the studies with NLG systems (Reiter et al., 2003; Lennox et al., 2001 and Jones et al., 1999; Jones et al., 2006) failed to demonstrate significant effects of tailoring. The most obvious differences in the studies is the difference in outcome measures: e.g. fat consumption, smoking cessation, mammography take up, anxiety levels, user preferences, hospital admissions. Clearly what is measured can make a difference. However, two studies that measured a similar outcome, smoking cessation, produced very different results. The small study of Strecher et al. (1994) reported cessation rates between 7 % and 19 %, whereas the much larger study of Lennox et al. (2001) reported rates of between 2.6 % and 4.4 %. Lennox et al. (2001) results are likely to be more reliable since they validated smokers' claims to have given up by administering biochemical tests; Strecher et al. (1994) did not do this, instead they relied on smokers' claims.

Another major difference is the area of health that these studies address. Leaving aside smoking cessation, the successful studies were on asthma education, health education and mammography screening advice and the unsuccessful studies were on cancer education. Perhaps this difference is crucial to the positive effects of tailoring. In asthma, the provision of self-help education can clearly have a great effect because patients can largely manage their own care, take steps to reduce their symptoms and improve their quality of life. And if asthma patients receive personalised information, the benefits in motivating them towards self-help are even greater. Cancer, on the other hand, seems to require more complex treatments and surgery which are often outside the patients' immediate control (apart from giving consent) and there may not be such direct ways in which education can help them. It is perhaps this aspect that could make cancer patients a difficult group with which to demonstrate the benefits of tailored information, even if we only try to demonstrate that personalised information is easier to understand than generic information. The unsuccessful studies of Jones et al.

(1999); Jones et al. (2006) warn us that it is hard even to demonstrate that cancer patients prefer personalised information.

The final difference is in types and quality of patient data. Obviously hand-crafted patient data will be of higher quality than automatically elicited data. There is clearly a need for further research into exactly what kinds of patient data are most effective for tailoring patient information.

7.3 Patients' understanding and recognition of medical terminology

Zeng, Tse, et al. (2005) took 34 concepts in their original medical terminology (from NLM MedlinePlus) and in corresponding "consumer-friendly" synonyms (from UMLS), e.g. exanthema/rash and tried them on 10 people in multiple-choice questions. The results showed that the subjects were more likely to understand and recognise the "consumer-friendly" versions (mean score 15.4 out of 34) than the original medical terms (mean score 6.0).

For more on this, see the comprehensive survey of this area in chapter 3 which shows that, on the whole, patients understand far fewer medical terms than one would expect.

7.4 Use of pictures in patient information materials

Hameen-Anttila, Kemppainen, Enlund, Bush, and Marja (2004) found that pictograms did not help children understand patient information. Possibly the results would have been different if the pictures had been better and they had been used in a better context, i.e. in real information leaflets.

No further literature could be found on this subject, perhaps more extensive searching would reveal some.

7.5 Clinicians' communications with patients

A good survey of this area can be found in Back, Arnold, Baile, Tulsky, and Fryer-Edwards (2005). Mostly, it is about face-to-face communications, but some of this evidence would be extremely useful for building interactive systems. For example, the section on "Making Anticancer Treatment Decisions" mentions a large study with 999 subjects (Lee, Back, Block, & Stewart, 2002) which indicates that there are essentially 5 types of patient when it comes to making medical decisions, ranging from those who would prefer to leave all decisions to their doctor to those who want to find out for themselves and make their own decision (the groups are: "paternalistic: physician makes decisions", "physician-as-agent: physician makes decisions after considering patient input", "shared decision making: physician and patient make decisions together", "informed decision making: patient makes decisions after considering physician input" and "consumerism: patient makes decisions"). The paper suggests many questions to ask patients to find out which type they are, such as "Are you the kind of person who likes to hear all the numbers?" See also Baker, Eash, Schuette, and Uhlmann (2002) for guidelines on writing letters to patients.

7.6 Summary

This survey has been very brief and we have merely provided pointers into some areas. The area that we have concentrated on most is information tailoring where results are conflicting, clearly showing the need for more research, especially into what kinds of patient data are most effective in information tailoring.

8 Corpus Annotation Tools

This chapter lists annotation tools for corpora that may be of interest to members of the WP27 work group.

Since syntactic and orthographic annotation tools are largely automated and many are configurable I simply present some options in tabular form listing some of the constraints on each tool (for example the tag set for a POS tagger) and the development platform on which they are supported.[1]

Semantic annotation tools largely assist manual annotation, although some can be run in an automated fashion with training and supervision. Additionally semantic annotation tasks tend to be project specific and are less reusable. I therefore present some semantic annotation tools that may be of interest directly or more likely by reference.

Finally I present a list of IDEs and workbenches that we may want to use to integrate some of these tools to make the pipeline simpler to manage. An integrated approach would be beneficial if we were to build up the corpora and annotate them piecemeal (thus requiring many iterations of the annotation cycle) but limits the choice of tools and presents overheads in terms of installation, configuration, training and managing bugs.

8.1 Requirements

A set of annotation tools that can be used to mark up a range of different types of information in the corpora that we collect for WP27.

The toolset must cover English, French, German and Swedish and must provide coverage of a range of annotations.

8.2 Forces

This area is not a research priority so we do not want to develop our own tools where we can reuse existing, standard approaches and software. Where possible we should build on the skill-set of the work group rather than learning to use new tools.

Hopefully we will find tools that we can download and use with little development required, however for some annotation tasks this may not be possible and we may only be able to access details of an approach. In such instances we will have to develop the code for the tool ourselves.

A pipeline approach gives us the greatest flexibility in our choice of tools allowing us to use particular tools that are well-known within the discipline or within the work group.

An integrated approach stifles choice and may require us to do some development ourselves, but if we continue to gather corpora throughout the project it may be easier to rerun the annotation if a single integrated controller has been constructed.

[1] By development platform I mean the combination of OS (where applicable) and development language/environment.

If we choose to integrate the toolset we might find an IDE or workbench helpful. Some of these have built-in annotation tools covering many of the tasks that are likely to be of interest to us.

Ideally all of the tools would be fully automated, in practice this is not going to be the case. Some tools will require us to create training sets and perform training and tuning iterations, others will require manual oversight or seeding.

Some annotations can only be determined if another annotation has already been marked up. Whatever architectural model we choose this means that the output from some tools must be the input to others. This raises questions of what schema we will use and what approach we will take to transforming these streams between formats.

8.3 Annotation Types

Broadly we might want to add the following annotations, note that there are dependencies (e.g. some semantic annotations require prior linguistic annotation to have been successful):

- Orthographic Annotations
  - Document: type, bibliography, encoding, source ...
  - Structure: volume, chapter, page, paragraph, sentence ...
  - Purpose: title, footnote, sidebar, heading, figure, table ...
- Linguistic Annotations
  - Tokenization, principally word boundary
  - Syntactic annotation: terms, quotations, dates, abbreviations, names ...
  - POS: stemming, lemmatisation ...
  - Morphology: case, tense, number, gender ...
  - Grammar: verb, object, indirect object ...
- Semantic Annotations
  - Named Entity Recognition
  - Discourse-Level Annotation
  - Term integration
  - Reference/Coreference/Anaphora
  - Word Sense Disambiguation
- Other Annotations
  - Alignment
  - Style

- Genre
- Dialogue
- Context

8.4 Schema

If we want to mark-up multiple features then we need to consider what schema we should use. If we use a lot of different tools in a pipeline then we may need to convert between mark-up schemas, also if we want to maintain a lot of different annotation information on the same corpus then we require a complex schema or multiple schemas and therefore multiple annotations.

There is also the question of whether we should use our own schema (perhaps basing the design on an existing comparable schema) or use an existing (possibly standard) one. Some references for interest:

- LMNL see Tennison (2002)
- XDML see Devillers, Vasilescu, and Lamel (2002)
- TIGER see Brants and Hansen (2002) and König and Lezius (2000)
- SUSANNE `http://www.grsampson.net/RSue.html`
- GENIA see Kim, Ohta, Tateisi, and Tsujii (2003)
- MedDoc DTD see Charlet et al. (1998) and see below under Hospitexte.

If we design our own schema, does it need to be compliant with standards?

8.5 Tool Overview

There are many freely available corpus annotation tools. They have different system requirements in terms of development platform, types of annotation supported, types of corpus supported, types of domain supported, and so on.

Initially I have examined the following features for each tool (where appropriate):

Availability Ideally we will download tools from the Internet and use them directly. In practice for some annotations this will not be possible and we may only be able to access details of the approach taken, leaving us to implement the tool ourselves.

Format The input and output format of the tool, we need to know this so that we can identify which tools that can be chained together, and also so that we can consider what schema we will use.

Multilingual Can this tool be used for many languages or has it only been used with one?

Platform The technical platform required for the tool to run. If our choice of tools cover a lot of different technical platforms then this will make it hard for us to rerun the annotations on new data being added to our corpus/corpora. In practice we may choose different platforms to annotate corpora at different partner sites.

URL A reference providing details of the tool in question.

Reference Sites Projects, papers or other references to where the tool has been used in research or for commercial projects.

8.5.1 Orthographic Annotations

Document information

I am not aware of any tools for this (although see CLaRK below), but we should hold some information about the document in a header, for example:

- Source (e.g. a URL, an ISBN)
- Media (e.g. Internet, book, ephemera)
- Domain (e.g. expert medical, blog, newsgroup posting)
- Title
- Author/Company
- Date (of publication)
- Length (in tokens or bytes)
- Language (might be multiple)
- Location (e.g. country of publication)
- Restrictions (e.g. restrictions on reproduction)
- Format (e.g. HTML, XML, Word Doc, RTF, printed material)

This sort of information can be useful for classifiers (for example classifying into lay and expert domains) and is also useful for maintaining and distributing corpora and for a variety of other tasks.

We could follow standards for this and use:

- TEI Header format: `http://www.tei-c.org/P4X/HD.html`
- CES (which I think amounts to the same thing as TEI for the header) `http://www.cs.vassar.edu/CES/CES1-3.html`
- XCES `http://www.cs.vassar.edu/XCES/schema/#header`

We would need to agree what data items we will hold in the header, and what format and values they can have.

See Habert, Grabar, Jacquemart, and Zweigenbaum (2001) on bibliographic annotations for a text corpus of medical language.

Document Structure

Many of the documents that we put into the corpora may already have some sort of document structure mark-up or annotation (for example they might be HTML, RTF or PDF documents). In these instances the annotation task involves transforming the information held in these formats into some annotation scheme of our own so that we can retain as much structural information as possible about the document.

Other documents will have no structural data in them and so we will have to hand annotate them.

I am not aware of any tools that automate this process, when I worked on the BNC we used Emacs regular expressions, Perl, C and command line sed scripts to perform these transformations. I suspect that this is still the case for most corpus construction.

See also Bouayad-Agha (2000) on annotation of logical structure in patient information leaflets as part of the PILS project.

See the section above on Annotation schemas.

8.5.2 Linguistic Annotations

Tokenizers and Segmenters

I don't think that tokenization presents an issue for English, French or German but it may be a more complex problem for Swedish. If we like we could use a freely available tokenizer such as lttoken or QToken, although I think that for English, French and German at any rate tokenization is just a step during the process of segmentation.

The term *segmenter* is ambiguous, sometimes it refers to a chunker or shallow parser. I use it here to refer to tools that perform one or more of the following functions (in English at any rate):

- Tokenize on white space
- Mark punctuation
- Mark abbreviations
- Mark numeric literals, dates, quantities etc
- Mark foreign words and characters
- Mark other symbols and non-lexical items
- Mark sentence boundaries
- Mark multi-word units

Segmenters are not generally particularly complex, so I recommend that we write or choose a segmenter that matches our POS tagger in platform and format (for example MtSeg is a segmenter that comes with the Multext tagging suite).

Stemmers

I think that this is a much more complex problem for Swedish than for English, French or German, so some areas of the corpus may require specialist tools for stemming.

There are a variety of stemming algorithms that are commonly used for English, principally:

- Lovins (Lovins, 1968)
- Porter (Porter, 1980)
- Lancaster (Paice, 1990)

In order to increase our flexibility in dealing with languages other than English I have also included links to the NLTK Python libraries and Snowball, a language for writing stemmers that compiles into C or Java, so that we could write our own stemmer if required, although any stemmers that use the Porter algorithm should cover additional languages including French (Natalia Grabar, personal communication).

Note that I have marked the format as any for all of these tools as the input format can be plain text and we they are all downloaded as source code so the output format is down to us to determine.

Name	Availability	Format	Platform	Notes
Porter[2]	Download	any	C, Java, Perl, C#, Ruby, Python, VB, Prolog, ...	Porter algorithm as source code in a variety of languages
Lancaster[3]	Download	any	C, Java, Perl, Pascal	Implementations of the Lancaster (Paice/Husk) stemmer
Lovins[4]	Download	any	Java, C, Perl	Implementations of the Lovins stemmer from SourceForge
UEA-Lite[5]	Download	any	Java, Perl	A stemmer written in 2005 at UEA (Jenkins 2005)
Morphix[6]	Download	any	Common Lisp	Morphological analyser and generator for German
NLTK tokenize[7]	Download	any	Python	Porter and simple regexp and affix algorithms are supported
Snowball[8]	Download	any	Java, C	A language for writing stemmers written by Martin Porter

[2] http://www.tartarus.org/~martin/PorterStemmer/index.html
[3] http://www.comp.lancs.ac.uk/computing/research/stemming/Links/implementations.htm
[4] http://sourceforge.net/projects/stemmers/
[5] http://www.cmp.uea.ac.uk/Research/stemmer/
[6] http://www.dfki.de/~neumann/morphix/morphix.html
[7] http://nltk.sourceforge.net/
[8] http://snowball.tartarus.org/

POS Taggers

Some of these taggers are specific to English, and all generate output based on particular tagsets. I suspect that we will want to augment this list with taggers/tagsets for the other languages in the work group proposal. Note that most multi-lingual taggers require a manually tagged training corpus for each language/domain.

There are a variety of standard approaches, some require manual training or a manually-tagged training corpus

- Rule-based taggers see Brill (1992); Brill (1995)
- N-gram frequency (Viterbi algorithm) see Brill and Marcus (1992)
- Hidden Markov Model see Dermatas and Kokkinakis (1995)
- Decision Trees see Schmidt (1997)
- Cyclic Dependency Network see Toutanova, Klein, Manning, and Singer (2003)
- Maximum Entropy Models see Ratnaparkhi (1997)

While it is important that we choose a tool that is flexible, multi-lingual and easy to use it is also important that we achieve high rates of precision and recall with it. Low recall would lead to overheads marking up the outstanding tokens and low precision would filter into the rest of the pipeline lowering the maximum precision available for other annotation tasks.

Most of the taggers use the Penn-Treebank tagset for English, as described in Marcus, Santorini, and Marcinkiewicz (1993). Taggers for German typically use the Stuttgart-Tübingen tagset (STTS) (Schiller, Teufel, Stöckert, & Thielen, 1995). I am not sure what tagsets are standard for other languages.

Some potential points of interest:

- The Stuttgart Tree Tagger has parameter files for English, German and French already
- The GENIA tagger has been trained on biomedical data (Medline extracts)
- TnT is performant, highly configurable and widely used
- RASP has been widely used and covers the whole linguistic annotation pipeline including parsing
- The Stanford POS tagger can be combined with the Stanford Parser into a simple Java linguistic annotation process

Name	Availability	Format	Tagset	Strategy	Lang.	Platform	Notes
RASP[9]	Download	SGML	C7 variant	Hybrid	English	Unix C and Common Lisp	See below on Parsers

[9] http://www.informatics.susx.ac.uk/research/nlp/rasp/

Name	Availability	Format	Tagset	Strategy	Lang.	Platform	Notes
LT POS[10]	Free Download	ASCII, SGML, XML	Configurable, sample has Penn	HMM	English	Unix command line tool	Configurable tagset although it looks like Penn comes as standard
CLAWS[11]	£750+VAT site licence	SGML	C7	Hybrid rule/HMM	English	SunOS4.x	
QTAG[12]	Site licence from Phrasys	Any		Some stochastic algorithm	All	Java (standalone or embedded)	Theoretically language independent – but we would have to devise resource files for languages other than English. Note that we need a tagged training corpus to build the resource files.
NLTK[13]	Free Download	Any	Any	Rule-based, Viterbi and HMM	All (?)	Python libraries	NLTK includes libraries for implementing Brill's transformational rule-based tagger as well as Viterbi and HMM algorithms.
Tree Tagger[14]	Free Download	ASCII input, CSV output	Penn-Treebank and other bespoke for other languages	Stochastic with Decision Trees	All	Perl, Python	Multi-lingual with a manually tagged training corpus, uses a decision tree algorithm see Schmidt (1997). Parameter files for the following languages are available: English, German, Italian, Spanish, Bulgarian, French.
GENIA[15]	Free Download	ASCII input, CSV output plus IOB2 chunk tags	Penn-Treebank (I think)	Cyclic Dependency Network	English	C (requires GNU gcc compiler)	Trained on the GENIA corpus (extracts from Medline) so supposed to deliver high precision for biomedical texts, see Tsuruoka et al. (2005)

[10] http://www.ltg.ed.ac.uk/software/pos
[11] http://www.comp.lancs.ac.uk/ucrel/claws/
[12] http://www.english.bham.ac.uk/staff/omason/software/qtag.html
[13] http://nltk.sourceforge.net/
[14] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
[15] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

Name	Availability	Format	Tagset	Strategy	Lang.	Platform	Notes
ACOPOST (formerly ICOPOST)[16]	Free Download	ASCII input, CSV output	Penn-Treebank	4 flavours: MaxEnt, HMM trigram, Brill, Example-Based	English	C on Unix	
Stanford POS Tagger[17]	Free Download (GPL)	ASCII input	Penn-Treebank	Log-linear – I think using a cyclic dependency network	All	Java 1.5+	See Toutanova and Manning (2000) for details. Requires training, has been trained on English.
Xerox Tagger[18]	Commercial licence	ASCII input	Any	HMM	All	C++	The Xerox POS tagger has now been sold to Temis and is available only as part of XeLDA on commercial terms.
TnT[19]	Non-commercial, non-profit license only	ASCII input, configurable output	Any	HMM using Viterbi algorithm	All	C on Unix	Performant, and can be trained on domain, language, corpus, tagset (Brants, 2000). This tagger is a popular choice
MTag (TATOO)[20]	Free Download	Expects output from MULTEXT segmenter as input	Any	HMM using Viterbi algorithm	All	Perl and Tcl/Tk on Unix	I think that MTag has now been subsumed into the TATOO tagger.
ATILF[21]	Unknown	Unknown	Unknown	Unknown, presumably rule-based	French	Unknown	A French version of the Brill Tagger is available from ATILF.
Brill Tagger[22]	Free Download	ASCII input	Penn-Treebank	Rule-based	English	C on Unix	Brill's own implementation in C of his widely used tagger. I think his version is for English and trained on the Wall St Journal.

[16] http://acopost.sourceforge.net/
[17] http://nlp.stanford.edu/software/tagger.shtml
[18] http://www.temis-group.com/fichiers/t_downloads/file_53_XeLDA_(en).pdf
[19] http://www.coli.uni-saarland.de/~thorsten/tnt/
[20] http://www.issco.unige.ch/staff/robert/tatoo/tatoo.html
[21] http://www.atilf.fr/
[22] http://research.microsoft.com/~brill/

Morphological Analysers

Some POS taggers include some morphological analysis (i.e. semantic analysis of individual words), but not all. If necessary we could use a morphological analyser to enhance the annotation with more detailed morphological information. This may be more of an issue in inflected languages than in English.

Name	Availability	Format	Platform	Notes
Morphix[23]	Download	any	Allegro Lisp	Morphological analyser and generator for German
MMorph[24]	Download	MULTEXT	Tcl/Tk	Multext morphological analyser
Flemm[25]	Download	Brill and TreeTagger tagsets only	Perl5 on any OS	Lemmatises output from the Brill tagger, also can be used to check output from the TreeTagger. See Namer (2000)
Derif[26]	Unknown	Unknown	Perl	Analyses derivational forms in French and can propose morphologically related word groups. See Namer (2002a)

Shallow Parsers and Chunkers

These tools mark up meaningful segments or chunks within the text, such as noun phrases for example, using a shallow parse of the text.

Shallow parsers will only work on text that has been annotated by a POS tagger. Where a shallow parser is associated with a particular POS tagging tool I have marked this relationship in the table below as we should use the two tools together.

I have included a few common tools here, I am not sure if this is an annotation that we will require as part of this work package.

Name	Availability	POS Tagger	Platform	Lang.s	Notes
GENIA Tagger[27]	Download	GENIA	C (GNU cc)	English	Shallow parsing is available as part of the GENIA tagger.
LT CHUNK[28]	Download	LT POS	Unix command line tool	English	Annotates output from LT POS with noun and verb phrase markers
Chunkie[29]	still available?	TnT	Unknown	All	Integrated with TnT and MMorph as part of ShProT by Vintar et al. (2002)

[23] http://www.dfki.de/~neumann/morphix/morphix.html
[24] http://www.issco.unige.ch/projects/MULTEXT.html
[25] http://www.univ-nancy2.fr/pers/namer/Telecharger_Flemm.html
[26] http://www.univ-nancy2.fr/pers/namer/
[27] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/
[28] http://www.ltg.ed.ac.uk/software/chunk/index.html
[29] http://www.coli.uni-saarland.de/~skut/chunker/

Name	Availability	POS Tagger	Platform	Lang.s	Notes
Segmenter[30]	Download		Perl	English	Segments text into topical chunks, see Kan et al. (1998)
Lexter[31]	Unknown			French, and others?	A term extractor that extracts candidate noun phrase terms from a corpus, see Bourigault (1995)
Acabit[32]	Download	Brill/Atilf and Celex/Flemm	Perl	English and French	Produces multi-word term candidates from a linguistically annotated corpus, see Daille (2003)
YamCha[33]	Download (GPL)	YamCha	C/C++	English	Chunking is performed as part of POS tagging, marked in additional IOB2 column

Parsers

I have not been able to locate any parsers that are specific to the medical domain, or that have been used specifically for projects in the medical domain, although it would make sense to use such a parser if we can.

A broad overview of a few parsers that are readily available includes:

Name	Availability	Platform	Lang.s	Notes
RASP[34]	Download	C / Common Lisp	English	RASP performs tokenisation, POS tagging, stemming and parsing. It uses a variant of the C7 CLAWS tagset.
Stanford Parser[35]	Download	Java 1.5+	All	Languages other than English and German may require a lot of additional configuration work. Works with the Stanford POS tagger.
Charniak Parser[36]	Download	C	English	A maximum entropy parser for English
Link Grammar Parser[37]	Download	C	English	Link grammar parser of English
PET[38]	Download	C++	All	A HPSG parser built at DFKI, can use LKB grammars

[30] http://www1.cs.columbia.edu/~min/research/segmenter/
[31] http://www-sira.montaigne.u-bordeaux.fr/IE10_FIN/bourigault/bourigault.htm
[32] http://www.sciences.univ-nantes.fr/info/perso/permanents/daille/acabit_en.html
[33] http://chasen.org/~taku/software/yamcha/
[34] http://www.informatics.susx.ac.uk/research/nlp/rasp/
[35] http://nlp.stanford.edu/software/lex-parser.shtml
[36] ftp://ftp.cs.brown.edu/pub/nlparser/
[37] http://www.link.cs.cmu.edu/link/
[38] http://wiki.delph-in.net/moin/PetTop

Name	Availability	Platform	Lang.s	Notes
ICETree[39]	£99 + VAT	Windows	All	Software top build and manipulate syntactic trees for large corpora.
NLTK[40]	Download	Python	All	Python class libraries for writing parsers
XTAG[41]	Download	C on Unix	English	I don't think that XTAG has been updated since 2001

8.5.3 Semantic Annotation

Once the corpus has had linguistic annotations applied to it, we can add semantic annotations of various sorts to mark up topic and genre, links to entities in some external lexicon or ontology, dialogue components or linguistic properties (such as coreference).

There are some tools around that can be downloaded, although they are usually quite specific to the problem domain of the institution that built the tool and will not necessarily port straightforwardly to our project. I have also included brief discussion of annotation projects that might be of interest to us.

8.5.4 Discourse-Level Annotation

There are a few tools for annotating discourse, RSTTool and Marcu's extension of it allow user annotations via a tool to plain text. SPADE and WordFreak have the option of running automated annotation. SPADE is tied to Charniak's parser, which I think only works with English, so this would tie SPADE to English. WordFreak is the tool being used to build the Penn Discourse Treebank. It can be used as a manual tool or automatically and has a plug-in architecture (like GATE) that allows POS taggers, parsers and other annotation tools to be integrated with it, see Carlson and Marcu (2001).

Name	Availability	Platform	Lang.s	Notes
RSTTool[42]	Download	Tcl/Tk	All	An interactive tool that annotates plain text with segments and allows the user to mark relationships between them
Marcu's RST Annotation Tool[43]	Download	Tcl/Tk	All	Marcu's extension to RSTTool with additional features such as logging and undo, see Marcu et al. (1999)
SPADE[44]	Download	Perl	English	An automatic discourse chunker and shallow annotator – input is assumed to have been parsed by Charniak's parser (mentioned above)

[39] http://www.ucl.ac.uk/english-usage/ice/annotate.htm
[40] http://nltk.sourceforge.net/
[41] http://www.cis.upenn.edu/~xtag/swrelease.html
[42] http://www.wagsoft.com/RSTTool/
[43] http://www.isi.edu/~marcu/
[44] http://www.isi.edu/~marcu/

Name	Availability	Platform	Lang.s	Notes
WordFreak[45]	Download	Java	All	Manual and automated tool to generate XML stand-off annotation for discourse – used by the Penn Discourse Treebank[46]. Note that the input must be POS tagged.

Muchmore Project

This is a project from Saarbrücken on Cross-lingual Information Retrieval in the medical domain.

In Vintar et al. (2002), a process for annotating a corpus of medical domain information (English-German parallel medical abstracts from SpringerLink) into UMLS is described in detail. Linguistic annotation is performed using TnT, MMorph and Chunkie (integrated into a single tool called ShProT), terms are then extracted into UMLS and links into specialist lexicon, Metathesaurus and a semantic network are annotated using unique keys.

The semantic annotation is performed using a tool written by DFKI, it analyses the output of the linguistic annotator and looks for unigrams, bigrams, trigrams that match terms within UMLS (based on stems).

Hospitexte Project

Charlet et al. (1998) describe a project to build electronic medical records using structured documents. The paper describes the process of semantic annotation of clinical information such as patient name, age, weight, drug names, diagnoses etc. using an SGML annotation schema described by a published DTD (called MedDoc).

The system does not link entities within the documents to some external data source (such as a database or ontology) but is designed to synthesise patient record documents that have been produced in standard ways (for example plain text documents) into structured medical record documents that can then be presented as HTML to the end user.

XDMLTool (eXtensible Dialogue Markup Language Tool)

This tool is part of the AMITIES project at Sheffield[47] and is detailed in Hardy et al. (2002) and Hardy et al. (2003).

The tool is for manual annotation of dialogues in French and English and is used in the AMITIES project to add semantic annotations in XML to dialogue transcriptions from European call centres. The tool is written in Java and presents each turn of the dialogue (received as input in plain text) to the annotator. Output is in XML according to a schema designed by Sheffield.

The output does not contain any other annotation information, just the plain text of the dialogue turns attached to an XML representation of the information captured by the annotator that feeds into a taxonomy based on the DAMSL schema (Allen & Core, 1997).

XDMLTool can be downloaded for research purposes and I believe that there is work ongoing at Sheffield to integrate this tool into GATE.

[45] `http://sourceforge.net/projects/wordfreak`
[46] `http://www.seas.upenn.edu/~pdtb/`
[47] http://www.dcs.shef.ac.uk/nlp/amities/

PALinkA (Perspicuous and Adjustable Links Annotator)

This tool allows the user to add referential link annotations to text, provided that the input is in XML format conformant with the tool.

It is written in Java and is available for download from .

For a detailed discussion of other architectures and tools for anaphora resolution see Mitkov (1999).

ACASD Semantic Tagging System

This tool, developed at UCREL, is described in Wilson and Rayson (1993). It takes output from the CLAWS tagging system and assigns semantic tags to each lexical item, using a set of semantic classifications specific to the project (a semantic classifier for annotating market research transcriptions) and based on the USAS classification system developed by UCREL. Since the system is from 1993 it is presumably dependent on an earlier version of CLAWS (perhaps C5?) than the current release.

I don't know if this tool is available for public or academic use, although it may be released as part of USAS[48].

See also Rayson and Wilson (1996) for a description of ACAMRIT including SEMTAG, I think this is a further development of the semantic tagging system at UCREL.

See also Pala and Smrz (2004) for a discussion of using the Top ontology to tag semantic roles in Czech verb valency frames.

Named Entity Recognition

- A few papers at the shared task on language independent named entity recognition from CoNLL-2002: `http://www.cnts.ua.ac.be/conll2002/ner/`
- Workshop "Beyond named entity recognition – semantic labelling for NLP tasks" at LREC 2004: `http://ai-nlp.info.uniroma2.it/ws_lrec04/`
- If we want to mine our own data then some of these corpora and text collections for biomedical NLP might be useful: `http://compbio.uchsc.edu/corpora/obtaining.shtml`

I think that this task extends beyond the remit of annotating the corpus, so I haven't looked at any systems, papers or tools in any detail.

8.6 Other tools of potential interest

8.6.1 MontyLingua

Includes a tokenizer, POS tagger, lemmatiser and chunker: `http://web.media.mit.edu/~hugo/montylingua/index.html`

Free download, Java/Python. Looks useful in terms of what it does, but I haven't seen it before so I have no idea how robust or precise it is.

[48] `http://www.comp.lancs.ac.uk/ucrel/usas/`

8.6.2 Unitex

A corpus text processing system written in Java and C++ that enables the production of lexica from corpus resources (`http://www-igm.univ-mlv.fr/~unitex/`). This may be of use to us for extracting terms from an annotated corpus. Unitex works with any language, and comes with a lot of dictionary and grammar resources in many common languages.

8.7 IDEs and Workbenches

Rather than building our own pipeline of tools to perform orthographic and linguistic analysis and annotation we could use a workbench or IDE that integrates a stack of tools for us. Some examples of such workbench or IDE toolsets include:

8.7.1 IMS Corpus Workbench (CWB)

`http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/`

8.7.2 Annotate

`http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/annotate.html`

Annotate is an Annotation tool developed against NEGRA, a treebank corpus of German newspaper text. It is written in C, requires MySQL and some Unix tools. It incorporates processes such as POS tagging (using TnT) and morphological analysis (using TigerMorph).

8.7.3 Alembic Workbench

`http://www.mitre.org/tech/alembic-workbench/`

The Alembic Workbench provides an integrated pipeline for the annotation of text on a Unix platform.

It includes a segmenter, a POS tagger, and a chunker. Text that has been annotated by all of these components can be processed semantically, for example by a inference component as described in Aberdeen et al. (1995).

8.7.4 CLaRK

`http://www.bultreebank.org/clark/`

The CLaRK system is a corpus development system developed for use on the BulTreeBank project and implemented in Java, see Simov et al. (2001) for details. The system provides tools and a GUI to enable the user to add orthographic and some linguistic annotations to a corpus using an XML schema of their own design.

8.7.5 GATE

`http://gate.ac.uk/`

GATE is a leading SDK and IDE for language processing tasks, including the collection and annotation of corpora, see Cunningham, Maynard, Bontcheva, and Tablan (2002). GATE aims to serve as an integration platform for language resources (such as text collections, lexicons, ontologies and corpora), processing resources (such as taggers, parsers or generators) and

visual resources (plug-ins to the GATE GUI to facilitate the visualisation of applications and processes).

GATE provides an annotation format (based on TIPSTER) that aims to capture all of the annotation information for language resources and provides built-in transformation to/from that format from standard document types (such as RTF or HTML) and standard schemas.

GATE also incorporates the following built-in processing resources: a tokenizer, a sentence splitter, a POS tagger, a gazetteer, a semantic tagger, an orthomatcher and a coreference solver. It is also possible to plug-in other tools provided that they are written in Java and a CREOLE interface can be devised. GATE also contains tools for performing evaluations and benchmarking on the tools within the IDE.

8.7.6 RASP

`http://www.informatics.susx.ac.uk/research/nlp/rasp/`

RASP is a pipeline of C and Common Lisp tools for annotating raw text to be used in corpora, as described in Briscoe and Carroll (2002). It is not an integrated workbench or IDE but I have included it here because it covers the whole linguistic annotation pipeline of tokenization, POS tagging, stemming and parsing (albeit only for English I believe).

8.7.7 Stanford Parser

`http://nlp.stanford.edu/software/lex-parser.shtml`

Like RASP, this system covers the whole linguistic annotation pipeline. It is written in Java, is recent and up-to-date and has been used to annotate a variety of corpora in German and English and provides the means to configure the system for other languages.

8.8 Survey of existing corpus annotation tools (Sweden)

A number of annotation tools for Swedish biomedical corpora have been developed or adapted to the sublanguage (Kokkinakis, 2006b). This set of tools includes:

- A Swedish (Brill-based) part-of-speech tagger, adapted to the medical vocabulary (and a heuristic lemmatiser based on the output of the part-of-speech tagger)
- Domain independent Named-entity recognition (persons, places, organisations, time and measure expressions)
- A Swedish MeSH tagger (level-1 tags)
- A Swedish NP-chunker, using finite state technology

9 Survey of Corpora of Patient Information

This survey points to a variety of resources, which constitute or might constitute usable patient information corpora. These are divided into corpora that have been previously used in research, and maybe even annotated, but that in many cases are not accessible, and corpora that have not been used or annotated, but that are easily accessible. As it is immediately apparent, it proved very difficult to find or even identify corpora that have been previously used and annotated and that could now be used for the purposes of the project. This is possibly due to the sometime sensitive nature of the information contained in corpora and to the fact that creating and annotating corpora implies high costs, all of which may make research institutions unwilling to share them. On the contrary, there seems to be a wealth of resources available on the World Wide Web, such as health sites and forums, which could be used to tailor new corpora for the purposes of the project. In this document, of those resources, only a few representative ones are listed in the following categories: corpora of medical information gathered by experts for laymen; corpora of medical information gathered by experts for other experts; corpora of medical information shared by laymen with other laymen – distinguishing corpora of medical information shared by laymen who have developed an expertise.

The findings mentioned above suggest that the Internet, given the numerous services that it offers and the wealth of information that it provides, be used as a source of corpora.

9.1 Previously used and/or annotated corpora

Below is a list of references to corpora that have been used, and sometime annotated, in previous research. In some cases, we know which ones they are and have at least some information about them, so we list them as *identified corpora*. In other cases, we only indirectly know – through literature – that they exist and have been used, so we refer to them as *unidentified corpora*.

9.1.1 Identified corpora

These are corpora that have been used in previous research and of which we have at least a basic description. However, so far, only the first two in the list are accessible to us and the third could be accessible if we were to pay for it.

The Patient Information Leaflet (PIL) Corpus

Description:

> The PIL corpus, developed as part of the ICONOCLAST and PILLS projects, consists of 471 documents, giving instructions to patients about their medication. The corpus, which is SGML annotated, has been used also in other projects between 2001 and 2004 (see references below).

Source: http://mcs.open.ac.uk/nlg/old_projects/pills/

References:

- ICONOCLAST: http://mcs.open.ac.uk/nlg/old_projects/iconoclast/
- Bouayad-Agha, Scott, & Power, 2000
- Bouayad-Agha, 2000

The British National Corpus (BNC)

Description:

The BNC has 119 files of spoken medical consultations (transcribed doctor-patient dialogues) totalling some 85,620 words. These files may be found by searching for 'medical consultation' in the titles (e.g. in the BNC bibliography)[1].

Source: http://www.natcorp.ox.ac.uk/

References:

- Somers & Lovel, 2003

THIN Patient Record Database

Description:

EPIC is an organisation that facilitates access to databases of patient records from THIN (The Health Improvement Network) patient record data from the UK and some European countries. The THIN database contains:

- primary care records of patients' GP consultations, diagnoses and prescriptions and some letters relating to hospital visits, tests, etc.
- over 5 million patient records in total and growing (in 2002 THIN covered 3.5 % of the UK population; in 2006 it covers 4 %)
- around 3 million active patient records (where "active" means the patient is still currently with the practice)
- records with time spans of up to 15 years
- old paper records converted and entered by hand
- all records are anonymised (names such as the patient name and surgery name are changed; locations and dates are changed) this is done by the GP practice and also by a manual anonymisation check at EPIC

THIN data is collected from over 300 GP practices in the UK, most data is entered using the VISION data system and collected by EPIC automatically. GPs are paid for the data, but payment depends on accurate and complete recording. Regarding ethics, THIN data

[1] information sent by David Lee <http://devoted.to/corpora> to the Corpora List on 2 Feb 2000, source: http://listserv.linguistlist.org/cgi-bin/wa?A2=ind0002&L=CORPORA&P=R681&I=-3

has MREC (Multi-centre Research Ethics Committee) approval and all research proposals on THIN data must have MREC type 3 approval (EPIC help with this). There is also approval for case validations where EPIC act as an intermediary to check data in individual records with the relevant GP practice and for patient questionnaires (patients may be approached). New in 2006 will be medicine "pack sizes" and socio-economic information. Pack size information will be, for instance, the actual number of tablets in a pack prescribed to a patient. Epidemiologists will be able to use this information, together with prescribed dosages, to estimate actual consumption of medicines. Socio-economic indicators will include a "deprivation score", ethnicity scores and environmental factors. These are based on the area in which the patient lives (although actual postcodes and locations are suppressed by anonymisation) this narrows down to groups of roughly 300 people. Thus socio-economic information can be used for "personalisation", but only in a probabilistic manner – e.g. it is known to what extent the general area in which the patient lives is deprived, but the individual circumstances of the patient are not known. In more detail, the socio-economic scores are based on 2001 UK population census data, grouped mostly on 5-point scales, as follows:

- Townsend deprivation index
- Population density grouping
- Ethnographic grouping
- Environmental grouping (based on air pollution, etc.)
- Incidence of long-term illness grouping

At present, THIN contains scores for only 20 practices, EPIC intend to add more during 2006. THIN contains some inconsistencies in the data (e.g. codes assigned to smokers may be out of date – currently 85 % of THIN patients have smoking codes, indicating that they are smokers or have been in the past, which far exceeds the percentage in the UK population as a whole. There can be problems making inferences based on data in fields of the database, e.g. a study on Deep Vein Thrombosis (DVT) excluded all patients who had ever taken the drug Warfarin, but it wasn't clear whether this was necessary, or not. Also discussed was whether it would be safe to infer that patients who had been prescribed anti-malarial drugs would have undertaken long-haul flights and consequently might be more likely to suffer from DVT. Such problems surely must occur with most database information and are not unique to THIN data.

Source: `www.epic-uk.org`

The CLAN based corpus of doctor patient interactions

Description:

This corpus gathers doctor patient-interactions, which were recorded using the cross platform database freeware CLAN. These records are part of the MOVIN (Micro-analysis of Verbal/Visual INteraction: Danish network in Conversation Analysis) corpus and are used to train medical students on doctor-patient communication, at the School of Health of the University of Southern Denmark.

References:

· Petersen & Wagner, 2005

The cancer treatment reports corpus of the Centre François Baclesse

Description:

This collects text of medical reports on cancer treatment from the Centre François Baclesse, the anti-cancer centre of the region of Lower Normandy. The corpus represents at least all the records from 1992 to 1994, its size totals approximately 180,000 words, with about 10,000 different words. All the reports are free texts with headers identifying the physician, patient, date, etc. They are medical descriptions variable in length and style.

References:

· Guedj & Nugues, 1994

The Medline and MedlinePlus corpora

Description:

This is a medical literature database collecting information from the National Library of Medicine, both for health professionals and consumers. It gathers information from the National Institutes of Health and other trusted sources on over 700 diseases and conditions. It also provides lists of hospitals and physicians, a medical encyclopedia and dictionary, information on drugs and links to clinical trials. Subsets of Medline and MedlinePlus resources have been used in different research projects (see references below)

References:

· Dykes, Curries, & Bakken, 2004

· Ferguson et al., 2002

· Liu & Friedman, 2003

9.1.2 Unidentified corpora

This section lists a number of references to publications about research that has made use of patient information corpora. Although at the moment we know little or nothing about those corpora, and have so far been unable to contact the researchers who used them, this list could constitute a resource in future.

· Cawsey et al., 1997

· Marttala, 1995[2]

[2]The corpus consists of authentic video-taped doctor-patient conversations, complemented by interviews and viewing sessions with comments by the participants.

- PERSIVAL corpus (2001) [3]
- Thomas & Wilson, 1996
- G. Thompson, 1999
- Wynn, 1999

9.2 Non-annotated corpora

This section lists a number of World Wide Web resources. With the exception of one (Medline), we are not aware of them being at all annotated and used within previous research. Here they are grouped in the following categories: corpora of medical information gathered by experts for laymen; corpora of medical information gathered by experts for other experts; corpora of medical information shared by novice laymen or laymen who have developed an expertise. Resources that fall into more than one category are repeatedly cited as appropriate.

9.2.1 From experts to laymen

These are medical websites or portals about specific or general health conditions, which are managed by authoritative institutions and dedicated to non-expert public and patients.

- Multiple Sclerosis Society website[4]:
  `http://www.mssociety.org.uk/`
- Multiple Sclerosis Federation website:
  `http://www.msif.org/`
- The Leukemia and Lymphoma Society website[5]:
  `http://www.leukemia.org/hm_lls`
- National Cancer Institute – Leukemia website:
  `http://www.cancer.gov/cancertopics/types/leukemia`
- MedicineNet website:
  `http://www.medicinenet.com/leukemia/article.htm`
- Breastcancer.org website[6]:
  `http://www.breastcancer.org/`
- Breast Cancer Care website:
  `http://www.breastcancercare.org.uk/`
- National Breast Cancer Foundation website:
  `http://www.nationalbreastcancer.org/`
- Cancer Backup website:
  `http://www.cancerbackup.org.uk/`

[3] An 85 million word corpus of medical journal articles that has been automatically annotated with medical terms. See `http://persival.cs.columbia.edu/reports/00-01.pdf`
[4] Google search for Multiple Sclerosis returned 36,600,000 hits
[5] Google search for for Leukemia returned 40,500,000 hits
[6] Google search for Breast Cancer returned 125,000,000 hits

- Maggie's Centres website:
  http://www.maggiescentres.org/
- Patient UK website:
  http://www.patient.co.uk/showdoc/26738895/
- Cancer Research UK website:
  http://www.cancerhelp.org.uk/
- The DIPEx.org website:
  http://www.dipex.org/
- Austrian, British, German, and Swedish Netdoctor websites:
  http://www.netdoktor.at/
  http://www.netdoctor.co.uk/
  http://www.netdoktor.de/
  http://www.netdoktor.se/ or http://netdoktor.passagen.se/

This category also includes general health and fitness online magazines, of which two examples are provided here.

- Healthy People 2010 website:
  http://www.healthypeople.gov/
- BBC Health website:
  http://www.bbc.co.uk/health/

A part of the Swedish Medlex Corpus incorporates a number of subcorpora of patient information sampled from various sites (e.g. NetDoktor.se), the size of the corpus is roughly one million tokens (Kokkinakis, 2006a).

The Swedish Carelink site has a very short description in English (more information in Swedish) about "National Patient Summary" which is intended both for health care professionals and patients (different views). See:

- http://www.carelink.se/pages/newsbill.asp?VersionID=1&Pages=1,124,321
- http://www.carelink.se/pages/newsbill.asp?VersionID=1&Pages=1,248,278,312

9.2.2 From expert to expert

These are websites managed by medical experts and expert discussion groups dedicated to medical experts.

- Doctor's Guide Global Edition website:
  http://www.pslgroup.com/dg/2e48e.htm
  http://www.rxlist.com/cgi/generic/optivar_pi.htm
- Up To Date Patient Information website:
  http://patients.uptodate.com/index.asp
- The medical literature database:
  http://www.medline.de/

- Pharmaceutical company Merck (especially oncology and cardiometabolic): `http://www.merck.de/`
- Usenet newsgroup `news:sci.med` and sub-groups such as `news:sci.med.diseases.cancer` (in English)
- Usenet newsgroup `news:de.sci.medizin` and subgroups (in German)

9.2.3 From novice laymen to novice laymen

These are resources for non-expert public and patients. Mostly, they are discussion groups in which people exchange experiences and information, but they can also be authoritative websites hosting forums or presenting interviews of patients.

- The DIPEx.org website: `http://www.dipex.org/`
- Usenet sub-groups of alt.support.*, e.g. `news:alt.support.cancer`[7]
- Usenet sub-groups of soc.support.*
- Usenet newsgroup `news:uk.people.support` and sub-groups
- Usenet newsgroup `news:misc.health` and sub-groups
- Usenet newsgroup `news:misc.kids.health` and `news:misc.kids.pregnancy`
- Usenet newsgroup `news:alt.health`

9.2.4 From expert laymen to novice laymen

These are resources provided by non-expert patients who have developed an expertise about their own condition and have set up websites to share their knowledge with other patients. Sometime, they also contribute to discussion groups by producing FAQs documentation.

- All About Multiple Sclerosis website: `http://www.mult-sclerosis.org/`
- Usenet sub-groups of alt.support.*, e.g. `news:alt.support.cancer`
- Usenet sub-groups of soc.support.*
- Usenet newsgroup `news:uk.people.support` and sub-groups
- Usenet newsgroup `news:misc.health` and sub-groups
- Usenet newsgroup `news:misc.kids.health` and `news:misc.kids.pregnancy`
- Usenet newsgroup `news:alt.health`

[7] You will find messages from patients with varying degrees of expertise and even medical professionals in these newsgroups. They will usually be in colloquial terms, so we have included them in section 9.2.3 as well as in section 9.2.4.

9.3 Survey of web portals addressing patient information needs (Sweden)

Bilingual terminological support for health care professionals in the first place and medically well informed consumers is provided by Karolinska Institutet, Stockhom, in form of the extended MeSH resource (the controlled vocabulary thesaurus of the NLM, U.S. National Library of Medicine), which has been supplemented by Swedish translations. In the present update, 22106 terms out of 23885 have been translated. <http://mesh.kib.ki.se/swemesh/swemesh.cfm>.

A list of 89 web sites dealing with Patient information services is provided by Karolinska Institutet at <http://www.mic.ki.se/Diseases/swefraga.html>. The range of subjects is very broad and covers medical legislative and social information, issues concerning health care provided by professionals and addresses to a number of patients' organisations. Interactive support of the type "Ask the doctor" services is provided by many sites listed in there.

The "Health Care Counselling" (Sjukvåardsråadgivningen) is one of the most widely used Swedish health sites <http://www.sjukvardsradgivningen.se/>. It is a non-commercial, public source site which provides both general health guidelines and interactive services. It has been many times awarded for its exemplary laymen oriented presentation, explanation and access to the data. The data there can be accessed either by a free text search, an alphabetic search or thematic search with 16 topics. The topics deal with diseases, symptoms and health troubles, injuries, dental care, medical inquiries, treatments, drugs and health wellness, to mention some. The text information is very well structured and the text segments and their subsections with informative titles and subtitles support navigation in the text. The readability level is adjusted to a layman, as the necessary medical terminology is introduced in a reader friendly way. The syntax is easy to follow, passive constructions are avoided and sentences have a transparent structure. The site aims in the first place at adult laymen, but there are also sections directed to children in form of cartoons and also guidelines to health care personal. For the informed patients who want to enter more deeply into the subject matters, there are links to relevant web sites dealing with particular subjects. The site visitors can look for answers in the bank of frequently asked questions. The FAQ bank contains over 2000 questions. The Sjukvåardsråadgivningen site promotes access to medical knowledge in a flexible and user-friendly way and thus in a very significant way contributes to the empowerment of patients. The services at Sjukvåårdsråadgivningen has recently been expanded with query services called "Fråaga sjukvåardsråådgivningen" (Ask Health Care Counselling), where professional medical staff (doctors and nurses) will answer written questions online. Questions are answered personally if answers cannot be found in the existing FAQ bank. If a question is similar to a previously asked question then an answer from the FAQ bank is provided to the user. There are plans to add an automatic interactive agent to this site where users would be able to get instant replies to their questions. The module is being evaluated at the moment.

A Swedish survey of web portals addressing patient information needs is provided in: "Health on the Internet. A survey of Swedish web sites 2002. English summary". Socialstyrelsen, The National Board of Health and Welfare. The document provides a survey of 35 Swedish web sites, which were selected according to the following criteria: whether, (i) the information on the web sites was written in Swedish; (ii) the web sites were oriented towards a wide public; (iii) they contained factual information about health and diseases available

without a password or company membership; (iv) they provided some form of interactivity (ask the doctor service or equivalent). Out of 35 sites nine were classed as non-commercial and 26 as commercial. The main purpose of the survey was to find out to what extent the health web sites comply with guidelines for quality grading as proposed by the EU Commission. Among the Commission's six quality criteria, the one that deals with Accessibility, namely search facilities, readability and lucidity, is of particular relevance for the issue of patient empowerment. However, in the report it is remarked that the assessment of accessibility has varied considerably among the individuals conducting the survey, as this judgement to a great extent depends on the person's reference frame, Internet experience and knowledge of the subject in question. None of the sites provided facilities for disabled persons. No web site had information in the most common immigrant languages, not even in easy Swedish for immigrants. Eight of the web sites provided information in other languages besides Swedish. Besides the quality matters examined in the survey, there were some other issues of relevance to patient empowerment which were inspected, namely patients rights, ask the doctor services, other interactivity in the form of discussion forums, tests and cartoon films for children and links from the web sites to some patient versions of national guidelines provided by the National Board of Health and Welfare. In the conclusions the authors of the report point to the need for further research on "how the medical health care information and services on the web are used, how this accessibility affects the contacts made by the general public with the medical health care services, and opportunities thus generated for participation and exerting influence as a patient."

10 Internet as Corpus

As we showed in chapter 9, it is still difficult to find annotated corpora for specific domains although corpus linguistics has been around for a while.

It comes as no surprise that using the Internet as a abundant source of textual information is an appealing approach to many who are in need of large text corpora.

A special issue of Computational Linguistics (Kilgarriff & Grefenstette, 2003) has appeared, workshops have been held, and very recently, a Special Interest Group of the ACL (SIGWAC) has formed – all under the title 'Web as corpus.'

In what follows, we take a fresh look at the variety of services that the Internet offers, and try to identify promising sources of information for our project. At the same time, we make a case for renaming this research effort to 'Internet as corpus' because as we will show, the Web is just one linguistically interesting part of the Internet.

There are about 500 transport protocols that correspond to the services or applications which comprise the Internet.[1] One such service is the Word Wide Web. Other services we will look at are email, the Usenet (newsgroups), and IRC (Internet Relay Chat). After characterising each service, we will assess the chances for finding patient information texts within this service.

10.1 World Wide Web

Characteristics

- transport protocol: HTTP (RFC 1945[2])
- Uniform Resource Locators: URL (RFC 1738[3])
- Hypertext Markup Language: HTML (W3C standards[4])
- size: billions of pages

The Word Wide Web (or WWW, or Web for short) has been used as a source for linguistic corpora, and re-usable technologies have been developed. They range from crawlers (aka spiders, harvesters, or web bots) that can retrieve web sites recursively[5] to tool chests for the Web as corpus[6].

[1] Wikipedia has more details: `http://en.wikipedia.org/wiki/Internet_protocol_suite`
[2] published in May 1996, see `http://www.ietf.org/rfc/rfc1945.txt`
[3] published in December 1994, see `://www.ietf.org/rfc/rfc1738.txt`
[4] early drafts published in 1993, see `http://www.w3.org/`
[5] A reliable and extremely scalable Web crawler is Heritrix which was developed for the Internet Archive (`http://crawler.archive.org/`).
[6] See e.g. the WaCky project (`http://wacky.sslmit.unibo.it/`) of Marco Baroni and others.

Examples

Although we listed several examples of promising websites in the previous section, two website shall be considered more closely at this point.

Netdoctor is interesting not only because it is a popular source of information for patients but also because it is available in several languages: Danish, German (both standard and Austrian), Spanish, Swedish, and English. However, most articles do not seem to be translations. They are usually written by local physicians. Even in the Austrian `netdoktor.at`, articles are written by Austrian doctors although articles from the German `netdoktor.de` would be perfectly intelligible to Austrian users. It seems that patient information depends much on national peculiarities. This entails that one can hardly speak of a parallel corpus in this case. Though the same topics are treated and the overall content is likely to be similar, syntactic alignment will not be possible, and semantic alignment will probably be too difficult.

Wikipedia is also available in several languages. While some articles start their life as translations from other Wikipedias, they soon develop a life of their own as they are continually revised by local users. So although Wikipedia is heavily multilingual, it constitutes a parallel corpus only in the widest possible interpretation of the term.

But how much patient and health information does it contain? An exact answer can only be given after we have downloaded the database dumps to a local server for closer inspection. Until then, the category system can provide some clues.

In the following, we will only consider the English language Wikipedia (`en.wikipedia.org`) although other Wikipedias such as the German language Wikipedia (`de.wikipedia.org`) are also very large and also have a complex category system.

Of the top level categories of the English Wikipedia[7], the *Health* category is most likely to match the domain we are interested in. There is also *Health Science* and *Medicine* but they are more likely to be targeted at researchers and professionals. *Health* has 25 immediate subcategories one of which is *Diseases*. *Diseases* and its subcategories hold a total of nearly 2000 articles. To find out the total number of articles in the *Health* category, we would need to run an SQL query on the database directly which will be possible after downloading it and installing a local Wikipedia database server.

Unfortunately, not every Wikipedia article has been assigned a category, so to identify more articles in the patient information or health domain, information extraction (IE) methods could be used.

Having database dumps freely available in the case of Wikipedia saves a lot of effort in creating a domain specific corpus. When downloading large parts of a website automatically, many issues must be dealt with that are beyond the scope of this paper.

Assessment

Using the Word Wide Web as an information source is an obvious choice. It should be born in mind though that like any corpus, it is only representative of itself, and that not everyone can access it. The Web still predominantly reflects the English-speaking, northern-hemisphere, rather well-off parts of the world.

[7] For an overview of the categories of the English Wikipedia, start here: `http://en.wikipedia.org/wiki/Wikipedia:Browse`

It is also important to note (as Kilgarriff & Grefenstette, 2003, do) that for various reason, Web search engines do not cover all the Web. Some parts are excluded because the web site owners have requested so (by means of the `robots.txt` convention or otherwise), other parts are excluded because they are not linked from more 'central' parts of the known Web, or because their content is generated on user request only (as in the case of many database Web interfaces).

Other parts of the Internet can be linked by means of Uniform Resource Locators (URL) but are strictly speaking not part of the World Wide Web. They usually cannot be accessed using a Web browser. It is those parts of the Internet that we will look at next.

10.2 Usenet

Characteristics

- transport protocol: NNTP (RFC 977[8], RFC 2980[9])
- size: more than a billion messages (according to Google) / about 10^{12} tokens

Usenet was created in 1979 as an alternative to the Arpanet which later became the Internet. It has been used for various kinds of data exchange but the 'killer application' of the Usenet is newsgroups. Despite the name, the majority of groups are discussion groups. A few newsgroups are used as announcement groups, a few are moderated discussion groups. A newsgroup is created after completion of a community process which mainly exists to make sure there is a real demand. After that, everybody with access to an NNTP server ('news server') can post messages (which look very much like emails) to any group. This freedom of access eventually lead to the first sightings of unsolicited commercial messages which soon became known as spam. Spam is still a problem in the Usenet, and corpus collectors will encounter annoying commercial messages in completely unrelated newsgroups. Fortunately, most of them can be filtered automatically because each message has a unique message ID.

Meanwhile, far more than 15,000 newsgroups exist on every conceivable topic. Usenet users still tend to be technically and highly educated, and the English language dominates the Usenet even more than it dominates the Word Wide Web. Having started as an alternative to the early Internet, the Usenet soon became part of it. When the Word Wide Wide became popular, most of its users were not aware of the Usenet, and there was no interface between both services. This changed only in 1995 when Deja News began to offer a very large, searchable archive of Usenet messages dating back to 1981. After a few years, the Deja News service went off-line. In 2001, this database was acquired by Google.

Examples

Some examples of suitable newsgroups can be found in the previous section.

[8] published in February 1986, see `http://www.ietf.org/rfc/rfc977.txt`
[9] published in October 2000, see `http://www.ietf.org/rfc/rfc2980.txt`

Assessment

Usenet newsgroups have been used as a corpus at least since 1997 when Radev and McKeown (1997) used NNTP as well as HTTP retrieval[10] to access newswire texts for building a knowledge source for a natural language generation system. It is still used as a resource e.g. by Hoffmann (2006). They developed a software for downloading messages from many newsgroups from an NNTP server, and they also describe how to handle 'messy' data like that. They were able to download about a message per second.

We have identified some very intensely used health-related newsgroups for the English language and also some for German. Newsgroups in the `alt.*` hierarchy tend to be used by laymen while newsgroups in the `sci.*` hierarchy tend to be reserved for academic discourse. Using the `alt.*` newsgroups for a corpus of patient texts seems certainly worthwhile. A comparison with expert language in `sci.*` could be illuminating. This could help us induce transfer rules from expert to layman speak. It should be noted that the medical expertise of users of the `alt.*` newsgroups varies over the full range from medically uneducated person to experienced patient to doctor and scientist. The language used however is likely to be informal and intelligible to a non-expert.

For downloading older messages that are no longer kept on the usual NNTP servers, it might be necessary to access a newsgroup archive such as `http://groups.google.com`. This in turn seems to be difficult (according to Hoffmann, cited above) for reasons that need to be discovered.

10.3 Email

Characteristics

- transport protocol: SMTP (RFC 2822[11])
- size: unknown

Email is arguably a more popular Internet service than the Word Wide Web. It might also constitute a larger corpus. Emails that have been sent to one or few recipients are generally not publicly available owing to privacy issues[12]. This situation is unlikely to change unless convincing methods are developed for anonymising emails while preserving their linguistic and empirical value.

A more accessible source of emails as a corpus is mailing lists. In this case, each email is sent to a special email address which hands on the message to a mailing lists software. This software maintains a list of recipients and takes care of forwarding the emails to them, acting on behalf of the original sender. Examples of such software are the classic Majordomo and more recent free software projects like Mailman and Sympa. They often offer the possibility to archive all messages. Some of these archives are freely accessible as web sites or by special emails containing commands.

A directory and search engine for mailing lists that anyone can subscribe to can be found here: `http://tile.net`. In addition to that, some commercial mailing list providers also

[10] They also envisaged SMTP retrieval, i.e. using email as corpus.

[11] published in April 2001, see `http://www.ietf.org/rfc/rfc2822.txt`

[12] The only publicly available email dataset is probably the ENRON dataset (`http://www.cs.cmu.edu/~enron/`) prepared by SRI's CALO Project, containing about 500,000 messages.

offer directories and search facilities, e.g. `http://groups.yahoo.com`. Recently, Google has also started a mailing list service which is marketed under the same name as Google's Usenet archive, Google Groups.

Examples

Examples of health related mailing lists include those listed under `http://lists.topica.com/channels/health/`. Mailing list service provider Topica.com lists some health-related mailing lists under 'Health and Fitness'[13], but others under 'Society and Culture'[14]. Some of them have public archives, others require prior subscription.

Yahoo Groups has a wealth of health-related mailing lists in the top-level category 'Health and Wellness'[15]. A surprisingly large proportion of these mailing lists have public archive that can be accessed via WWW.

The mailing lists mentioned so far are mainly discussion lists where patients and health-aware people report their experiences and support each other.

Other mailing lists are used by medical students, researchers and practitioners[16]

Assessment

A large number of publicly accessible mailing list archives are available on the Web. Downloading some health-related ones would be an interesting if difficult enterprise.

Mailing list user span a wider range than Usenet users because email clients software is more wide-spread than Usenet client software. Apart from that, the sorts of text produced by mailing lists users are expected to be similar of those produced by Usenet users. That is because the process of creating both kinds of messages and the message formats are very similar. In fact, some email client software can be used to post messages to the Usenet.

Like in the case of the Usenet, spam is a huge problem, especially in Yahoo Groups. If these messages can be downloaded with intact header lines, there is a good chance that the usual spam-filtering tools will be able to filter the spam in a post-processing step.

10.4 Internet Relay Chat

Characteristics

- transport protocol: RFC 1459 [17]
- $>$ 3,500 networks hosting $>$ 500,000 channels

Internet Relay Chat (IRC[18]) relates to instant messaging much like the Usenet relates to email. Here we have instant messaging that goes to a group of people who have joined a certain channel (elsewhere known as chat room). The sorts of text produced are even more spontaneous than in the case of email and newsgroups. When using IRC clients, people rarely

[13] for Topica's 'Health and Fitness' category, see `http://lists.topica.com/dir/?cid=32`
[14] for Topica's 'Diseases' category under 'Society and Culture', see `http://lists.topica.com/dir/?cid=2188`
[15] see `http://health.dir.groups.yahoo.com/dir/Health___Wellness`
[16] for example, see this list of mailing lists for practitioners: `http://www.athealth.com/practitioner/imaillist/`
[17] `http://www.ietf.org/rfc/rfc1459.txt`
[18] see Wikipedia for a more complete introduction: `http://en.wikipedia.org/wiki/Internet_Relay_Chat`

correct typing errors; the software only allows for correcting the current line because previous lines have already been sent to the channel. A line roughly corresponds to a turn, but longer sentences or turns are usually sent in several lines (which may be interspersed with lines sent by other channel members).

There is an unknown number of IRC networks hosting an unknown number of channels. We have found an IRC search engine but we have no estimate of its coverage. `http://searchirc.com` is a search engine and directory of IRC channels. Currently it knows 7,751 servers constituting 3,480 IRC networks with more than a million users in more than 500,000 channels. On most networks, channels can easily be created by anyone at any time (if you 'join' a channel that does not exist yet, it is automatically created). Some channels are never used by more than one person and thus never get used. Others continuously attract dozens or even hundreds of users.

It is easy for participants to archive an IRC channel. However some channels do not approve of being archived, and those channels that are officially archived have their archives all in different places. This contrasts with mailing lists where the archive location can often be deduced from the location of the mailing list provider.

Werry (1996) (reprinted and carefully updated in Werry, 2004) has described this medium in more depth, and has examined it from a linguist's point of view.

Examples

Due to the non-hierarchical structure of the IRC networks, it is not easy to find relevant channels. Not all channels strictly stick to one topic; some are simply virtual meeting places for a more or less loosely defined group.

The IRC search engine and directory we found has a suitable but rather incomplete listing in a category called 'Health / Fitness / Medicine'[19]. It currently only contains 14 entries, many of which are not very popular.

The following is an incomplete list of some channels where patients are likely to meet. The text behind 'Topic' is a self-description of the channel which can change at any time.

- `irc://irc.freenode.net:6667/wrongplanet`
  Topic: #wrongplanet was founded in August 2004. This channel is for anyone with autism and aspergers.
- `irc://irc.servercentral.net:6667/autism`
  Topic: `http://www.autism-awareness.org.uk/` | `http://www.autism-society.org/`
- `irc://irc.starlink-irc.org:6667/lyme`
  Topic: Lyme Disease Support – Tuesdays and Fridays at 10 p.m. EDT / 7 p.m. PDT
- `irc://irc.purplesurge.com:6667/the-clinic`
  Topic: Got a medical problem or questions? Drop into #the-clinic to discuss it!
- `irc://irc.chatautism.com:6667/chatautism`
  Topic: autism asperger Age 16+ chat No ASL

The large share of autism related topics (such as Asperger) in health related IRC channels has not been fully explained. People on #wrongplanet pointed out that IRC (and especially

[19] see `http://searchirc.com/dir/Health-Fitness-Medicine`

the Freenode network) is popular among software developers, and Asperger frequently occurs in software developers. Of course this might be nothing more than hearsay.

Assessment

IRC is a very large source for a sort of spontaneous written text. For our immediate purpose however, IRC is not a useful resource because not enough people discuss health-related topics here – autism being a notable exception.

10.5 Conclusion

Our initial investigations have indicated that besides the World Wide Web, it also worth taking a closer look at newsgroups and mailing lists as potential sources for patient information corpora.

Index of Acronyms

AAC, 32
Acabit, 77
ACAMRIT, 80
ACASD, 80
ACL, 92
ACOPOST, 75
Alembic Workbench, 81
AMITIES, 79
AP-HP, 9
ASCII, 74, 75
ASL, 97
ATILF, 75
Atilf, 77

BabyTalk, 51, 53, 54, 58
BBC, 88
BEST, 60
BIOSYS, 27
BMA, 13–16
BNC, 71, 84
Brill, 75, 77
BulTreeBank, 81

C5, 80
C7, 73, 74, 77
CALO, 95
Celex, 41, 77
CES, 70
Charniak Parser, 77
Chunkie, 76, 79
CHV, 32
CINAHL, 27
CISMeF, 21, 29
CLAN, 85
CLaRK, 70, 81
CLAWS, 74, 77, 80
CLEF, 51, 53, 54, 58
CM, 29
CMA, 27
CREOLE, 82
CRM, 34
CSV, 74, 75
CT, 34
CWB, 81
Cyclic Dependency Network, 73

DAMSL, 79
Data Protection Act 1998, 13
Decision Trees, 73
Derif, 76
DFKI, 77, 79
DTD, 45, 50, 69, 79
DVT, 85

EBMT, 38
ED, 19
EDT, 97
EHR, 51, 54
ELIZA, 39
ENRON, 95
EPIC, 84, 85
EU, 26

FAQ, 39, 89
findaffix, 41
Flemm, 76, 77
FUF, 50, 52

GALEN, 33, 34, 37
GATE, 78, 79, 81, 82
GENIA, 69, 73, 74, 76
GNU, 74, 76
GP, 16, 19, 84, 85
GPL, 75, 77
GRAIL, 34, 37
GRASSIC, 44, 49, 51, 52, 54, 55, 58
GUI, 81, 82

HealthDoc, 44, 49, 51, 52, 54, 56
Heritrix, 92
Hidden Markov Model, 73
HIV, 20
HMM, 74, 75
Hospitexte, 69, 79
HPSG, 77
HTML, 70, 71, 79, 82, 92
HTTP, 92, 95

IC, 19
ICD-10, 31, 32
ICD-9, 22, 27, 29
ICETree, 78

ICF, 32
ICONOCLAST, 83, 84
ICOPOST, 75
ICU, 45
ID, 94
IDE, 8, 67, 68, 81, 82
IE, 93
IOB2, 74, 77
IR, 21, 28, 29
IRC, 92, 96–98
ISBN, 70
IT, 12

KA, 52, 53
KL-ONE, 36
KS, 28
KSH97-P, 32

Lancaster, 72
Lexter, 77
Link Grammar Parser, 77
LKB, 77
LMNL, 69
LOINC, 23, 30
Lovins, 72
LREC, 80
LT CHUNK, 76
LT POS, 74, 76
lttoken, 71

MAGIC, 50–54, 56, 58, 59
Mailman, 95
MailMerge, 52
Majordomo, 95
MaxEnt, 75
Maximum Entropy Models, 73
MDA, 50, 52–54, 56
MDS, 20, 21
MedDoc, 69, 79
Medical BeliefNet, 37
Medical FactNet, 37
Medical WordNet, 37, 60
Medical Wordnet, 37
MedLee, 35
Medlex, 88
Medline, 18, 35, 73, 74, 86, 87
MedlinePlus, 65, 86
MedSyndikate, 36
MedView, 50, 51, 53, 54, 57–59
MeSH, 21, 29, 32–34, 82, 90
MetaMap, 35
Metathesaurus, 7, 27–29, 33, 35, 79
MIGRAINE, 49, 52, 53, 55
MMorph, 76, 79
MontyLingua, 80
Morphix, 72, 76
MorTAL, 41
MOVIN, 85
MREC, 85
MTag, 75
MtSeg, 71
Muchmore, 79
MULTEXT, 75, 76
Multext, 71
MySQL, 81

N-gram, 73
NANDA, 30
NCSP, 32
NEGRA, 81
Netdoctor, 37, 93
NLG, 7, 43, 44, 46–48, 56, 58, 59, 62–64
NLM, 65
NLP, 33, 36, 37, 46, 60, 80
NLTK, 72, 74, 78
NNTP, 94, 95

ONIONS, 36
OR, 45
OS, 67, 76

PALinkA, 80
PDF, 71
PDT, 97
PENMAN, 44
PERSIVAL, 21, 46, 50–52, 54, 57, 59, 87
PET, 77
PIGLIT, 49, 51, 52, 54, 56
PIL, 83
PILLS, 46, 47, 50–54, 57, 83
PILS, 71
Porter, 72
POS, 8, 35, 67, 68, 71, 73, 75–82
PubMed, 27

QA, 37
QTAG, 74
QToken, 71

RASP, 73, 77, 82
Read, 29, 52
RFC, 92, 94–96
RST Annotation Tool, 78
RSTTool, 78
RTF, 70, 71, 82
Rule-based taggers, 73

SBU, 31
SDK, 81
Segmenter, 77
SEMTAG, 80
SGML, 73, 74, 79, 83
ShProT, 76, 79
SIGWAC, 92
SMTP, 95
SNOMED, 29, 30, 33, 34, 36, 37
Snowball, 72
SourceForge, 72
SPADE, 78
SPECIALIST, 7, 34
SQL, 93
SRI, 95
Standford POS tagger, 73
Stanford Parser, 77, 82
STOP, 47, 50–54, 57, 58, 63, 64
STTS, 73
Stuttgart Tree Tagger, 73
SumTime-Neonate, 51, 53, 54, 58
SUREGEN-2, 50, 53, 54, 58
SURGE, 50, 52
SUSANNE, 69
Sympa, 95

TAS, 51, 53, 54, 58
TATOO, 75
TEI, 70
TermColl, 32
TermExtractor, 36
THIN, 84, 85
TIGER, 69
TigerMorph, 81
TIPSTER, 82
TLFi, 41
TnT, 73, 75, 76, 79, 81
TraumAID, 49, 52, 54, 56
TraumGEN, 49, 52, 54, 56
TreeTagger, 76

UCREL, 80
UEA, 72
UEA-Lite, 72
UK, 7, 13, 18, 37, 52, 84, 85, 88
UMLS, 8, 21, 27–29, 33–36, 47, 52, 53, 65, 79
UMS, 7
Unitex, 81
URL, 70, 92, 94
US, 32
USA, 18
USAS, 80

VAT, 74, 78
VB, 72
Verbaction, 41
VISION, 84
Viterbi, 73

WaCky, 92
Wikipedia, 34, 92, 93
WordFreak, 78
WordNet, 37
WP27, 7, 32, 33, 35, 36, 67
WWW, 92, 96
WYSIWYM, 47

XCES, 70
XDML, 69
XDMLTool, 79
XeLDA, 75
Xerox Tagger, 75
XML, 35, 50, 70, 74, 79–81
XTAG, 78

YamCha, 77

Index of Authors

Abdalla, M. I., 120
Aberdeen, J., 81, 109
Aberle, D. R., 22, 111
Abraham, C., 24, 112
Abrams, R. S., 29, 122
Adachi, S., 19, 119
Ahmad, S., 119
Ahmed, A., 18, 119
Aizono, T., 42, 116
Allen, B., 45, 119
Allen, J., 79, 109, 113
Altman, R. B., 37, 120
Amir, L., 124
Ammerman, A. S., 111
Amorrort, E., 118
Ananiadou, S., 35, 123, 124
Andén, F., 116
Angus, K., 23, 123
Archibald, J., 38, 111
Arnold, R. M., 65, 109
Aronson, A. R., 28, 29, 110
Ash, N., 26, 125
Assal, J. P., 22, 25, 109
Aufseesser, M., 25, 109
Aufseesser-Stein, M., 22, 109
Avenarius, H. J., 26, 109

Bachimont, B., 112
Back, A. L., 65, 109, 117
Badloo, K., 121
Baile, W. F., 65, 109
Baker, D. L., 65, 109
Baker, K., 115
Bakken, S., 22, 30, 86, 113, 121
Balon, R., 18, 119
Barrett, A., 116
Batra, V., 119
Baud, R., 29, 36, 38, 113, 118, 121
Baud, R. H., 31, 109
Bauer, B. A., 113
Beattie, J. A. G., 120
Bechhofer, S., 121
Belz, A., 47, 110
Bental, D., 116
Berche, A., 41, 109
Bergstrom, L. R., 113
Berry, D. C., 26, 117
Bickmore, T., 39, 109
Bindels, P. J., 22, 117
Binsted, K., 46, 109, 111
Binyet, S., 25, 109
Blachar, Y., 124
Blake, D. R., 24, 109
Blaylock, N., 113
Block, S. D., 65, 117
Bodenreider, O., 33, 35, 36, 109
Boer, D. J. D., 123
Boisvieux, J.-F., 112
Bonneau-Maynard, H., 115
Bontcheva, K., 81, 112
Borg, A., 32, 110
Bosch, A. van den, 41, 110
Bouayad-Agha, N., 47, 71, 84, 110
Bouhaddou, O., 18, 110
Bourigault, D., 42, 77, 110
Branson, R., 120
Brants, T., 69, 75, 110
Bray, P., 113
Brenna, P. F., 28, 29, 110
Brennan, P. F., 35, 116
Brill, E., 73, 110
Briscoe, E., 82, 110
Broadhead, W. E., 29, 125
Broberg, C., 32, 111
Brouwer, H. J., 22, 117
Brown, S. H., 113
Browne, A. C., 125
Bruijn, J. de, 120
Brun, C., 45, 111
Brunie, V., 112
Bryett, A., 120
Bucknall, C., 37, 121
Bui, A. A., 22, 29, 111
Buitelaar, P., 124, 125
Buller, J., 18, 125
Burén, A., 111
Burger, J., 109
Burnage, G., 41, 111

Bush, P. J., 65, 115
Byron, D., 113

Cameron, M. J., 20, 111
Campbell, A., 120
Campbell, D. A., 35, 111
Campbell, J. R., 29, 111
Campbell, M. K., 63, 64, 111
Canning, Y., 38, 111
Carberry, S., 48, 111
Carenini, G., 46, 119
Carlson, L., 78, 111
Carlsson, E., 111
Carroll, J., 38, 82, 110, 111
Carruth, W., 113
Cawsey, A. J., 43, 46, 48, 86, 109, 111, 116
Cecil, R. R., 22, 114
Cederberg, S., 125
Chambers, N., 113
Chan, C.-K., 125
Chanda, G., 119
Chandrasekar, R., 38, 111, 112
Chang, S.-F., 119
Chapman, K., 24, 112
Chapman, W. W., 27, 35, 122
Charlet, J., 69, 79, 112
Chemo, M., 124
Chen, E., 119
Chen, M.-J., 35, 125
Cheng, S. H., 23, 112
Chiao, Y.-C., 42, 112
Chiao, Y. C., 31, 112
Chodkiewicz, C., 42, 110
Chung, K. P., 23, 112
Church, K., 42, 112
Church, K. W., 42, 114
Cimino, J. J., 119
Claveau, V., 31, 112
Cloete, I., 41, 124
Conley, E. S., 42, 112
Corbin, D., 41, 112
Core, M., 79, 109
Covvey, D., 113
Cowan, D., 113
Craig, N., 116
Crawley, R., 38, 111
Croft, B. W., 40, 125
Crowell, J., 32, 125
Cunningham, H., 81, 112
Curries, L., 86, 113

Daelemans, W., 41, 110
Dagan, I., 42, 112
Daille, B., 77, 112
Dal, G., 41, 112, 115
Darmoni, S. J., 21, 123
Daumke, P., 31, 112
Davies, M. J., 20, 112
Day, D., 109
Deber, R. B., 18, 113
Degerstedt, L., 116
Déjean, H., 41, 113
Delacey, S. L., 20, 112
Dermatas, E., 73, 113
Despont-Gros, C., 29, 113
DeVellis, B. M., 111
DeVellis, R. F., 111
Devillers, L., 69, 113, 115
Devlin, S., 38, 111
Dibble, E., 125
DiCiccio, V., 113
DiFlorio, I., 25, 113
DiMarco, C., 44, 113, 116
Dimitrov, M., 122
Divita, G., 125
Dordoni, A., 22, 111
Dorian, C., 38, 111
Douyère, M., 21, 123
Driscoll, C., 22, 119
Dunn, E. B., 26, 113
Dyche, L., 21, 113
Dykes, P. C., 86, 113
Dymetman, M., 45, 111
Dzikovska, M., 113

E, E. M., 18, 119
Eadie, D., 23, 123
Eash, T., 65, 109
Egerod, I., 23, 113
Ehnfors, M., 111
Eijk, P. van der, 42, 113
Elhadad, N., 46, 48, 113
Elkin, P. L., 37, 113
El-Saden, S., 22, 111
Engel, H. J., 19, 115
Enlund, H., 65, 115
Evens, M., 45, 118
Evoy, D., 24, 118

Fabry, P., 29, 113
Falkman, G., 45, 124
Fallowfield, L., 24, 112
Febles, A., 120
Feier, C., 120
Feiner, S., 119
Fellbaum, C., 37, 60, 113, 122

Ferguson, G., 86, 113
Ferguson, I., 120
Fletcher, K. E., 24, 109
Florin, J., 111
Flycht-Eriksson, A., 61, 113, 116
Foster, G. F., 42, 122
Freidin, R. B., 22, 114
Friedman, C., 35, 86, 118, 119
Friend, J. A., 117, 120
Fryer-Edwards, K., 65, 109
Fujita, A., 38, 116
Fung, P., 42, 114
Furukawa, M., 19, 119

Gale, W. A., 42, 114
Galeazzi, E., 34, 121
Gangemi, A., 34, 36, 114, 121
Gaussier, E., 40, 114
Geissbühler, A., 38, 121
Giorgino, T., 39, 109
Gittelman, M. A., 24, 114
Glowinski, A., 34, 121
Goble, C. A., 121
Goldberg, H., 28, 114
Goldberg, H. S., 28, 114
Goldman, L., 22, 114
Goldsmith, D., 114
Golodetz, A., 20, 114
Gonzales, M., 42, 124
Good, B. M., 36, 114
Goodare, H., 18, 115
Gordon, D., 23, 27, 114, 123
Gosselink, J., 114
Grabar, N., 36, 40, 42, 71, 114, 115, 126
Gravano, L., 119
Greenes, R. A., 26, 125
Grefenstette, G., 92, 94, 117
Grehn, L., 32, 114
Groundmeijer, H. G., 22, 117
Guedj, P.-O. E., 86, 114

Haag, K., 42, 114
Haas, S. W., 29, 124
Habeck, D., 19, 115
Habert, B., 71, 115
Hadlow, J., 25, 115
Hadouche, F., 40, 126
Hahn, U., 30, 31, 36, 37, 113, 115, 118, 121, 122
Hail, J. J., 29, 122
Halaschek-Wiener, C., 38, 116
Hales, J. W., 28, 115
Hallett, C., 43, 115
Hameen-Anttila, K., 65, 115
Hamilton, M., 37, 115
Hamm, R. M., 21, 123
Hammond, W. E., 29, 125
Hansen, S., 69, 110
Hardardottir, G. A., 35, 116
Hardy, H., 79, 115
Harvey, T., 48, 111
Hathout, N., 40–42, 109, 112, 115, 123
Hatzivassiloglou, V., 42, 122
Heidrich, F., 26, 123
Heja, G., 26, 123
Helgesson, G., 23, 115
Henley, L. D., 25, 115
Henry, S. B., 30, 115
Herxheimer, A., 18, 115
Hewlett, D., 38, 116
Hiemstra, D., 42, 116
Hier, D., 45, 118
Hill, I. D., 25, 115
Hirschman, L., 109
Hirst, G., 37, 44, 113, 116, 120
Hitt, B., 20, 116
Hjorth, K., 111
Ho, Y. C., 23, 112
Hoffmann, S., 95, 116
Holzemer, W. L., 30, 115
Horrocks, I., 121
Hospers, H., 64, 122
Hospers, H. J., 123
Hovy, E., 44, 113, 116
Hovy, E. H., 60, 116
Hsieh, Y., 35, 116
Hsu, C., 28, 114
Humbley, J., 42, 110
Hunter, J., 43, 48, 116, 123
Hüske-Kraus, D., 48, 116
Husser, C. S., 113
Hutnik, C. M., 19, 121
Hyland, M. E., 22, 116

Iida, R., 38, 116
Inui, K., 38, 116
Isabelle, P., 42, 122
Iwakura, T., 38, 116

Jacquemart, P., 71, 115
Jacquemin, C., 40, 116
Jang, F.-L., 37, 125
Janicke, D. M., 23, 118
Jehle, D. V., 23, 118
Jenkins, V., 24, 112
Johnson, S. B., 35, 111
Jones, R., 18, 125

Jones, R. B., 43, 46, 62–65, 109, 111, 116
Jönsson, A., 61, 116
Jordan, D., 45, 48, 113, 119
Junnila, J., 18, 125
Jurafsky, D., 41, 122

Kaji, H., 42, 116
Kalyanpur, A., 38, 116
Kan, M.-Y., 77, 116
Karla, E. K., 20, 117
Kassar, S. el, 112
Kaufman, D., 48, 113
Keck, K., 114
Kehler, M., 18, 117
Kemppainen, K., 65, 115
Kessler, D. P., 28, 115
Khan, S., 18, 119
Kilgarriff, A., 92, 94, 117
Kilpatrick, S., 22, 119
Kim, E., 32, 125
Kim, J.-D., 69, 117, 124
Kirkman, R., 20, 119
Kiryakov, A., 122
Klar, R., 36, 121
Klavans, J. L., 116
Klein, D., 73, 124
Klein, M., 60, 117
Knapp, P., 26, 117
Kobrin, S., 123
Kogan, S., 26, 125
Kokkinakis, D., 82, 88, 117, 118
Kokkinakis, G., 73, 113
König, E., 69, 117
Kornai, A., 35, 117
Koskenniemi, K., 41, 117
Kothari, R., 119
Kouylekov, M., 122
Kovlovski, V., 38, 116
Kraetschmer, N., 18, 113
Kreuter, M., 123
Krovetz, R., 39, 40, 117

Laarschot, R. van, 60, 117
Lacroix, A., 22, 25, 109
Lamel, L., 69, 113, 115
Larsen, J. H., 21, 117
Lauteslager, M., 22, 117
Law, V., 28, 114
Lecomte, J., 41, 109
Lee, S. J., 65, 117
Lennox, A. S., 47, 63, 64, 117
Lerner, E. B., 23, 118
Levow, G.-A., 30, 118
Lezius, W., 69, 117
Li, P., 45, 118
Lindsley, O. R., 26, 118
Liu, H., 35, 86, 118
Longwell, J., 42, 124
Lovel, H., 84, 123
Lovins, J., 72, 118
Lovis, C., 29, 31, 36, 109, 113, 118
Ludvigson, J., 23, 115
Lux, V., 45, 111
Lyman, M., 37, 121

Mackay, I. R., 20, 123
Madsen, P., 20, 118
Mahabee-Gittens, E. M., 24, 114
Maharaj, D., 121
Mani, I., 38, 118
Manning, C., 73, 124
Manning, C. D., 75, 124
Manov, D., 120
Marcinkiewicz, M., 73, 118
Marcu, D., 78, 111, 118
Marcus, M., 73, 110, 118
Marja, A., 65, 115
Markó, K., 30, 31, 112, 118, 122
Martín-Recuerda, F., 120
Marttala, U. M., 86, 118
Maynard, D., 81, 112
McArthur, G., 37, 120
McCann, I., 117
McCormack, D., 24, 118
McCue, K., 26, 119
McGregor, S., 116
McKeown, K. R., 21, 42, 45, 46, 48, 95, 113, 114, 116, 119, 121, 122
McNaught, J., 124
Medelyan, A., 30, 118
Merkel, M., 116, 118
Michel, P., 36, 118
Michel, P. A., 31, 109
Michener, J. L., 29, 125
Milhous, R. L., 20, 114
Miller, G. A., 37, 119
Miller, J., 22, 119
Mills, K., 20, 119
Minnen, G., 38, 111
Mitkov, R., 80, 119
Mittal, V. O., 46, 119
Miyawaki, S., 19, 119
Mohrs, J., 22, 117
Moore, J. D., 46, 119
Mosati, R. M., 23, 118
Mougin, F., 41, 109

Mukherjea, S., 36, 119
Mulcahy, D., 24, 118
Mulhall, K. J., 18, 119
Mullin, N., 20, 119
Munstermann, W., 19, 115
Musen, M. A., 35, 120

Namer, F., 41, 76, 112, 115, 119
Naraynsingh, V., 121
Naseem, A., 18, 119
Neuberger, J., 18, 120
Ngo, L., 125
Nhan, N., 37, 121
Niu, Y., 37, 120
Nohama, P., 31, 122
Norberg, S., 116
Nordby, H., 23, 120
Norman, R. J., 20, 112
Nowlan, W., 121
Nugues, P., 86, 114
Nyman, H., 32, 120

Oard, D. W., 30, 118
Ogden, J., 24, 30, 120, 125
Ohta, T., 69, 117, 124
Okumura, M., 38, 122
Oliver, D. E., 35, 37, 120
Osman, L., 47, 121
Osman, L. M., 44, 62, 64, 117, 120, 121
Ownsby, R. L., 32, 120

Paice, C., 72, 120
Pakhomov, S. V., 35, 36, 109
Pala, K., 80, 120
Pan, S., 45, 119
Parsons, K., 44, 116
Payne, T. H., 29, 111
Pearce, D., 111
Pearson, J., 46, 111, 116
Peckham, T. J., 25, 120
Peev, Z., 122
Penney, G. C., 20, 111
Pentheroudakis, J., 40, 120
Perry, A., 121
Perry, G., 19, 121
Peters, S., 125
Petersen, M., 86, 120
Pirrelli, V., 41, 120
Pisanelli, D. M., 36, 114
Pitts, M., 25, 115
Plovnick, R. M., 21, 28, 120, 125
Porter, M., 39, 72, 120
Power, R., 47, 84, 110, 120
Predoiu, L., 61, 120
Prescher, D., 124
Putnam, S., 21, 117

Quillen, E. J., 22, 119

Radev, D. R., 95, 121
Raileanu, D., 124
Ramdass, M. J., 18, 121
Rassinoux, A., 36, 118
Rassinoux, A. M., 31, 109
Ratnaparkhi, A., 73, 121
Raynor, D. K., 26, 117
Rayson, P., 80, 121, 125
Rector, A., 34, 37, 121
Reiter, E., 43, 47, 48, 62–64, 116, 117, 121, 123
Resnik, P., 30, 118, 121
Rey, J. G. del, 24, 114
Rice, A., 18, 117
Richardson, T., 26, 121
Ripplinger, B., 124
Risor, O., 21, 117
Robertson, R., 47, 117, 121
Robinson, P., 109
Rodning, C. B., 26, 121
Rodriguez-Gianolli, P., 37, 120
Rogers, J., 34, 37, 121
Romacker, M., 36, 115, 121
Romera, M., 118
Ross, S. J., 120
Rosset, S., 115
Rossi-Mori, A., 34, 121
Roth, D. H., 29, 122
Roth, L., 125
Ruch, P., 29, 38, 113, 121
Ruess, J., 20, 114
Ruland, C. M., 22, 30, 121
Russell, I. T., 120

Sacaleanu, B., 124
Sadan, B., 30, 121
Safran, C., 28, 114
Sager, N., 37, 121
Sandler, R. S., 111
Sankararaman, S., 119
Santorini, B., 73, 118
Sbrissia, E., 31, 122
Scanlan, D., 19, 121
Scharffe, F., 120
Scherpbier, H. J., 29, 122
Scherrer, J., 36, 118
Scherrer, J. R., 31, 109
Schiller, A., 73, 122

Schmidt, H., 73, 74, 122
Schnattinger, K., 36, 121
Schoeffler, K. M., 28, 115
Schone, P., 41, 122
Schuette, J. L., 65, 109
Schulz, S., 30, 31, 36, 112, 115, 118, 121, 122
Scott, D., 43, 47, 84, 110, 115, 120
Senda, Y., 38, 122
Shackle, E. M., 20, 122
Shahar, Y., 35, 120
Sharpe, N., 18, 113
Shaw, J., 45, 119
Shehata, M., 114
Shortliffe, E. H., 35, 120
Siddhartan, A., 38, 122
Siddiqui, F., 19, 121
Siegal, G., 124
Siegal, N., 124
Simard, M., 42, 122
Simov, A., 122
Simov, K., 81, 122
Singer, Y., 73, 124
Singhera, G. K., 114
Singleton, A., 18, 125
Sinohara, Y., 38, 122
Skinner, C. S., 64, 122, 123
Slaughter, L., 34, 123
Smadja, F., 42, 122
Smith, B., 37, 60, 113, 122
Smith, C. A., 27, 34, 35, 122
Smith, S. L., 21, 123
Smrz, P., 80, 120
Soderland, S. G., 38, 123
Soergel, D., 29, 34, 123, 124
Solomon, W., 121
Somers, H., 84, 123
Soualmia, L. F., 21, 29, 123
Spadaro, R., 62, 123
Spasic, I., 35, 123
Spees, C. M., 25, 123
Spiro, D., 26, 123
Spri, 31, 123
Srinivas, B., 38, 111, 112
Sripada, S., 43, 48, 116, 123
Stagg, R., 20, 123
Stahl, E., 22, 116
Stavri, P. Z., 27, 35, 122
Stead, M., 23, 123
Steensma, D. P., 20, 123
Steffen, D., 125
Steimann, F., 38, 123
Steve, G., 36, 114
Stevens, D. P., 20, 123
Stewart, S. K., 65, 117
Stöckert, C., 73, 122
Stolt, U. G., 23, 115
Stone, L., 35, 117
Strecher, V. J., 63, 64, 111, 122, 123
Strzalkowski, T., 115
Subramaniam, L. V., 119
Surjan, G., 26, 123
Swidenski, D., 21, 113

Tablan, V., 81, 112
Taira, R. K., 22, 38, 111, 123
Tait, J., 38, 111
Takada, K., 19, 119
Takahashi, T., 38, 116
Tan, P. C., 114
Tanguy, L., 42, 123
Tateishi, Y., 124
Tateisi, Y., 69, 117
Teelucksingh, S., 121
Tennison, W. P. J., 69, 123
Teufel, S., 73, 122
Theron, P., 41, 124
Thielen, C., 73, 122
Thirion, B., 21, 123
Thomas, J. A., 87, 124
Thompson, G., 87, 124
Thompson, H. J., 27, 124
Tick, L. J., 37, 121
Toporowska-Gronostaj, M., 118
Torgersson, O., 45, 124
Toutanova, K., 73, 75, 124
Tranfield, E. M., 114
Travers, D. A., 29, 124
Tse, C. K., 29, 125
Tse, T., 21, 29, 32, 34, 65, 123–125
Tse, T. A., 32, 125
Tsujii, J., 69, 117
Tsuruoka, Y., 74, 124
Tulsky, J. A., 65, 109
Tuttle, M., 114

Uhlmann, W. R., 65, 109
Urowitz, S., 18, 113

Vanderwende, L., 40, 120
Vasilescu, I., 69, 113
Viegas, E., 42, 124
Vilain, M., 109
Vintar, S., 42, 76, 79, 124

Wagner, J., 86, 120
Waisman, Y., 19, 124
Walsh, M., 24, 118

Wanner, L., 44, 113
Warner, H., 18, 110
Webber, B. L., 43, 111
Weber, B. M., 24, 109
Weijters, T., 41, 110
Weizenbaum, J., 39, 124
Werry, C. C., 97, 124
White, J., 116
White, P., 18, 125
Widdows, D., 35, 125
Wilkinson, J., 44, 113
Williams, N., 30, 125
Williams, S., 18, 120
Williams, S. H., 38, 125
Wilson, A., 80, 87, 121, 124, 125
Wittich, A. C., 18, 125
Wolfe, J. J., 26, 113
Wu, C.-H., 35, 37, 125
Wynn, R., 87, 125

Xu, J., 40, 125

Yarnall, K. S., 29, 125
Yeh, J.-F., 35, 125
Yu, J., 48, 123
Yu, L.-C., 35, 37, 125
Yvon, F., 41, 120

Zeng, Q., 21, 26, 28, 32, 65, 120, 125
Zielstorff, R. D., 30, 126
Zur-Mühlen, B. von, 32, 126
Zweigenbaum, P., 31, 36, 40, 42, 71, 112, 114, 115, 118, 126

Bibliography

Aberdeen, J., Burger, J., Day, D., Hirschman, L., Robinson, P., & Vilain, M. (1995, November). MITRE: Description of the Alembic system as used for MUC-6. In *Proceedings of the sixth message understanding conference (MUC-6)* (pp. 141–155). Columbia, Maryland, USA: Morgan Kaufmann Publishers.

Allen, J., & Core, M. (1997). *Draft of DAMSL: Dialog act markup in several layers.* (Available from `http://www.cs.rochester.edu/research/cisd/resources/damsl/` [accessed 8 May 2006])

Assal, J. P., Lacroix, A., & Aufseesser-Stein, M. (1992). Better listening, better information, better teaching. Lessons of our incompetence. *Schweiz Rundsch Med Prax*, *81*(6), 147-51.

Aufseesser, M., Lacroix, A., Binyet, S., & Assal, J. P. (1995). Diabetic retinopathy. interpretation of medical terms by patients. *J Fr Ophtalmol*, *18*(1), 27–32.

Avenarius, H. J. (1994). Language use in medicine. *Wien Med Wochenschr*, *144*(18–19), 460–3.

Back, A. L., Arnold, R. M., Baile, W. F., Tulsky, J. A., & Fryer-Edwards, K. (2005). Approaching difficult communication tasks in oncology. *CA: A Cancer Journal for Clinicians*, *55*, 164–177.

Baker, D. L., Eash, T., Schuette, J. L., & Uhlmann, W. R. (2002). Guidelines for writing letters to patients. *J Genetic Counselling*, *11*(5), 399–418.

Baud, R. H., Lovis, C., Rassinoux, A. M., Michel, P. A., & Scherrer, J. R. (1998, August). Automatic extraction of linguistic knowledge from an international classification. In *MEDINFO'98. proceedings of the 9th world congress on medical informatics* (Vol. 1, pp. 581–585). Seoul.

Berche, A., Mougin, F., Hathout, N., & Lecomte, J. (1997). *Verbaction : constitution d'un lexique déverbal du français* (Rapport technique). INIST.

Bickmore, T., & Giorgino, T. (2004). Some novel aspects of health communication from a dialogue systems perspective. In *Proceedings of the AAAI fall symposium on dialogue systems for health communication.* Washington, DC.

Binsted, K., Cawsey, A. J., & Jones, R. B. (1995). Generating personalised patient information using the medical record. In *Proceedings of artificial intelligence in medicine europe.* Pavia, Italy.

Binyet, S., Aufseesser, M., Lacroix, A., & Assal, J. P. (1994). The diabetic foot: various interpretations by patients of some terms used by physicians in podiatric consultation. *Diabete Metab*, *20*(3), 275–81.

Blake, D. R., Weber, B. M., & Fletcher, K. E. (2004). Adolescent and young adult women's misunderstanding of the term Pap smear. *Arch Pediatr Adolesc Med*, *158*(10), 966–70.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, *32*.

Bodenreider, O., & Pakhomov, S. V. (2003, July). Exploring adjectival modification in

biomedical discourse across two genres. In *Proceedings of the ACL 2003 workshop on natural language processing in biomedicine.*

Borg, A. (2005). *Bipacksedlars läsbarhet. En studie av bipacksedlar med focus på läsbarhet.* Unpublished master's thesis, Luleå tekniska universitet. (Available from `http://epubl.ltu.se/1404-5516/2005/48/LTU-HV-EX-0548-SE.pdf` [accessed 2 May 2006])

Bosch, A. van den, Daelemans, W., & Weijters, T. (1996). Morphological analysis as classification: an inductive-learning approach. In *International conference on computational linguistics (COLING).*

Bouayad-Agha, N. (2000). Layout annotation in a corpus of patient information leaflets. In *LREC 2000* (pp. 507–510). Athen, Greece.

Bouayad-Agha, N., Power, R., Scott, D., & Belz, A. (2002). PILLS: Multilingual generation of medical information documents with overlapping content. In *Proceedings of the 3rd international conference on language resources and evaluation (LREC 2002)* (pp. 2111–2114). (Available from `http://mcs.open.ac.uk/rp3242/papers/ITRI-02-04.pdf` [accessed 8 May 2006])

Bouayad-Agha, N., Scott, D., & Power, R. (2000). Integrating content and style in documents: a case study of patient information leaflets. *Information Design Journal*, *9*(2).

Bouayad-Agha, N., Scott, D., & Power, R. (2001). The influence of layout on the interpretation of referring expressions. In L. Degand, Y. Bestgen, W. Spooren, & L. van Waes (Eds.), *Multidisciplinary approaches to discourse* (pp. 133–141). Amsterdam: Stiching Neerlandistiek VU. (Available from `http://mcs.open.ac.uk/rp3242/papers/mad_bsp.pdf` [accessed 8 May 2006])

Bouhaddou, O., & Warner, H. (1995). An interactive patient information and education system (Medical HouseCall) based on a physician expert system (Iliad). In *Medinfo* (pp. 1181–5).

Bourigault, D. (1995). Lexter: A terminology extraction software for knowledge acquisition from texts. In *KAW'95.*

Bourigault, D., Chodkiewicz, C., & Humbley, J. (1999, 10–11 mai). Construction d'un lexique bilingue des droits de l'homme à partir de l'analyse automatique d'un corpus aligné. In *Terminologie et intelligence artificielle (TIA).* Nantes.

Brants, T. (2000). TnT – a statistical part-of-speech tagger. In *6th ANLP conference.* Seattle, WA.

Brants, T., & Hansen, S. (2002). Developments in the TIGER annotation scheme and their realization in the corpus. In *Third conference on language resources and evaluation LREC 2002.* Las Palmas de Gran Canaria, Spain.

Brenna, P. F., & Aronson, A. R. (2003). Towards linking patient and clinical information: detecting UMLS concepts in e-mail. *J Biomed Inform*, *36*(4–5), 334–341.

Brill, E. (1992). A simple rule-based part of speech tagger. In *Third annual conference on applied natural language processing.*

Brill, E. (1995). Unsupervised learning of disambiguation rules for part of speech tagging. In *Proceedings of the third workshop on very large corpora* (pp. 1–13). (Available from `http://www.cs.jhu.edu/~brill/acl-wkshp.ps` [accessed 8 March 2006])

Brill, E., & Marcus, M. (1992). Tagging an unfamiliar text with minimal human supervision. In *Proceedings of the AAAI symposium on probabilistic approaches to natural language* (pp. 10–16). (Available from `http://citeseer.ifi.unizh.ch/brill92tagging.html` [accessed 11 May 2006])

Briscoe, E., & Carroll, J. (2002). Robust accurate statistical annotation of general text. In

Third international conference on language resources and evaluation (LREC 2002) (pp. 1499–1504). Las Palmas, Canary Islands.

Broberg, C., Burén, A., Carlsson, E., Ehnfors, M., Florin, J., Hjorth, K., et al. (2006). Sjukvården behöver ett tvärprofessionellt språk. *Dagens Medicin*, *4*.

Brun, C., Dymetman, M., & Lux, V. (2000). Document structure and multilingual text authoring. In *Proceedings of INLG'2000.*

Bui, A. A., Taira, R. K., El-Saden, S., Dordoni, A., & Aberle, D. R. (2004). Automated medical problem list generation: towards a patient timeline. In *Medinfo* (pp. 587–91).

Burnage, G. (1990). *CELEX – A guide for users.* University of Nijmegen: Centre for Lexical Information.

Cameron, M. J., & Penney, G. C. (2005). Terminology in early pregnancy loss: what women hear and what clinicians write. *J Fam Plann Reprod Health Care*, *31*(4), 313–4.

Campbell, D. A., & Johnson, S. B. (1999). A technique for semantic classification of unknown words using umls resources. In *Proceedings of the AMIA annual symposium* (pp. 716–720).

Campbell, J. R., & Payne, T. H. (1994). A comparison of four schemes for codification of problem lists. In *Proc annu symp comput appl med care* (pp. 201–5).

Campbell, M. K., DeVellis, B. M., Strecher, V. J., Ammerman, A. S., DeVellis, R. F., & Sandler, R. S. (1994). Improving dietary behavior. The effectiveness of tailored messages in primary care settings. *American Journal of Public Health*, *84*(5), 783–787.

Canning, Y., Tait, J., Archibald, J., & Crawley, R. (2000). Cohesive regeneration of syntactically simplified newspaper text. In *Proceedings of ROMAND 2000.* Lausanne.

Carberry, S., & Harvey, T. (1997). Generating coherent messages in real-time decision support. In *Proc. cognitive science conference.*

Carlson, L., & Marcu, D. (2001). *Discourse tagging reference manual* (Tech. Rep.). Univ. of Southern California / Information Sciences Institute.

Carroll, J., Minnen, G., Canning, Y., Devlin, S., & Tait, J. (1998). Practical simplification of English newspaper text to assist aphasic readers. In *AAAI-98 workshop on integrating artificial intelligence and assistive technology.* Madison, Wisconsin.

Carroll, J., Minnen, G., Pearce, D., Canning, Y., Devlin, S., & Tait, J. (1999). Simplifying text for language-impaired readers. In *Proceedings of EACL'99.* Bergen.

Cawsey, A. J., Binsted, K., & Jones, R. B. (1995a, April). An on-line explanation of the medical record to patients via an artificial intelligence approach. In B. Richards (Ed.), *Proceedings of healthcare computing 1995* (pp. 269–275).

Cawsey, A. J., Binsted, K., & Jones, R. B. (1995b). Personalised explanations for patient education. In *Proceedings of the fifth european workshop on natural language generation, 20–22 may 1995* (pp. 59–74). Leiden, the Netherlands.

Cawsey, A. J., Jones, R. B., & Pearson, J. (2000). The evaluation of a personalised information system for patients with cancer. *User Modeling and User-Adapted Interaction*, *10*(1), 47–72.

Cawsey, A. J., Webber, B. L., & Jones, R. B. (1997, Nov/Dec). Natural language generation in healthcare. *Journal of the American Medical Informatics Association*, *4*(6), 473–482. (Available from `http://citeseer.ifi.unizh.ch/cawsey97natural.html`)

Chandrasekar, R., Dorian, C., & Srinivas, B. (1997). Motivations and methods for text simplification. In *International conference on computational linguisitics (COLING 96)* (pp. 1041–1044). Copenhagen.

Chandrasekar, R., & Srinivas, B. (1997). Automatic rules for text simplification. *Knowledge-Based Systems*, *10*, 183–190.

Chapman, K., Abraham, C., Jenkins, V., & Fallowfield, L. (2003). Lay understanding of terms used in cancer consultations. *Psychooncology*, *12*(6), 557–66.

Charlet, J., Bachimont, B., Brunie, V., Kassar, S. el, Zweigenbaum, P., & Boisvieux, J.-F. (1998). Hospitexte: towards a document-based hypertextual electronic medical record. In *Proc. AMIA symp.* (pp. 713–717).

Cheng, S. H., Ho, Y. C., & Chung, K. P. (2002). Hospital quality information for patients in taiwan: can they understand it? *Int J Qual Health Care*, *14*(2), 155–60.

Chiao, Y.-C., & Zweigenbaum, P. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. In *International conference on computational linguistics (COLING)*. Taipei, Taiwan.

Chiao, Y. C., & Zweigenbaum, P. (2002, November). Looking for French-English translations in comparable medical corpora. In I. S. Kohane (Ed.), *AMIA 2002. proceedings of the annual symposium of the american medical informatics association. biomedical informatics: One discipline* (pp. 150–154). San Antonio, TX.

Claveau, V., & Zweigenbaum, P. (2005, July). Translating biomedical terms by inferring transducers. In *Artificial intelligence in medicine. proceedings of the 10th conference on artificial intelligence in medicine in europe. AIME 2005* (Vol. 3581, pp. 236–240). Aberdeen, Scotland.

Conley, E. S. (2002). *Seq_align: A parsing-independent bilingual sequence alignment algorithm.* Unpublished master's thesis, Bar-Ilan University, Ramat-Gan, Israël. (Available from `http://www.cs.biu.ac.il/~konli/Msc_thesis.html` [accessed 31 August 2002])

Corbin, D. (1987). *Morphologie dérivationnelle et structuration du lexique* (Vol. 1). Lille: Presse universitaire de Lille.

Corbin, D. (1991). La formation des mots : structures et interprétations. *Lexique*, *10*, 7–30.

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). GATE: A framework and graphical development environment for robust NLP tools and applications. In *40th anniversary meeting of the association for computational linguistics (ACL'02)*. Philadelphia, US.

Dagan, I., & Church, K. (1994). *Termight*: Identifying and translating technical terminology. In *Proceedings of the fourth conference on applied natural language processing (ANLP'94)* (pp. 34–40). Institut for Computational Linguistics. University of Stuttgart, Germany.

Daille, B. (2003). Conceptual structuring through term variations. In F. Bond, A. Korhonen, D. MacCarthy, & A. Villacicencio (Eds.), *ACL 2003 workshop on multiword expressions: Analysis, acquisition and treatment* (pp. 9–16).

Dal, G., Namer, F., & Hathout, N. (1999). Construire un lexique dérivationnel : théorie et réalisations. In P. Amsili (Ed.), *Traitement automatique des langues naturelles (TALN)* (pp. 115–124). Cargèse.

Daumke, P., Schulz, S., & Markó, K. (forthcoming). Subword approach for acquiring and cross-linking multilingual specialized lexicons. In *LREC 2006 workshop: Acquiring and representing multilingual, specialized lexicons: the case of biomedicine.*

Davies, M. J., Delacey, S. L., & Norman, R. J. (2005). Towards less confusing terminology in reproductive medicine: clarifying medical ambiguities to the benefit of all. *Hum Reprod*, *20*(10), 2669–71.

Deber, R. B., Kraetschmer, N., Urowitz, S., & Sharpe, N. (2005). Patient, consumer, client, or customer: what do people want to be called? *Health Expect*, *8*(4), 345–51.

Déjean, H. (1998). Morphemes as necessary concept for structures discovery from untagged corpora. In *Workshop on paradigms and grounding in natural language learning* (pp. 295–299). Adelaide.

Dermatas, E., & Kokkinakis, G. (1995). Automatic stochastic tagging of natural language texts. *Computational Linguistics*, *21*(2), 137–163.

Devillers, L., Vasilescu, I., & Lamel, L. (2002). Annotation and detection of emotion in a task-oriented human-human dialog corpus. In *ISLE workshop on dialogue tagging.* Edinburgh.

DiFlorio, I. (1991). Mothers' comprehension of terminology associated with the care of a newborn baby. *Pediatr Nurs*, *17*(2), 193–6.

DiMarco, C., Bray, P., Covvey, D., Cowan, D., DiCiccio, V., Hovy, E., et al. (2005). Authoring and generation of tailored preoperative patient education materials. In *Workshop on personalisation in e-health, user modelling conference.* Edinburgh, Scotland.

DiMarco, C., Hirst, G., Wanner, L., & Wilkinson, J. (1995). Healthdoc: Customizing patient information and health education by medical condition and personal characteristics. In A. J. Cawsey (Ed.), *Proceedings of the workshop on patient education.* Glasgow.

Dunn, E. B., & Wolfe, J. J. (2001). Let go of latin! *Vet Hum Toxicol*, *43*(4), 235–6.

Dyche, L., & Swidenski, D. (2005). The effect of physician solicitation approaches on ability to identify patient concerns. *J Gen Intern Med*, *20*(3), 267–70.

Dykes, P. C., Curries, L., & Bakken, S. (2004). Patient education and recovery learning system (PEARLS) pathway: a tool to drive patient centered evidence-based practice. *Journal of Healthcare Information Management*, *18*(4).

Egerod, I. (2002). Uncertain terms of sedation in ICU. how nurses and physicians manage and describe sedation for mechanically ventilated patients. *J Clin Nurs*, *11*(6), 831–40.

Eijk, P. van der. (1993). Automating the acquisition of bilingual terminology. In *Proceedings of the 6th conference of the european chapter of the association for computational linguistics (ACL).* Utrecht, Pays-Bas. (Available from `http://home.multiweb.nl/~pvde/PDF/phralign.pdf` [accessed 31 August 2002])

Elhadad, N., & McKeown, K. R. (2001). Towards generating patient specific summaries of medical articles. In *Proceedings of NAACL workshop on automatic summarization* (pp. 31–39). Pittsburgh.

Elhadad, N., McKeown, K. R., Kaufman, D., & Jordan, D. (2005). Facilitating physicians' access to information via tailored text summarization. In *AMIA annual symposium.*

Elkin, P. L., Brown, S. H., Bauer, B. A., Husser, C. S., Carruth, W., Bergstrom, L. R., et al. (2005, May). A controlled trial of automated classification of negation from clinical notes. *BMC Medical Informatics and Decision Making*, *5*(13).

Fabry, P., Baud, R., Ruch, P., Despont-Gros, C., & Lovis, C. (2005). Methodology to ease the construction of a terminology of problems. *Int J Med Inform.*

Fellbaum, C., Hahn, U., & Smith, B. (forthcoming). Towards new information resources for public health. From WordNet to Medical WordNet. *Journal of Biomedical Informatics.*

Ferguson, G., Allen, J., Blaylock, N., Byron, D., Chambers, N., Dzikovska, M., et al. (2002). *The medication advisor project* (Tech. Rep. No. 766). University of Rochester, Computer Science Department.

Flycht-Eriksson, A. (2004). *Design and use of ontologies in information-providing dialogue systems.* Unpublished doctoral dissertation, School of Engineering at Linköping Uni-

versity. (Thesis No. 874, Linköping Studies in Science and Technology. ISBN: 91-7373-947-2.)

Freidin, R. B., Goldman, L., & Cecil, R. R. (1980). Patient-physician concordance in problem identification in the primary care setting. *Ann Intern Med*, *93*(3), 490–3.

Fung, P., & McKeown, K. R. (1997). Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th annual workshop on very large corpora.* Hong Kong. (Available from `http://www.ee.ust.hk/~pascale/wvlc97.ps` [accessed 31 August 2002])

Gale, W. A., & Church, K. W. (1991). A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th annual meeting of the association for computational linguistics (ACL).* Toulouse, France. (Available from `http://cm.bell-labs.com/cm/ms/departments/sia/doc/93.7.ps/` [accessed 31 August 2002])

Gangemi, A., Pisanelli, D. M., & Steve, G. (1998). Ontology integration: Experiences with medical terminologies. In N. Guarino (Ed.), *Formal ontology in information systems* (pp. 163–178). IOS Press.

Gaussier, E. (1999, June). Unsupervised learning of derivational morphology from inflectional lexicons. In A. Kehler & A. Stolcke (Eds.), *ACL workshop on unsupervised methods in natural language learning.* College Park, Md.

Gittelman, M. A., Mahabee-Gittens, E. M., & Rey, J. G. del. (2004). Common medical terms defined by parents: are we speaking the same language? *Pediatr Emerg Care*, *20*(11), 754–8.

Goldberg, H., Goldsmith, D., Law, V., Keck, K., Tuttle, M., & Safran, C. (1998). An evaluation of UMLS as a controlled terminology for the problem list toolkit. In *Medinfo* (pp. 609–12).

Goldberg, H. S., Hsu, C., Law, V., & Safran, C. (1998). Validation of clinical problems using a UMLS-based semantic parser. In *Annual symposium of the american medical informatics association (AMIA)* (pp. 805–9).

Golodetz, A., Ruess, J., & Milhous, R. L. (1976). The right to know: giving the patient his medical record. *Arch Phys Med Rehabil*, *57*(2), 78–81.

Good, B. M., Tranfield, E. M., Tan, P. C., Shehata, M., Singhera, G. K., Gosselink, J., et al. (2006). Fast, cheap and out of control: A zero curation model for ontology development. *Pacific Symposium on Biocomputing*, *11*, 128–139.

Gordon, D. (1996). MDs' failure to use plain language can lead to the courtroom. *CMAJ*, *155*(8), 1152–4.

Grabar, N., & Haag, K. (2003). Des textes parallèles vers une terminologie trilingue. In *Terminologie et intelligence artificielle (TIA)* (pp. 102–111). Strasbourg.

Grabar, N., & Zweigenbaum, P. (2000). A general method for sifting linguistic knowledge from structured terminologies. *JAMIASUP*, 310–314. (Available from `http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=11079895&dopt=Abstract` [accessed 16 March 2006])

Grehn, L. (2004). *Granskning av bipacksedlar för förebyggande och substituerande läkemedel.* (Receptarieprogrammet FAR 010, Naturvetenskapliga institutionen, GU – NEPI. Available from `http://www.nepi.net/linda.pdf` [accessed 2 May 2006])

Guedj, P.-O. E., & Nugues, P. (1994). A chart parser to analyze large medical corpora. In *Proceedings of the 16th annual international conference of the IEEE.* (Available from `http://citeseer.ifi.unizh.ch/elguedj94chart.html` [accessed 8 May 2006])

Habeck, D., Engel, H. J., & Munstermann, W. (1977). Patient's opinions on doctors' explanations. *MMW Munch Med Wochenschr*, *119*(25), 861–4.

Habert, B., Grabar, N., Jacquemart, P., & Zweigenbaum, P. (2001). Building a text corpus for representing the variety of medical language. In *Corpus linguistics.* Lancaster.

Hadlow, J., & Pitts, M. (1991). The understanding of common health terms by doctors, nurses and patients. *Soc Sci Med*, *32*(2), 193–6.

Hahn, U., Romacker, M., & Schulz, S. (1999). How knowledge drives understanding- matching medical ontologies with the needs of medical language processing. *Artificial Intelligence in Medicine*, *15*, 25–51.

Hahn, U., Romacker, M., & Schulz, S. (2000). MedSyndikate – design considerations for an ontology-based medical text understanding system. *International Journal of Medical Informatics*, *67*, 63–74.

Hahn, U., Romacker, M., & Schulz, S. (2002). Creating knowledge repositories from biomedical reports: The MedSyndikate text mining system. In *Proceedings of PSB* (pp. 338–349).

Hales, J. W., Schoeffler, K. M., & Kessler, D. P. (1998). Extracting medical knowledge for a coded problem list vocabulary from the UMLS knowledge sources. In *Annual symposium of the american medical informatics association (AMIA)* (pp. 275–9).

Hallett, C., & Scott, D. (2005). Structural variation in generated health reports. In *Proceedings of the 3rd international workshop on paraphrasing.*

Hameen-Anttila, K., Kemppainen, K., Enlund, H., Bush, P. J., & Marja, A. (2004). Do pictograms improve children's understanding of medicine leaflet information? *Patient Education and Counseling*, *55*(3), 371–8.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery and Psychiatry*, *23*, 56–62.

Hardy, H., Baker, K., Bonneau-Maynard, H., Devillers, L., Rosset, S., & Strzalkowski, T. (2003). Semantic and dialogic annotation for automated multilingual customer service. In *Eurospeech.* Geneva, Switzerland.

Hardy, H., Baker, K., Devillers, L., Lamel, L., Rosset, S., Strzalkowski, T., et al. (2002). Multi-layer dialogue annotation for automated multilingual customer service. In *SLE workshop on dialogue tagging for multi-modal human-computerinteraction.* Edinburgh, UK.

Hathout, N. (2001). Analogies morpho-syntaxiques. In *Traitement automatique des langues naturelles (TALN).* Tours.

Hathout, N., Namer, F., & Dal, G. (2001). An experimental constructional database: the MorTAL project. In P. Boucher (Ed.), *Morphology book.* Cambridge, MA: Cascadilla Press.

Helgesson, G., Ludvigson, J., & Stolt, U. G. (2005). How to handle informed consent in longitudinal studies when participants have a limited understanding of the study. *J Med Ethics*, *31*(11), 670–3.

Henley, L. D., & Hill, I. D. (1990). Errors, gaps, and misconceptions in the disease-related knowledge of cystic fibrosis patients and their families. *Pediatrics*, *85*(6), 1008–14.

Henry, S. B., & Holzemer, W. L. (1994). Can SNOMED international represent patients' perceptions of health-related problems for the computer-based patient record? In *Proc annu symp comput appl med care* (pp. 184–7).

Herxheimer, A., & Goodare, H. (1999). Who are you, and who are we? looking through some key words. *Health Expect*, *2*(1), 3–6.

Hewlett, D., Kalyanpur, A., Kovlovski, V., & Halaschek-Wiener, C. (2005). Effective natural language paraphrasing of ontologies on the semantic web. In *End user semantic web interaction workshop, international semantic web conference (ISWC).* Galway, Ireland.

Hiemstra, D. (1998). Multilingual domain modeling in Twenty-One: automatic creation of a bi-directional translation lexicon from a parallel corpus. In *Proceedings of the eighth CLIN meeting.* Leuven, Belgique. (Available from `http://wwwhome.cs.utwente.nl/~hiemstra/papers/clin.ps.gz/` [accessed 31 August 2002])

Hirst, G., DiMarco, C., Hovy, E., & Parsons, K. (1997). Authoring and generating health-education documents that are tailored to the needs of the individual patient. In A. Jameson, C. Paris, & C. Tasso (Eds.), *User modelling: Proceedings of the sixth international conference (UM97).* Vienna, New York: Springer.

Hitt, B. (1998). It's only words ... or is it. *Posit Dir News*, *10*(2), 33–4.

Hoffmann, S. (2006). Processing Internet-derived text. Creating a corpus of Usenet messages. *Journal of Literary and Linguistic Computing.* (In press)

Hovy, E. H. (2003). Using an ontology to simplify data access. In *Communications of the ACM, special issue on digital government.* ACM.

Hsieh, Y., Hardardottir, G. A., & Brennan, P. F. (2004). Linguistic analysis: Terms and phrases used by patients in e-mail messages to nurses. In *Medinfo 2004* (pp. 511–515).

Hunter, J., Reiter, E., & Sripada, S. (2006). *Babytalk. Generating textual summaries of clinical temporal data.* (Scientific Outline. Available from `http://www.csd.abdn.ac.uk/research/babytalk/` [accessed 2 May 2006])

Hüske-Kraus, D. (2003). Suregen-2: A shell system for the generation of clinical documents. In *Proceedings of EACL'03* (pp. 215–218).

Hyland, M. E., & Stahl, E. (2004). Asthma treatment needs: a comparison of patients' and health care professionals' perceptions. *Clin Ther*, *26*(12), 2141–52.

Inui, K., Fujita, A., Takahashi, T., Iida, R., & Iwakura, T. (2003). Text simplification for reading assistance: A project note. In *Second international workshop on paraphrasing: Paraphrase acquisition and applications (IWP2003).* (Available from `http://citeseer.ifi.unizh.ch/article/inui03text.html` [accessed 2 May 2006])

Jacquemin, C. (1997, July). Guessing morphology from terms and corpora. In *ACM SIGIR.*

Jones, R. B., Pearson, J., Cawsey, A. J., Bental, D., Barrett, A., White, J., et al. (2006, April). Effect of different forms of information produced for cancer patients on their use of the information, social support, and anxiety: randomised trial. *British Medical Journal*, *332*, 942–948. (Available from `http://bmj.bmjjournals.com/cgi/content/full/332/7547/942` [accessed 2 May 2006])

Jones, R. B., Pearson, J., McGregor, S., Cawsey, A. J., Barrett, A., Craig, N., et al. (1999). Randomised trial of personalised computer based information for cancer patients. *British Medical Journal*, *319*(7219), 1241–1247.

Jönsson, A., Andén, F., Degerstedt, L., Flycht-Eriksson, A., Merkel, M., & Norberg, S. (2004). Experiences from combining dialogue system development with information extraction techniques. In M. T. Maybury (Ed.), *New directions in question answering.* AAAI/MIT Press.

Kaji, H., & Aizono, T. (2001). Extracting word correspondences from bilingual corpora based on word co-occurrence information. *Information Processing Society of Japan (IPSJ) Journal 42(9).* (Available from `http://www.ipsj.or.jp/members/Journal/Eng/4209/` [accessed 31 August 2002])

Kan, M.-Y., Klavans, J. L., & McKeown, K. R. (1998). Linear segmentation and segment

relevance. In *6th international workshop of very large corpora (WVLC-6)* (pp. 197–205). Montréal, Canada.

Karla, E. K. (2003). Can a change in the name of depresseion result in its reduced negativity? *Med Hyoitheses*, *60*(6), 897–9.

Kehler, M., & Rice, A. (2004). How would patients like to be addressed? a brief survey. *Br J Gen Pract*, *54*(506), 704.

Kilgarriff, A., & Grefenstette, G. (2003). Web as corpus. Introduction to the special issue on the Web as corpus. *Computational Linguistics*, *29*(3), 333–347.

Kim, J.-D., Ohta, T., Tateisi, Y., & Tsujii, J. (2003). GENIA corpus – a semantically annotated corpus for bio-textmining. *Bioinformatics*, *19(Suppl. 1)*, i180-i182.

Klein, M. (2001, August 4–5,). Combining and relating ontologies: an analysis of problems and solutions. In A. Gomez-Perez, M. Gruninger, H. Stuckenschmidt, & M. Uschold (Eds.), *Workshop on ontologies and information sharing, IJCAI'01.* Seattle, USA. (Available from `http://citeseer.ifi.unizh.ch/klein01combining.html`)

Knapp, P., Raynor, D. K., & Berry, D. C. (2004). Comparison of two methods of presenting risk information to patients about the side effects of medicines. *Qual Saf Health Care*, *13*(3), 176–80.

Kokkinakis, D. (2006a). Collection, encoding and linguistic processing of a Swedish medical corpus – the MEDLEX experience. In *Proceedings of the 5th international conference on language resources and evaluation (LREC 2006).* Genoa.

Kokkinakis, D. (2006b). Developing resources for Swedish bio-medical text mining. In *Second international symposium on semantic mining in biomedicine.* Jena.

König, E., & Lezius, W. (2000). A description language for syntactically annotated corpora. In *COLING-2000* (pp. 1056–1060). Saarbrücken, Germany.

Kornai, A., & Stone, L. (2004). Automatic translation to controlled medical vocabularies. In A. Abraham & L. Jain (Eds.), *Innovations in intelligent systems and applications* (pp. 413–434). Springer Verlag.

Koskenniemi, K. (1983). *Two-level morphology: a general computational model for word-form recognition and production.* Unpublished doctoral dissertation, University of Helsinki Department of General Linguistics, Helsinki.

Krovetz, R. (1993). Viewing morphology as an inference process. In *Proceedings of the 16th annual international ACM-SIGIR conference on research and development in information retrieval* (pp. 191–202).

Laarschot, R. van. (2005). *Ontology-based knowledge modelling in Dutch civil law.* Unpublished master's thesis, Vrije Universiteit Amsterdam.

Larsen, J. H., Risor, O., & Putnam, S. (1997). P-r-a-c-t-i-c-a-l: a step-by-step model for conducting the consultation in general practice. *Fam Pract*, *14*(4), 295–301.

Lauteslager, M., Brouwer, H. J., Mohrs, J., Bindels, P. J., & Groundmeijer, H. G. (2002). The patient as a source to improve the medical record. *Fam Med*, *19*(2), 167–71.

Lee, S. J., Back, A. L., Block, S. D., & Stewart, S. K. (2002). Enhancing physician–patient communication. In *Hematology. american society of hematology education program book* (Vol. 1, pp. 464–483). American Society of Hematology. (Available from `http://www.asheducationbook.org/cgi/content/full/2002/1/464` [accessed 5 May 2006].)

Lennox, A. S., Osman, L. M., Reiter, E., Robertson, R., Friend, J. A., McCann, I., et al. (2001). The cost-effectiveness of computer-tailored and non-tailored smoking cessation letters in general practice. A randomised controlled trial. *British Medical Journal*, *322*, 1396–1400.

Lerner, E. B., Jehle, D. V., Janicke, D. M., & Mosati, R. M. (2000). Medical communication: do our patient understand? *Am J Energ Med*, *18*(7), 764–6.

Levow, G.-A., Oard, D. W., & Resnik, P. (2005). Dictionary-based techniques for cross-language information retrieval. *Information Processing and Management: an International Journal*, *41*(3), 523–547.

Li, P., Evens, M., & Hier, D. (1986). Generating medical case reports with the linguistic string parser. In *Proceedings of 5th national conference on artificial intelligence (AAAI-86)* (pp. 1069–1073). Philadelphia, PA.

Lindsley, O. R. (1991). From technical jargon to plain english for application. *J Appl Behav Anal*, *24*(3), 449–58.

Liu, H., & Friedman, C. (2000). A method for vocabulary development and visualization based on medical language processing and XML. In *Proceedings of the AMIA annual symposium* (pp. 502–506).

Liu, H., & Friedman, C. (2003). Mining terminological knowledge in large biomedical corpora. In *Pacific symposium on biocomputing.*

Lovins, J. (1968). Development of a stemming algorithm. *Mechanical translation and computational linguistics*, *11*, 22–31.

Lovis, C., Baud, R., Rassinoux, A., Michel, P., & Scherrer, J. (1998). Medical dictionaries for patient encoding systems: a methodology. *Artificial Intelligence in Medicine*, *14*, 201–214.

Madsen, P. (2005). Fertility speack: language and the patient. *Fertil Steril*, *82*(1), 844–5.

Mani, I. (2001). *Automatic summarization* (No. 3). Amsterdam: John Benjamins. (ISBN 1-58811-060-5)

Marcu, D., Amorrort, E., & Romera, M. (1999, June). Experiments in constructing a corpus of discourse trees. In *Proceedings of the ACL'99 workshop on standards and tools for discourse tagging.* University of Maryland.

Marcus, M., Santorini, B., & Marcinkiewicz, M. (1993). Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, *19*, 313–330.

Markó, K., Baud, R., Zweigenbaum, P., Merkel, M., Toporowska-Gronostaj, M., Kokkinakis, D., et al. (forthcoming). Cross-lingual alignment of medical lexicons. In *LREC 2006 workshop: Acquiring and representing multilingual, specialized lexicons: the case of biomedicine.*

Markó, K., Schulz, S., & Hahn, U. (2005, July). Unsupervised multilingual word sense disambiguation via an interlingua. In *AAAI 2005 – proceedings of the 20th national conference on artificial intelligence & IAAI'05 – proceedings of the 17th innovative applications of artificial intelligence conference* (pp. 1075–1080). Pittsburgh, PA, USA.

Markó, K., Schulz, S., Medelyan, A., & Hahn, U. (2005, August). Bootstrapping dictionaries for cross-language information retrieval. In *SIGIR 2005 – proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 528–535). Salvador, Brazil.

Marttala, U. M. (1995). *Innehåll och perspektiv i samtal mellan läkare och patient: En språklig och samtalsanalytisk undersökning (Content and perspective in doctor-patient conversations: a linguistic and conversation analytic investigation)* (Skrifter utgivna av Institutionen för nordiska språk vid Uppsala universitet No. 39). Uppsala: Uppsala universitet. (Summary: `http://helmer.aksis.uib.no/nlb/nlb-enews_1995_9.txt`)

McCormack, D., Evoy, D., Mulcahy, D., & Walsh, M. (1997). An evaluation of patient

comprehension of orthopeadic terminology: implications for informed consent. *J R Coll Surg Edinb*, *42*(1), 33–35.

McCue, K. (1991). Labels. *Child Health Care*, *20*(4), 248–9.

McKeown, K. R., Chang, S.-F., Cimino, J. J., Feiner, S., Friedman, C., Gravano, L., et al. (2001). PERSIVAL, a system for personalized search and summarization over multimedia healthcare information. In *ACM/IEEE joint conference on digital libraries* (pp. 331–340).

McKeown, K. R., Jordan, D., Feiner, S., Shaw, J., Chen, E., Ahmad, S., et al. (2002). A study of communication in the cardiac surgery intensive care unit and its implications for automated briefing. In *Proc. of the american medical informatics association 2000 symposium.*

McKeown, K. R., Pan, S., Shaw, J., Jordan, D., & Allen, B. (1997). Language generation for multimedia healthcare briefings. In *Proc. of the fifth conference on applied natural language processing* (pp. 277–282).

Miller, G. A. (1995, November). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Miller, J., Driscoll, C., Kilpatrick, S., & Quillen, E. J. (2003). Management of prenatal care information: integration of the problem list and clinical comments. *Top Health Inf Manage*, *24*(1), 42–9.

Mitkov, R. (1999). *Anaphora resolution: The state of the art* (Working paper). University of Wolverhampton. (Available from `http://citeseer.ifi.unizh.ch/mitkov99anaphora.html`)

Mittal, V. O., Carenini, G., & Moore, J. D. (1994). Generating patient specific explanations in migraine. In *Proceedings of the eighteenth annual symposium on computer applications in medical care.* McGraw-Hill Inc.

Miyawaki, S., Takada, K., Furukawa, M., & Adachi, S. (1995). An interactive consultation multimedia software for orthodontic patients. In *Medinfo* (p. 1308).

Mukherjea, S., Subramaniam, L. V., Chanda, G., Sankararaman, S., Kothari, R., Batra, V., et al. (2004). Enhancing a biomedical information extraction system with dictionary mining and context disambiguation. *IBM Journal of Research and Development*, *48*(5/6), 693–701.

Mulhall, K. J., Ahmed, A., & E, E. M. (2002). The "doctor-customer" relationship: Hippocrates in the modern marketplace. *Int J Health Care Qual Assur Inc Leadersh Health Serv*, *15*(1), 9–10.

Mullin, N., Mills, K., & Kirkman, R. (2004). Coil or intrauterine device? patient preferences for contraceptive terminology. *J Fam Plann Reprod Health Care*, *30*(1), 46–8.

Namer, F. (2000). FLEMM : un analyseur flexionnel du français à base de règles. *Traitement Automatique des Langues (TAL)*, *41*(2), 523–547.

Namer, F. (2002a). Acquisition automatique de sens à partir d'opérations morphologiques en français : étude de cas. In *Traitement automatique de la langue naturelle (TALN)* (pp. 235–244). Nancy.

Namer, F. (2002b). Acquisition automatique de sens à partir d'opérations morphologiques en français : étude de cas. In *Traitement automatique de la langue naturelle (TALN)* (pp. 235–244). Nancy.

Naseem, A., Balon, R., & Khan, S. (2001). Customer, client, consumer, recipient, or patient. *Ann Clin Psychiatry*, *13*(4), 239–40.

Neuberger, J., & Williams, S. (2001). Is the word 'patient' outmoded? *Nurs Times*, *97*(5), 16.

Niu, Y., Hirst, G., McArthur, G., & Rodriguez-Gianolli, P. (2003, July). Answering clinical questions with role identification. In *Proceedings of the ACL 2003 workshop on natural language processing in biomedicine.*

Nordby, H. (2003). Doctor-patient-interaction is non-holistic. *Med Health Care Philos*, *6*(2), 145–52.

Nyman, H. (1996). *Medicinens språk.* Almqvist & Wiksell Medicin / Liber.

Ogden, J., Branson, R., Bryett, A., Campbell, A., Febles, A., Ferguson, I., et al. (2003). What's in a name? an experimental study of patients' views of the impact and function of a diagnosis. *Fam Pract*, *20*(3), 248–53.

Oliver, D. E., & Altman, R. B. (1994). Extraction of SNOMED concepts from medical record texts. In *Proceedings of the 18th annual SCAMC* (pp. 179–183). Washington: McGraw Hill.

Oliver, D. E., Shahar, Y., Shortliffe, E. H., & Musen, M. A. (1999). Representation of change in controlled medical terminologies. *Artificial Intelligence in Medicine*, *15*, 53–76.

Osman, L. M., Abdalla, M. I., Beattie, J. A. G., Ross, S. J., Russell, I. T., Friend, J. A., et al. (1994). Reducing hospital admission through computer supported education for asthma patients. *British Medical Journal*, *308*, 568–571.

Ownsby, R. L. (2005). Influence of vocabulary and sentence complexity and passive voice on the readability of consumer-oriented health information on the Internet. In *AMIA 2005 symposium proceedings* (pp. 585–588).

Paice, C. (1990). Another stemmer. *SIGIR Forum*, *24*, 56–61.

Pala, K., & Smrz, P. (2004). Top ontology as a tool for semantic role tagging. In *Proceedings of the 4th international conference on language resources and evaluation (LREC 2004).* Lisbon.

Peckham, T. J. (1994). 'doctor, have i got a fracture or a break'? *Injury*, *25*(4), 221–2.

Pentheroudakis, J., & Vanderwende, L. (1993). Automatically identifying morphological relations in machine-readable dictionaries. In *Ninth annual confenrence of the UW center for the new OED and text research* (pp. 114–131).

Petersen, M., & Wagner, J. (2005). Digital corpora of interaction data for research and education. In P. J. Henrichsen (Ed.), *CALL for the nordic languages.* Copenhagen Business School. (Available from `http://www.sdu.dk/Hum/graduateschool/digitalcorpora1.pdf` [accessed 7 May 2006])

Pirrelli, V., & Yvon, F. (1999). The hidden dimension: a paradigmatic view of data-driven NLP. *JETAI*, *11*, 391–408.

Plovnick, R. M., & Zeng, Q. (2004). Reformulation of consumer health queries with professional terminology: as pilot study. *J Med Internet Res*, *6*(3), e27.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137.

Power, R., & Scott, D. (1998). Multilingual authoring using feedback texts. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th annual meeting of the Association for Computational Linguistics* (pp. 1053–1059). (Available from `http://mcs.open.ac.uk/rp3242/papers/coling98.pdf` [accessed 8 May 2006])

Predoiu, L., Feier, C., Scharffe, F., Bruijn, J. de, Martín-Recuerda, F., Manov, D., et al. (2004). *State-of-the-art survey on ontology merging and aligning V2.* (Deliverable D4.2.2. Available from `http://www.sekt-project.org/rd/deliverables/wp04/`

`sekt-d-4-2-2-SOA%20Survey%20on%20Ontology%20Merging%20and%20Aligning%20v2.pdf` [accessed 5 May 2006])

Radev, D. R., & McKeown, K. R. (1997). Building a generation knowledge source using Internet-accessible newswire. In *Proceedings of the 5th conference on applied natural language processing* (p. 221-228). Washington, DC.

Ramdass, M. J., Naraynsingh, V., Maharaj, D., Badloo, K., Teelucksingh, S., & Perry, A. (2001). Question of 'patients' versus 'clients'. *J Qual Clin Pract*, *21*(1–2), 14–5.

Ratnaparkhi, A. (1997). A maximum entropy model for part-of-speech tagging. In *EMNLP 1997*.

Rayson, P., & Wilson, A. (1996). The ACAMRIT semantic tagging system: progress report. In L. J. Evett & T. G. Rose (Eds.), *Language engineering for document analysis and recognition, LEDAR, AISB96 workshop proceedings* (pp. 13–20). Brighton, England.

Rector, A., Bechhofer, S., Goble, C. A., Horrocks, I., Nowlan, W., & Solomon, W. (1997). The GRAIL concept modelling language for medical terminology. *Artificial Intelligence in Medicine*, *9*, 139–171.

Rector, A., Gangemi, A., Galeazzi, E., Glowinski, A., & Rossi-Mori, A. (1994). The GALEN CORE model schemata for anatomy: Towards a re-usable application-independent model of medical concepts. In *Twelfth international congress of the european federation for medical informatics, mie-94* (pp. 229–233). Lisbon, Portugal.

Rector, A., & Rogers, J. (2006, December). *Ontological & practical issues in using a description logic to represent medical concepts: Experience from GALEN* (Tech. Rep.). School of Computer Science, the University of Manchester. (Preprint Series, CSPP-35)

Reiter, E., & Osman, L. M. (1997). Tailored patient information: Some issues and questions. In *ACL workshop: From research to commercial applications, making technology work in practice.*

Reiter, E., Robertson, R., & Osman, L. (2003). Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, *144*, 41–58.

Resnik, P. (1999, June). Mining the web for bilingual text. In *ACL'99 – proceedings of the 37th annual meeting of the association for computational linguistics* (pp. 527–534). College Park, MD, USA.

Richardson, T. (1996). The terminology of patient-focused care nouns as verbs, adjectives as nouns. *Revolution*, *6*(1), 31–34.

Rodning, C. B. (1992). Coping with ambiguity and uncertainty in patient-physician relationships: II. traditio argumentum respectus. *J Med Humanit*, *13*(3), 147–56.

Romacker, M., Schnattinger, K., Hahn, U., Schulz, S., & Klar, R. (2004). *A natural language understanding system for knowledge-based analysis of medical texts.*

Ruch, P., Baud, R., & Geissbühler, A. (2002). Evaluating and reducing the effect of data corruption when applying bag of words approaches to medical records. *International Journal of Medical Informatics*, *67*, 75–83.

Ruland, C. M., & Bakken, S. (2001). Representing patient preference-related concepts for inclusion in electronic health records. *J Biomed Inform*, *34*(6), 415–22.

Sadan, B. (2002). Patient impowerment and the asymetry of knowledge. *Stud Health Technol Inform*, *90*, 514–518.

Sager, N., Lyman, M., Bucknall, C., Nhan, N., & Tick, L. J. (1994, Mar/Apr). Natural language processing and the representation of clinical data. *Journal of the American Medical Informatics Association*, *1*(2).

Scanlan, D., Siddiqui, F., Perry, G., & Hutnik, C. M. (2003). Informed consent for cataract

surgery: what patients do and do not understand. *J Cataract Refract Surg*, *29*(10), 1904–12.

Scherpbier, H. J., Abrams, R. S., Roth, D. H., & Hail, J. J. (1994). A simple approach to physician entry of patient problem list. In *Proc annu symp comput appl med care* (pp. 206–10).

Schiller, A., Teufel, S., Stöckert, C., & Thielen, C. (1995). *Guidelines für das Tagging deutscher Textcorpora mit STTS* (Tech. Rep.). Universität Stuttgart and Universität Tübingen. (Available from `http://www.sfs.nphil.uni-tuebingen.de/Elwis/stts/stts.html` [accessed 15 March 2006])

Schmidt, H. (1997). Probabilistic part-of-speech tagging using decision trees. In D. Jones & H. Somers (Eds.), *New methods in language processing studies in computational linguistics* (pp. 154–164). London: UCL Press.

Schone, P., & Jurafsky, D. (2001). Knowledge-free induction of inflectional morphologies. In *NAACL.*

Schulz, S., Markó, K., Sbrissia, E., Nohama, P., & Hahn, U. (2004, August). Cognate mapping: A heuristic strategy for the semi-supervised acquisition of a Spanish lexicon from a Portuguese seed lexicon. In *COLING 2004 – proceedings of the 20th international conference on computational linguistics* (Vol. 2, pp. 813–819). Geneva, Switzerland.

Senda, Y., Sinohara, Y., & Okumura, M. (2004). A support system for revising titles to stimulate the lay reader's interest in technical achievements. In *Proceedings of COLING 2004* (pp. 155–161). Geneva.

Shackle, E. M. (1985). Psychiatric diagnosis as an ethical problem. *J Med Ethics*, *11*(3), 132–4.

Siddhartan, A. (2002). An architecture for a text simplification system. In *Proceedings of LREC 2002.* Las Palmas.

Simard, M., Foster, G. F., & Isabelle, P. (1992). Using cognates to align sentences in bilingual corpora. In *Proceedings of the 4th international conference on theoretical and methodological issues in machine translation.* Montréal, Canada. (Available from `http://www-rali.iro.umontreal.ca/Publications/sfiTMI92.ps/` [accessed 31 August 2002])

Simov, K., Peev, Z., Kouylekov, M., Simov, A., Dimitrov, M., & Kiryakov, A. (2001). CLaRK – an XML-based system for corpora development. In *Corpus linguistics 2001* (pp. 558–560). Lancaster, UK.

Skinner, C. S., Strecher, V. J., & Hospers, H. (1994). Physicians' recommendations for mammography: Do tailored messages make a difference? *American Journal of Public Health*, *84*(1), 43–49.

Smadja, F., McKeown, K. R., & Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, *21*(4).

Smith, B., & Fellbaum, C. (2004, August). Medical WordNet: A new methodology for the construction and validation of information. In *Proceedings of coling: The 20th international conference on computational linguistics* (pp. 371–382). Geneva.

Smith, C. A., Stavri, P. Z., & Chapman, W. W. (2002a). In their own words? a terminological analysis of e-mails to a cancer information service. In *Annual symposium of the american medical informatics association (AMIA)* (pp. 697–701).

Smith, C. A., Stavri, P. Z., & Chapman, W. W. (2002b). In their own words? A terminological analysis of e-mail to a cancer information service. In *Proceedings of the AIMA annual symposium* (pp. 697–701).

Smith, S. L., & Hamm, R. M. (1998). Patrient certification through mutual problem lists. *Mil Med*, *163*(11), 786–8.

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

Somers, H., & Lovel, H. (2003). Computer-based support for patients with limited english. In *Proceedings of the 7th international EAMT workshop on MT and other language technology tools.*

Soualmia, L. F., Darmoni, S. J., Douyère, M., & Thirion, B. (2003). Modelisation of consumer health information in a quality-controlled gateway. *Stud Health Technol Inform*, *95*, 701–6.

Spadaro, R. (2003, March). *European Union citizens and sources of information about health.* (EUROBAROMETER 58.0. European Opinion Research Group (EORG). Available from `http://europa.eu.int/comm/health/ph_information/documents/eb_58_en.pdf` [accessed 8 May 2006])

Spasic, I., & Ananiadou, S. (2005). A flexible measure of contextual similarity for biomedical terms. In *Pacific symposium on biocomputing* (Vol. 10, pp. 197–208).

Spees, C. M. (1991). Knowledge of medical terminology among clients and families. *Image J Nurs Sch*, *23*(4), 225–9.

Spiro, D., & Heidrich, F. (1983). Lay understanding of medical terminology. *J Fam Pract*, *17*(2), 277–9.

Spri. (1999). *Metoder och principer i terminologiarbetet.* (Spri rapport 481. Available from `http://www.sos.se/epc/klassifi/filer/Terminologi/rap481.pdf` [accessed 5 May 2006])

Sripada, S., Reiter, E., Hunter, J., & Yu, J. (2003). Summarizing neonatal time series data. In *Proceedings of the research note sessions of the EACL'03* (pp. 167–170).

Stead, M., Eadie, D., Gordon, D., & Angus, K. (2005). "hello, hello–it's english i speak!": a qualitative exploration of patients' understanding of the science of clinical trials. *J Med Ethics*, *31*(11), 664–9.

Steensma, D. P. (2006). Are myelodysplastic syndromes "cancer"? unexpected adverse consequences of linguistic ambiguity. *Leuk Res.*

Steimann, F. (1998). Dependency parsing for medical language and concept representation. *Artificial Intelligence in Medicine*, *12*, 77–86.

Stevens, D. P., Stagg, R., & Mackay, I. R. (1977). What happens when hospitalized patients see their own records. *Ann Intern Med*, *86*(7), 474–7.

Strecher, V. J., Kreuter, M., Boer, D. J. D., Kobrin, S., Hospers, H. J., & Skinner, C. S. (1994). The effects of computer-tailored smoking cessation messages in family practice settings. *J Fam Pract.*, *39*(3), 262–270.

Surjan, G., & Heja, G. (2003). About the language of hungarian discharge reports. *Stud Health Technol Inform*, *95*, 869–73.

Taira, R. K., & Soderland, S. G. (1999). A statistical natural language processor for medical reports. In *Proceedings of the AMIA annual symposium* (pp. 970–974).

Tanguy, L., & Hathout, N. (2002). Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du web. In *Traitement automatique des langues naturelles (TALN)* (pp. 245–254). Nancy.

Tennison, W. P. J. (2002). The layered markup and annotation language. In *Extreme markup languages.* Montréal, Canada.

Theron, P., & Cloete, I. (1997). Automatic acquisition of two-level morphological rules. In *ANLP* (pp. 103–110).

Thomas, J. A., & Wilson, A. (1996). Methodologies for studying a corpus of doctor-patient interaction. In J. A. Thomas & S. M. H. (Eds.), *Using corpora for language research: Studies in the honour of Geoffrey Leech* (pp. 92–109). London: Longman.

Thompson, G. (1999). Acting the part: lexico-grammatical choices and contextual factors. In M. Ghadessy (Ed.), *Text and context in functional linguistics* (pp. 101–124). Amsterdam, Philadelphia, USA: John Benjamins.

Thompson, H. J. (2005). Fever: a concept analysis. *J Adv Nurs*, *51*(5), 484–92.

Torgersson, O., & Falkman, G. (2002). Using text generation to access clinical data in a variety of contexts. In G. Surján, R. Engelbrecht, & P. McNair (Eds.), *Health data in the information society. proceedings of MIE2002* (Vol. 90, pp. 460–465). IOS Press.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature rich part-of-speech tagging with a cyclic dependency network. In *HLT-NAACL 2003* (pp. 252–259).

Toutanova, K., & Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Joint SIGDAT conference on empirical methods in natural language processing and very large corpora (EMNLP/VLC-2000).* Hong Kong.

Travers, D. A., & Haas, S. W. (2003). Using natural language entries to build a concept-oriented terminology for patients' chief complaints in the emergency department. *J Biomed Inform*, *36*(4–5), 260–70.

Tse, T., & Soergel, D. (2003). Exploring medical expressions used by consumers and the media: an emerging view of consumer health vocabularies. In *Annual symposium of the american medical informatics association (AMIA)* (pp. 674–8).

Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S., et al. (2005). Developing a robust part-of-speech tagger for biomedical text. In *Advances in informatics – 10th panhellenic conference on informatics* (pp. 382–392).

Viegas, E., Gonzales, M., & Longwell, J. (1996). Morpho-semantics and constructive derivational morphology: A transcategorical approach. In *Memoranda in computer and cognitive science.* (Available `http://crl.nmsu.edu/Research/Projects/mikro/htmls/misc-htmls/mikro-pub.html` [accessed 17 August 1999])

Vintar, S. (2001). Using parallel corpora for translation-oriented term extraction. *Babel Journal*, *47*(2). (Available from `http://lojze.lugos.si/~spela/` [accessed 15 March 2006])

Vintar, S., Buitelaar, P., Ripplinger, B., Sacaleanu, B., Raileanu, D., & Prescher, D. (2002). An efficient and flexible format for linguistic and semantic annotation. In *Third international language resources and evaluation conference.* Las Palmas, Spain.

Waisman, Y., Siegal, N., Chemo, M., Siegal, G., Amir, L., Blachar, Y., et al. (2003). Do parents understand emergency department discharge instructions? a survey analysis. *Isr Med Assoc J*, *5*(8), 567–70.

Weizenbaum, J. (1966, January). ELIZA – a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.

Werry, C. C. (1996). Linguistic and interactional features of Internet Relay Chat. In S. Herring (Ed.), *Computer-mediated communication: Linguistic, social and cross-cultural perspectives* (pp. 47–63). John Benjamins.

Werry, C. C. (2004). Linguistic and interactional features of Internet Relay Chat. In G. Samp-

son & D. McCarthy (Eds.), *Corpus linguistics. readings in a widening discipline* (pp. 340–352). London, New York: continuum.

White, P., Singleton, A., & Jones, R. (2004). Copying referral letters to patients: the views of patients, patient representatives and doctors. *Patient Educ Couns*, *55*(1), 94–8.

Widdows, D., Peters, S., Cederberg, S., Chan, C.-K., Steffen, D., & Buitelaar, P. (2003, July). Unsupervised monolingual and bilingual word-sense disambiguation of medical documents uisng UMLS. In *Proceedings of the ACL 2003 workshop on natural language processing in biomedicine.*

Williams, N., & Ogden, J. (2004). The impact of matching the patient's vocabulary: a randomized control trial. *Fam Pract*, *21*(6), 630–5.

Williams, S. H. (2004). *Natural Language Generation (NLG) of discourse relations for different reading levels.* Unpublished doctoral dissertation, University of Aberdeen.

Wilson, A., & Rayson, P. (1993). Automatic content analysis of spoken discourse: a report on work in progress. In C. Souter & E. Atwell (Eds.), *Corpus based computational linguistics* (pp. 215–226). Amsterdam: Rodopi.

Wittich, A. C., Junnila, J., & Buller, J. (2003). Would your patients prefer to be your clients? *J Am Osteopath Assoc*, *103*(10), 485–7.

Wu, C.-H., Yu, L.-C., & Jang, F.-L. (2005, Nov/Dec). Using semantic dependencies to mine depressive symptoms from consultation records. *IEEE Intelligent Systems*, *20*(6), 50–57.

Wynn, R. (1998). *Provider–patient interaction: A corpus-based study of doctor–patient and student–patient interaction.* Unpublished doctoral dissertation, Bergen University.

Wynn, R. (1999). *Provider–patient interaction: A corpus-based study of doctor–patient and student–patient interaction* (No. 8). Kristiansand, Norway: Høyskoleforlaget. (ISBN: 82-7634-172-1. First published as a PhD dissertation in 1998 (Wynn, 1998). Summary: `http://helmer.aksis.uib.no/nlb/nlb-enews_1999_9.txt`)

Xu, J., & Croft, B. W. (1998). Corpus-based stemming using co-occurrence of word variants. *ACM Transactions on Information Systems*, *16*(1), 61–81.

Yarnall, K. S., Michener, J. L., Broadhead, W. E., Hammond, W. E., & Tse, C. K. (1995). Computer-prompted diagnostic codes. *J Fam Pract*, *40*(3), 257–62.

Yeh, J.-F., Wu, C.-H., Chen, M.-J., & Yu, L.-C. (2004, August). Automated alignment and extraction of bilingual ontology for cross-language domain-specific applications. In *Proceedings of the 20th international conference on computational linguistics.* Geneva.

Zeng, Q., Crowell, J., Plovnick, R. M., Kim, E., Ngo, L., & Dibble, E. (2006). Assisting consumer health information retrieval with query recommendadtions. *J Am Med Inform Assoc*, *13*(1), 80–90.

Zeng, Q., Kim, E., Crowell, J., & Tse, T. A. (2005). Text corpora-based estimation of the familiarity of health terminology. In *Proceedings of the sixth international symposium on biological and medical data analysis (ISBMDA'05).*

Zeng, Q., Kogan, S., Ash, N., & Greenes, R. A. (2001). Patient and clinician vocabulary: how different are they? In *Medinfo* (pp. 399–403).

Zeng, Q., & Tse, T. (2005). Case for developing an open-source first-generation consumer health vocabulary. *J Am Med Inform Assoc.*

Zeng, Q., & Tse, T. (2006). Exploring and developing consuming health vocabulary. *J Am Med Inform Assoc*, *13*(1), 24–9.

Zeng, Q., Tse, T., Crowell, J., Divita, G., Roth, L., & Browne, A. C. (2005). Identifying

consumer-friendly display (CFD) names for health concepts. In *Annual symposium of the american medical informatics association (AMIA).*

Zielstorff, R. D. (2003). Controlled vocabularies for consumer health. *J Biomed Inform*, *36*(4–5), 326–33.

Zur-Mühlen, B. von. (2005/05). Nu fråntas läkarna språket i de digitala datajournalerna. *Sjukhusläkaren.* (Available from `http://www.sjukhuslakaren.org/505_spraket.html` [accessed 5 May 2006])

Zweigenbaum, P., & Grabar, N. (2000). *A contribution of medical terminology to medical language processing resources: Learning morphological knowledge from thesauri.* (Available from `http://citeseer.ifi.unizh.ch/zweigenbaum00contribution.html` [accessed 2 May 2006])

Zweigenbaum, P., Hadouche, F., & Grabar, N. (2003). Apprentissage de relations morphologiques en corpus. In *Traitement automatique des langues naturelles (TALN).*